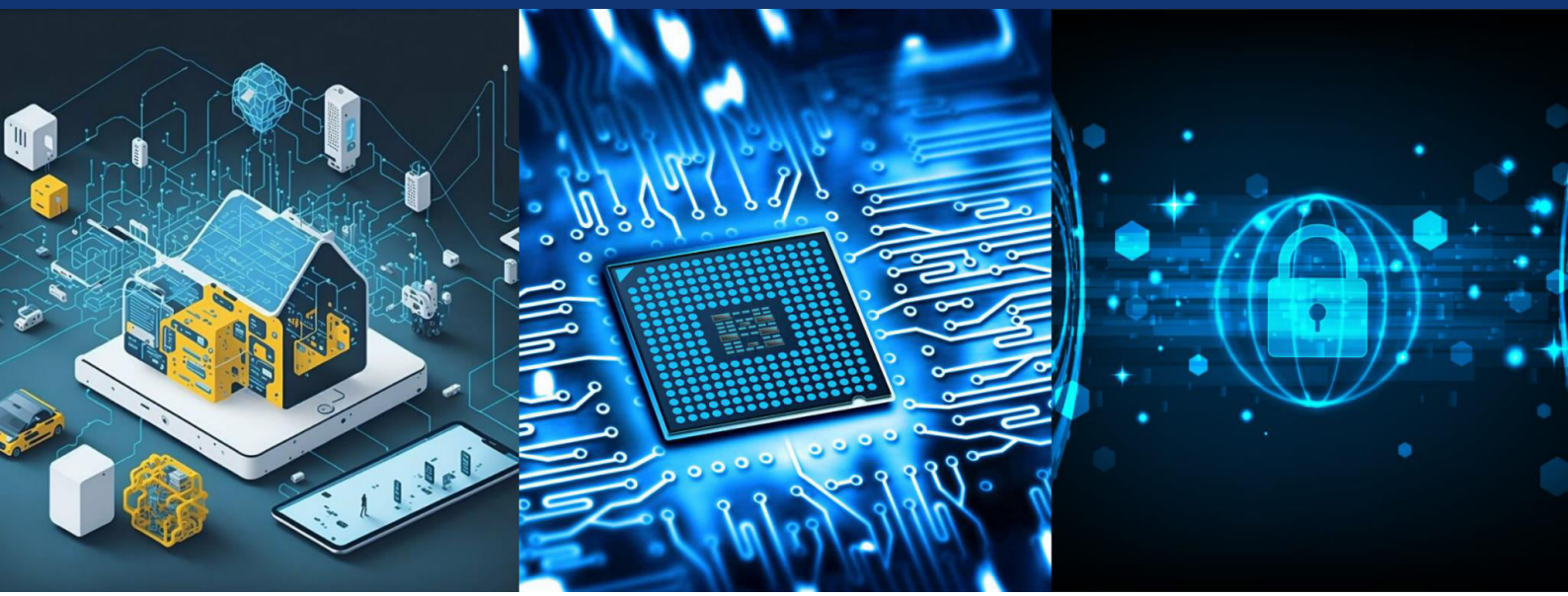
IJIRCCE

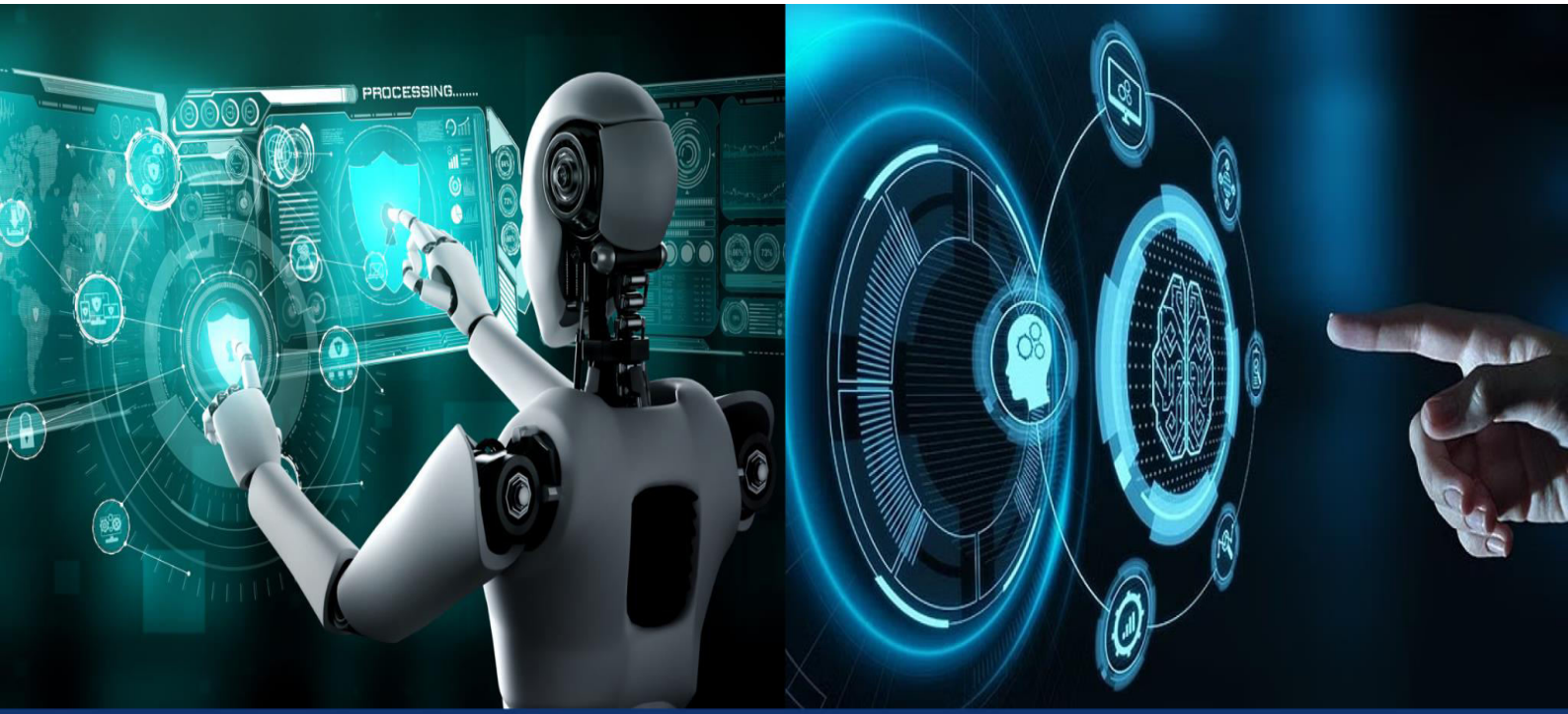
ISSN(O): 2320-9801

ISSN(P): 2320-9798



# International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)




**Impact Factor: 9.927**

**Volume 14, Issue 6, June 2026**

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

# Combating Online Job Scams Using a Multi-Algorithm Ensemble-Based AI Detection System

**Mr. Palakuri Swamy, Dr. P. Sammulal**

Post Graduate Student, Department of Computer Science and Engineering(Data Science), Jawaharlal Nehru Technological University, Hyderabad, India

Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

**ABSTRACT:** In addition to increasing employment prospects, the quick expansion of internet job portals has led to a surge in fraudulent job ads. This project introduces a Fake Job Detection System based on Artificial Intelligence (AI) that automatically detects bogus job postings. Optical Character Recognition (OCR)-based picture extraction, Uniform Resource Locator (URL), and text input are all accepted by the system. Text preprocessing and feature extraction are done using Natural Language Processing (NLP) methods and Term Frequency–Inverse Document Frequency (TF-IDF). Additionally, fraud indications including money demands, exaggerated salary, and urgent language are identified by a rule-based engine. Long Short-Term Memory (LSTM), Random Forest, and Logistic Regression models are used to analyze the generated data. To increase accuracy and dependability, an ensemble learning technique is used to provide the final forecast. Through a Flask web application, the system gives real-time prediction with an accuracy of about 98%. Additionally, Streamlit is used to deploy the solution on Snowflake for scalable cloud accessibility. This study shows how deep learning (DL), machine learning (ML), and artificial intelligence (AI) may successfully shield people from online recruiting frauds.

**KEYWORDS:** Fake Job Detection, picture extraction, Uniform Resource Locator.

## I. INTRODUCTION

By making job hunting quicker and simpler, online job portals like Indeed, LinkedIn, Glassdoor, and Naukri.com have revolutionized the hiring process. Every day, millions of job openings are posted in various regions and sectors. These platforms facilitate effective connections between companies and qualified applicants. Using digital platforms, job searchers may apply from any location. Both companies and candidates now have more accessibility and convenience thanks to the expansion of online hiring. Despite these advantages, the rise of online recruitment has also increased fraudulent job activities. Cybercriminals create fake job postings to cheat job seekers and steal sensitive information. Many fraudulent advertisements appear genuine by using professional language and company details. Victims are often asked to pay registration fees, security deposits, or share personal documents. These scams can lead to financial loss, identity theft, and emotional stress for applicants. Detecting such fake job postings manually has become increasingly difficult due to their realistic appearance. Traditional fraud detection methods mainly depend on manual review, user complaints, and keyword-based filtering systems. However, these approaches are not effective for handling the huge number of job postings published every day. Fraudsters continuously change their strategies to bypass existing detection mechanisms. Manual verification also requires significant time and human effort. As a result, an automatic and intelligent method for precisely identifying bogus job postings is desperately needed. Machine learning and artificial intelligence offer effective methods for resolving this issue. The AI-based Fake Job Detection System proposed in this research automatically determines if job advertisements genuine. Additionally, a rule-based engine recognizes typical fraud indications such as unreasonable salary offers and payment demands. Text input, URL-based extraction, and image-based OCR processing are all supported by the system. For scalable and user-friendly access, the finished solution is implemented using Flask and Snowflake with Streamlit.

## II. RELATED WORK

For the purpose of detecting phony jobs, a number of academics have suggested deep learning and machine learning methods. Gopi et al. (2025) enhanced detection accuracy by addressing class imbalance using Random Forest, MLP, and

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Naive Bayes with ADASYN. In order to find significant fraud signs in job advertising, Khalifa et al. (2022) used feature engineering and data analysis. By using BERT and RoBERTa models, Akram et al. (2024) were able to achieve good performance through contextual text interpretation. TF-IDF and Bag-of-Words methods for text-based categorization were investigated by Fating et al. (2023). Pranay et al. (2020) used the EMSCAD dataset and conventional machine learning techniques to obtain baseline findings. An LSTM-based model was presented by Ranparia et al. (2024) to identify sequential patterns in job descriptions. Udayshankar and associates. (2023) used data mining methods to find fraud patterns, including clustering and association rule mining. Nevertheless, the majority of current systems lack explainability and rely on a single strategy. By integrating LSTM, Logistic Regression, Random Forest, and Rule-Based Analysis in an ensemble framework for precise and comprehensible fake job identification, the suggested approach overcomes these drawbacks.

## III. PROPOSED ALGORITHM

### A. SYSTEM ARCHITECTURE:

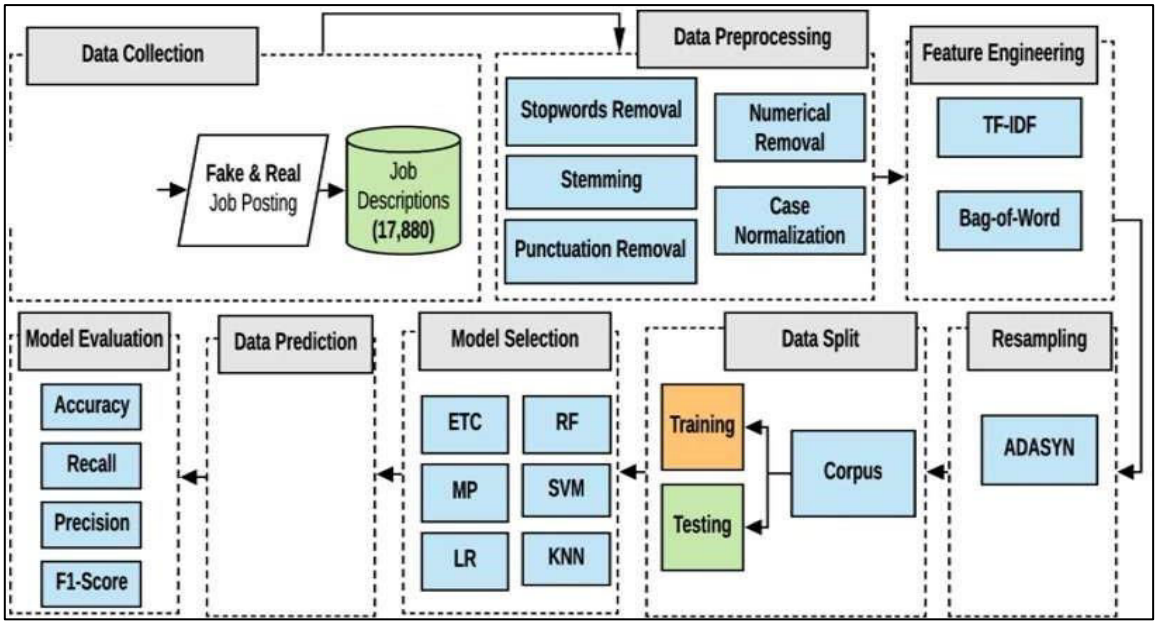



**Fig. 1. Proposed System Architecture**

### B. *DESIGN CONSIDERATIONS:*

- The accepts a variety of input forms, including text, URLs, and job ads with images.
- enhances data quality by using text preparation methods like as tokenization, stop-word removal, lemmatization, and text normalization.
- uses ADASYN oversampling to address dataset imbalance and enhance the identification of bogus job listings.
- Combines Logistic Regression, Random Forest, LSTM, and Rule-Based Analysis to increase prediction accuracy.
- makes forecasts clear and easy to interpret by offering fraud indicators and confidence scores.
- enables real-time deployment with Streamlit and Flask, guaranteeing scalability and accessibility.

### C. *DESCRIPTION OF THE PROPOSED ALGORITHM:*

The proposed algorithm aims to accurately identify fake job advertisements and classify them as genuine or fraudulent. The algorithm consists of five major stages: data collection, data preprocessing, feature extraction, classification using Logistic Regression, Random Forest, LSTM, and Rule-Based Analysis, and ensemble prediction. Finally, the system generates the classification result along with confidence scores and fraud indicators to help users identify suspicious job postings.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

**1. Data Collection Module**

The Data Collection Module is responsible for gathering job advertisement data from multiple sources. The system uses the Fake Job Postings Dataset obtained from Kaggle, which contains information such as job title, company profile, job description, requirements, benefits, employment type, salary range, and fraud labels. In addition to the training dataset, the system accepts user inputs in the form of direct text, URLs, and images. When a URL is provided, web scraping techniques are used to extract the job content, while OCR technology is applied to extract text from images. This module ensures that all collected data is valid and suitable for further processing.

➢ **Mathematical Representation:**

- D = {J1, J2, J3, ..., Jn}

Where:

- D = Dataset
- Ji = Individual Job Posting
- n = Total Number of Job Postings

Input Set:

- I = {T, U, M}

Where:

- T = Text Input
- U = URL Input
- M = Image Input

**2. DATA PREPROCESSING MODULE**

The Data Preprocessing Module cleans and transforms raw job posting content into a structured format suitable for machine learning analysis. Online job advertisements often contain HTML tags, URLs, special symbols, numbers, and irrelevant words that can negatively affect model performance. To address this issue, preprocessing operations such as lowercasing, tokenization, stop-word removal, punctuation removal, and lemmatization are applied. These techniques improve text quality and reduce noise, enabling more accurate feature extraction and classification.

**Mathematical Representation:**

- Jp = f(J)

Where:

- J = Raw Job Posting
- Jp = Preprocessed Job Posting
- f = Preprocessing Function

Preprocessing Function:

- f(J) = Lowercase + Tokenization
  + Stopword Removal
  + Lemmatization
  + Cleaning

**3. FEATURE EXTRACTION MODULE**

The Feature Extraction Module converts preprocessed textual content into numerical feature vectors that can be processed by machine learning algorithms. The proposed system uses the TF-IDF (Term Frequency-Inverse Document Frequency) technique to measure the importance of words within job advertisements. TF-IDF assigns higher weights to informative words while reducing the influence of frequently occurring terms. For deep learning models such as LSTM, Keras Tokenization is used to convert text into numerical sequences, followed by sequence padding to maintain a fixed input length.

**Mathematical Representation:**

Term Frequency:

- TF(t,d) =

Frequency of term t in document d

---

Total terms in document d

Inverse Document Frequency:

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- IDF(t) =

log(N / DF(t))

TF-IDF Score:

- TF-IDF(t,d) = TF(t,d) × IDF(t)

Feature Vector:

- V = TF-IDF(Jp)

Where:

- V = Feature Vector
- Jp = Preprocessed Text

**4. FAKE JOB DETECTION AND CLASSIFICATION MODULE**

The Fake Job Detection and Classification Module is the core component of the system. It analyzes extracted features using multiple machine learning and deep learning models to determine whether a job posting is genuine or fraudulent. Logistic Regression performs binary classification efficiently on TF-IDF vectors. Random Forest uses multiple decision trees to improve robustness and reduce overfitting. LSTM captures contextual and sequential relationships within textual data. The Rule-Based Engine identifies suspicious indicators such as advance payment requests, salary offers, urgency phrases, and missing company information.

The outputs generated by all models are combined using an Ensemble Learning approach. The final prediction is produced based on weighted voting, which improves overall detection accuracy and reliability. The system also provides fraud indicators and confidence scores to improve transparency and user trust.

**Mathematical Representation:**

**Individual Predictions:**

P1 = LogisticRegression(V)

P2 = RandomForest(V)

P3 = LSTM(V)

P4 = RuleBased(V)

Ensemble Prediction:

F = w1P1 + w2P2 + w3P3 + w4P4

Where:

w1 + w2 + w3 + w4 = 1

Final Decision:

If F ≥ 0.5

Fake Job

Else

Real Job

The ensemble model achieves approximately 98% accuracy, making it highly effective in detecting fraudulent job advertisements.

## IV. PSEUDO CODE

**Step 1:** Load the Fake Job Postings dataset containing both genuine and fraudulent job advertisements.

**Step 2:** Perform data preprocessing by:

- Removing missing and duplicate values.
- Removing HTML tags, URLs, and special characters.
- Converting text into lowercase.
- Removing stop words.
- Applying lemmatization.

**Step 3:** Separate the dataset into input features (X) and target labels (Y).

**Step 4:** Split the dataset into training and testing datasets.

**Step 5:** Apply TF-IDF Vectorization to convert textual data into numerical features.

**Step 6:** Initialize the Logistic Regression classifier with suitable hyperparameters.

**Step 7:** Train the Logistic Regression model using the training dataset.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

**Step 8:** Initialize the Random Forest classifier with suitable hyperparameters.
**Step 9:** Train the Random Forest model using the training dataset.
**Step 10:** Initialize and train the LSTM model using tokenized text sequences.
**Step 11:** Define Rule-Based Analysis using suspicious keywords and fraud indicators.
**Step 12:** Test the trained models using the testing dataset.
**Step 13:** Calculate performance metrics such as Accuracy, Precision, Recall, F1-Score, ROC-AUC Score, and Confusion Matrix.
**Step 14:** Upload or receive a new job advertisement for analysis.
**Step 15:** Determine the input type.
**IF Input = URL**
Extract job content using Web Scraping.
**ELSE IF Input = Image**
Extract text using OCR.
**ELSE**
Use direct text input.
**Step 16:** Apply the same preprocessing and feature extraction steps to the incoming job advertisement.
**Step 17:** Pass the processed data to the trained Logistic Regression, Random Forest, LSTM, and Rule-Based models.
**Step 18:** Generate prediction scores from all models.
**Step 19:** Perform Ensemble Prediction by combining all model outputs.
**Step 20:** Predict the job advertisement class.
**IF Prediction = Genuine**
Display **"Real Job Posting"**
Display confidence score.
**ELSE**
Display **"Fake Job Posting Detected"**
Display confidence score and fraud indicators.
**Step 21:** Display prediction results, model analysis, and fraud indicators on the Flask/Streamlit dashboard.
**Step 22:** Store analysis results for future monitoring and model improvement.
**Step 23:** Continue analyzing new job advertisements.
**Step 24:** End.
.

## V. SIMULATION RESULTS

The proposed Fake Job Detection System was implemented using Python and evaluated using the Fake Job Postings dataset obtained from Kaggle. The dataset contains both genuine and fraudulent job advertisements collected from online recruitment platforms. Data preprocessing techniques such as text cleaning, tokenization, stop-word removal, and lemmatization were applied to improve text quality. TF-IDF feature extraction was used to convert textual content into numerical representations suitable for machine learning models. The system utilizes Logistic Regression, Random Forest, LSTM, and Rule-Based Analysis for detecting fraudulent job advertisements. Performance evaluation was carried out using Accuracy, Precision, Recall, F1-Score, ROC-AUC Score, and Confusion Matrix. Experimental results showed that the ensemble learning approach achieved high classification accuracy and improved fraud detection performance.

The system effectively identified fake job postings while maintaining low false positive and false negative rates. The proposed solution supports text, URL, and image-based job advertisement analysis through web scraping and OCR technology. A Flask and Streamlit-based dashboard was developed to display prediction results, confidence scores, fraud indicators, and model performance metrics. The dashboard provides an interactive and user-friendly interface for analyzing job advertisements. The system successfully helps users identify suspicious job postings and avoid recruitment fraud. The combination of machine learning, deep learning, and rule-based techniques improves reliability and robustness. The experimental results demonstrate the effectiveness of the proposed approach in real-world recruitment environments. The system provides a scalable and efficient solution for fake job detection. Overall, the proposed framework enhances online recruitment security and protects job seekers from fraudulent job advertisements.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

## EVALUATION AND INTERPRETATION OF SYSTEM INTERFACES:

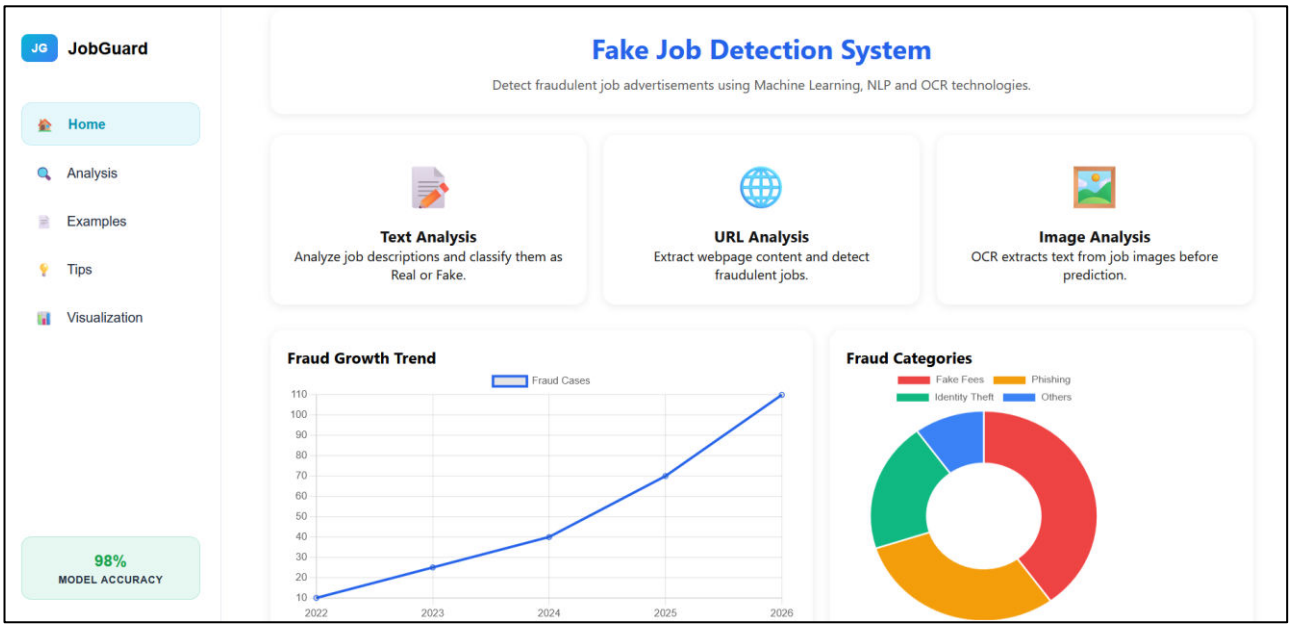



**Fig.2. Home Page**

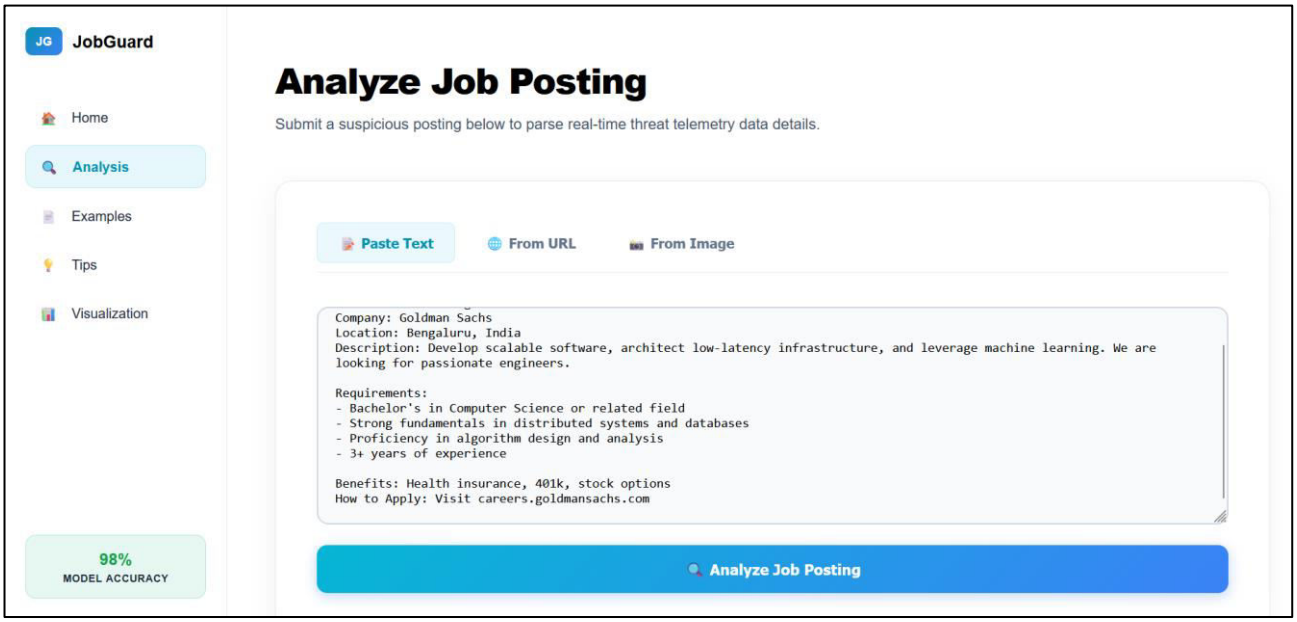



**Fig.3. Analyse Job Posting Based On Text**

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406058

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

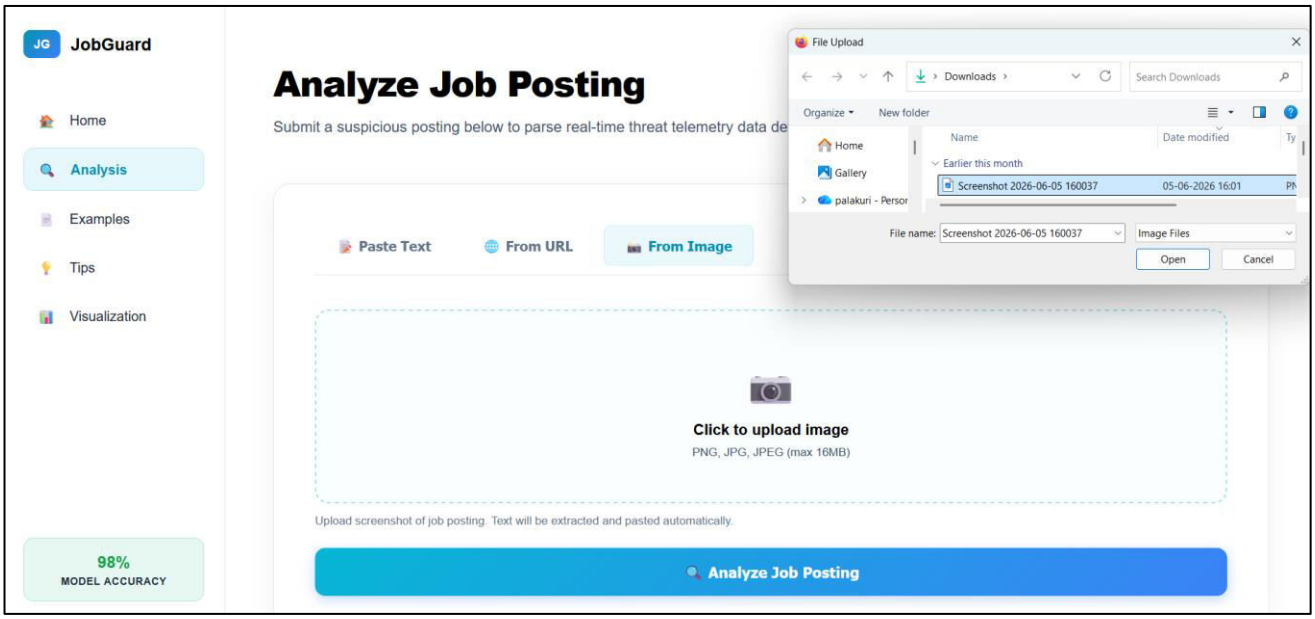



**Fig.4. Analyse Job Posting Based On Images**

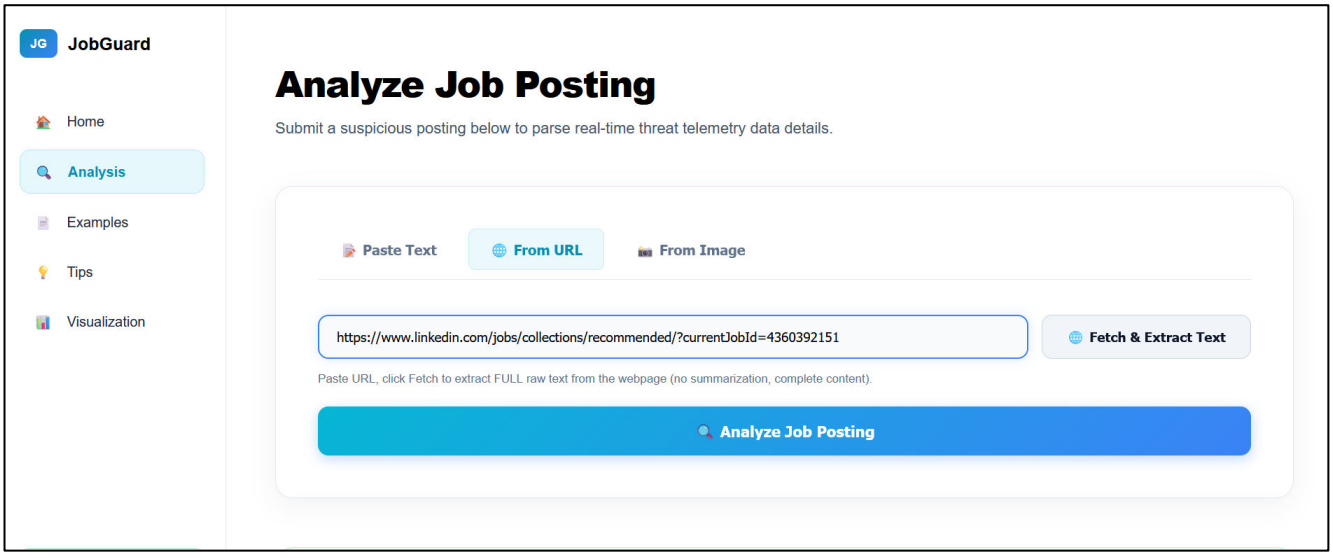



**Fig.5. Analyse Job Posting Based On URL**

### B. Metrics for Evaluation

Standard classification measures such as Accuracy, Precision, Recall, and F1-Score are used to assess the effectiveness of the suggested AI-Based Fake Job Detection System. These measures aid in evaluating how well the ensemble architecture differentiates between legitimate and fraudulent job advertisements.

- TP stands for True Positives, or the accurate identification of bogus job ads.
- TN stands for True Negatives, which are legitimate job advertisements that have been accurately.
- False Positives (FP) are real job ads that are mistakenly labeled as fraudulent.
- False Negatives (FN) are fake job postings that are mistakenly labeled as legitimate.
- **Accuracy**

The ratio of accurately categorized job postings to all job listings examined is known as accuracy.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- It evaluates the system's overall efficacy in distinguishing between legitimate and fraudulent job postings. Equation (5): Accuracy = (TP + TN) / (TP + TN + FP + FN)
- **Precision**

The ratio of accurately predicted fraudulent job listings to the total number of postings indicated as fraudulent is known as precision.

Equation (6) uses mining and clustering to find fraud patterns. Precision = TP / (TP + FP). Nevertheless, the majority of current systems lack explainability and rely on a single strategy.

- **Efficiency**

Precision is defined as the ratio of correctly predicted fraudulent job advertising to all postings anticipated to be fake. Precision is equal to TP/TP + FP in equation (6).

- **Recall**

The percentage of real fraudulent job posts that the algorithm accurately identifies is known as recall.

Equation (7): Recall = TP / (TP + FN)

- **F1-Score**

The F1-Score offers a fair assessment of the classifier's performance and is calculated as the harmonic mean of Precision and Recall.

Equation (8) states that F1-Score is equal to 2 × (Precision × Recall) / (Precision + Recall).

The performance of the suggested model is compared using these assessment measures, which also guarantee accurate identification of bogus job postings.

## VI. CONCLUSION AND FUTURE WORK

The proposed Using machine learning and deep learning approaches, the proposed AI-Based Fake Job Detection System successfully detects bogus job postings. To differentiate between real and fraudulent recruitment ads, the system uses preprocessing, feature extraction, and classification techniques to examine the content of job postings. The framework increases detection accuracy and reliability by integrating models like LSTM, Random Forest, Logistic Regression, and Rule-Based Analysis. The technique increases confidence in online job sites and helps job seekers avoid recruiting frauds. Based on Accuracy, Precision, Recall, and F1-Score metrics, experimental findings show that the suggested method provides successful classification performance.

Future work will incorporate real-time job monitoring from online recruiting platforms, enhance multi-modal analysis utilizing text, URLs, and photos, and integrate sophisticated transformer-based models like BERT and RoBERTa. Larger datasets, cloud-based deployment, and explainable AI methods can all enhance scalability, interpretability, and detection performance in practical settings.

## REFERENCES

1. R. Gopi et al., "Fake Job Post Detection Using Machine Learning: An ADASYN Approach", 2025.
2. Ahamed Khalifa Z. et al., "Online Recruitment Fraud Detection", 2022.
3. Natasha Akram et al., "Online Recruitment Fraud Detection Using Deep Learning Approaches", 2024.
4. Jyoti Fating et al., "Fake Job Listing Detection Using Machine Learning Approach", 2023.
5. Varma Pranay et al., "Fake Job Recruitment Detection Using Machine Learning", 2020.
6. Dhruvil Ranparia et al., "Fake Job Prediction Using Sequential Network", 2024.
7. Shri Udayshankar et al., "Fake Job Post Prediction Using Data Mining", 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details