Enhancing Fraudulent Job Recruitment Detection: A Comprehensive Ensemble Modeling Approach

1st B.Mayuri
Computer Science and Engineering
Vignan University
Vadlamudi ,Andhra Pradesh,India
mayuri4613@gmail.com

2nd Bala Sahithi
Computer Science and Engineering
Vignan University
Vadlamudi ,Andhra Pradesh,India
balasahithi7@gmail.com

3rdAbhiram
Computer Science and Engineering
Vignan University
Vadlamudi ,Andhra Pradesh,India
abhiramchowdary937@gmail.com

4th Harsha
Computer Science and Engineering
Vignan University
Vadlamudi ,Andhra Pradesh,India
vardhan3645@gmail.com

5th Nithin Reddy
Computer Science and Engineering
Vignan University
Vadlamudi ,Andhra Pradesh,India
nithinreddyallamsd@gmail.com

***Abstract*—In Recent years ,Fraud job postings has been increasing in number and there are many studies,researches done so far to prevent this from happening by removing the fake online job postings all over from organizational venues who are publishing these online. Almost every model has an average accuracy equal to greater than 95. We have worked on ensemble model for training of EMSCAD dataset which contains about 18000 fraudulent job postings . The ensmble model is made of ,A BoW model using simple Count vectorizer and linear support vector machine(SVM), A BoW model using TF-IDF vectorizer and random forest model and A XGBoost classifier model on the non-textual features which caries the accuracy of nearly 98.**

***Index Terms*—Deep Learning,Logistic Regression,Decision Tree,Support Vector Machine ,Random Forest,AdaBoost,XGBoost Classifier,TF-IDF vectorizer**

I. Introduction

”In recent years, with the proliferation of modern technology and social communication, the prevalence of advertising new job opportunities has escalated. Consequently, detecting fake job postings has emerged as a pressing concern in today’s society. Online job scams have proliferated, posing significant challenges in distinguishing between legitimate and fraudulent job advertisements. These scams often lure unsuspecting individuals into applying for seemingly genuine positions, only to discover later that they have fallen victim to fraudsters seeking to pilfer personal information such as residential addresses, email IDs, contact numbers, and even banking details. The digital realm provides a fertile ground for criminals to operate across borders, evading detection until it’s too late.

Similar to other classification tasks, predicting fake job postings presents numerous challenges. This paper proposes leveraging various data mining techniques and classification algorithms, including K-Nearest Neighbors (KNN), decision trees, support vector machines, naive Bayes classifiers, random forest classifiers, multilayer perceptrons, and deep neural networks, to distinguish between authentic and fraudulent job posts. Experimentation is conducted on the Employment Scam Aegean Dataset (EMSCAD), comprising 18,000 samples. Job seekers rely on these job advertisements to explore their options based on factors such as time availability, qualifications, experience, and suitability. The recruitment process is increasingly influenced by the internet and social media, with the impact of social media on job advertisement being particularly significant. The rapid proliferation of job posting opportunities has, however, led to a rise in the percentage of fraudulent job postings, causing distress to job seekers.

Fraudsters seeking to acquire individuals’ personal information, such as insurance details, banking information, income tax records, dates of birth, and national identification, create counterfeit job advertisements. Advance fee scams involve fraudsters requesting money for purported administrative charges, information security checks, or management costs. In some cases, fraudsters pose as employers and request passport details, bank statements, or driving licenses as part of a pre-employment verification process.Many students who ought to get placed in off-campus jobs also get troubled by these kind of fraud job postings online. Illegal money laundering scams involve convincing individuals, particularly students, to transfer money into fraudulent accounts and then transfer it back. Scammers often utilize tactics such as creating fake company websites, cloning bank websites, and forging official-looking documents to ensnare job seekers. They predominantly target individuals through email rather than face-to-face communication and frequently leverage social media platforms like LinkedIn to pose as recruitment agencies or headhunters. Their aim is to portray their company profiles or websites as realistic as possible to lure job seekers into their trap, collecting information and exploiting it for financial gain or other nefarious purposes.”

Identify applicable funding agency here. If none, delete this.

II. BACKGROUND

We have used ensemble model for our model to train and test the dataset to give the maximum accuracy and work more efficiently. Our ensemble model consists of linear Support Vector Machine algorithm, random Forest Classifier and XGBoost classifier. Combining these three efficient algorithms together gave us the accuracy of approximately 98.

A. Linear SVM

Linear SVM algorithm works by finding the optimal hyperplane that best separates the classes in the input feature space.

y=sign(wx+b)

B. Random Forest Algorithm

This algorithm constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

$=1 = 1()_{y=N1i=1Nfi(x)}$

C. XGBoost Classifier

This is an library designed which uses a gradient boosting framework, combining the outputs from multiple weak learners (typically decision trees) to make accurate predictions.

Formula: $= = 1()_{y=k=1Kfk(x)}$

III. RELATED WORK

In this part of section we are going to review the some of the previous works done on fraud job postings detection .

Prashanth. c and his team in [1] presented Online Fake Job Advert Detection Application using Machine learning(2022) in which first they have first trained few of the machine learning models such as Random Forest, Logistic Regression,SVM and Blockchain technology along with NLP then have trained using Block chain technology to create an online tool called REVEAL.

Habiba and her team in [2] presented A comparative study on fake job post prediction using Data Mining Techniques(2021), in this paper,their model used softmax function for classification technique.They also used ensemble classifier (Bi-GRU CNN, Bi-GRULSTM CNN) using majority voting technique to increase classification accuracy. They found 66 classification accuracy using TextCNN and 70 accuracy for Bi-GRU- LSTM CNN individually. This classification task performed best with ensemble classifier having an accuracy of 72.4.

Nessa and members in [3] have done Recruitment Scam Detection Using Gated Recurrent Unit in this first they have tested the data with 9 models such as e Logistic Regression , K-Nearest Neighbors (KNN), Decision Tree (DT) , Na¨ıve Bayes (NB) , Random Forest (RF), Gradient Boosting Machines. (GBM) , Light Gradient Boosting Machine (LightGBM) , Extreme Gradient Boosting (XGBoost).

Tabassum, Hridita, et al .[4] presented ”Detecting online recruitment fraud using machine learning.”Achieved accuracy of different prediction models, where LightGBM (95.17) and Gradient Boosting (95.17) give the highest accuracy .Utilized TF-IDF for feature extraction, combined with SMOTE and ADASYN for data balancing, and employed KNN and Random Forest classifiers.

Naude´, Marcel, Kolawole John Adebayo, and Rohan Nanda.[5] ’s paper ”A machine learning approach to detecting fraudulent job types.” which consists of The hybrid model achieved 97.94 accuracy, outperforming state-of-the-art baselines, with features which are identified as most effective.Propose and compute features like organization type, job details, and compensation indicators, followed by a two-step feature selection strategy and building a robust XGBoost-based fraud detection model.

Anita, C. S., et al. [6] in her paper ”Fake job detection and analysis using machine learning and deep learning algorithms.” attained The accuracy of the algorithms can be fine-tuned by cleaning and preprocessing the data in a proper way.variables, and perform feature engineering if needed. Exploratory Data Analysis (EDA): Explore the dataset to gain insights into the distribution of features, correlations, and potential patterns. Random Forest, Support Vector Machines (SVM), Logistic Regression, and neural networks

Swetha, K., et al.[7] ”FAKE JOB DETECTION USING MACHINE LEARNING APPROACH.” paper Highlight the best-performing model(s) and their ability to accurately detect fake job postings, and provided their limitations along with results. Models trained on a specific dataset may not generalize well to new or unseen data. Overfitting to the training data or capturing noise in the dataset can result in poor performance on real-world job postings.

Mahbub, Syed, Eric Pardede, and A. S. M. Kayes.[8] ”Online recruitment fraud detection: A study on contextual features in Australian job industries.” consists of the experimentation conducted with decision tree reveals improved accuracy of 91.64 in experiment B, over 88.89. The results indicate that the Gradient Boosting classifer with empirical rule-set based features, part-of-speech tags and bag-of-words vector achieved the best performance with an F1-score of 0.88.

IV. DATASET

The Employment Scam Aegean Dataset (EMSCAD) is a labelled dataset designed for detecting fraudulent job postings. It contains features extracted from job advertisements, such as job title, location, company name, salary range, job description, and requirements. Each posting is labelled as either legitimate or fraudulent based on manual verification. The dataset exhibits mixed data types, including textual and numerical features, and may have missing values. With a potential class imbalance between legitimate and fraudulent postings, EMSCAD is suitable for fraud detection model development. Preprocessing steps include handling missing data, encoding categorical variables, and potentially performing feature engineering. EMSCAD serves various use cases, including fraud detection model training, data analysis to understand scam patterns, and feature extraction for model enhancement.

V. Proposed Methodology

We have used ensemble model which is done using SVM,random forest algorithm and XGBoost to attain the maximum accuracy in training the dataset and detet the fraudulent job postings more efficiently. The model we are using have the good accuracy when compared with other models which are not ensembled to train the model and the dataset. Our dataset consists of aroumd 18000 job postings. According to the latest creation of online tool REVEAL which detects the fraudness in postings of job also has an higher accuracy but still fail to work efficiently unline other models in instance. See how this model gave the results using our model.

A. FLOW DIAGRAM

The Flow Diagram in this paper first identifies the data and then preprocess it and then split the data into training and testing sets and use those sets for the creation of models and prediction the fraudulence of the transaction and finally use the model to predict that weather the job posting is fraud or not.

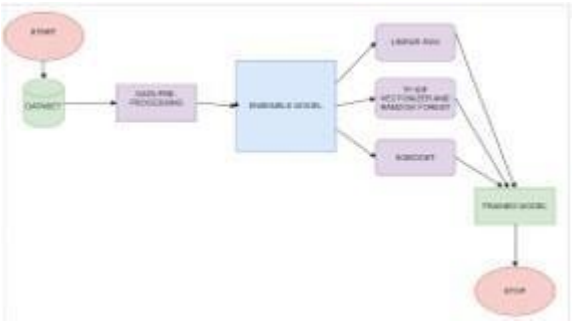

Fig. 1. Architecture and Proposed Model

B. DATA PREPROCESSING AND CLEANING

The method of transforming raw data to useful and efficient format is noted as data preprocessing. One of the main forms of preprocessing is to separate unusable data. It is necessary to make computers understand the natural language as an individual understands. NLP could be a process which will process computers to grasp the language. During this work, we have established a sequential NLP that can be employed in GRU algorithms by using word tokenization, padding data and truncating data techniques. Detecting fraud in job recruitment using machine learning typically involves preprocessing the data to prepare it for model training. Below, I'll outline the steps for data preprocessing, including integration and label encoding:

1. Data Integration:

• Gather all relevant data sources related to job recruitment, such as job postings, resumes, candidate profiles, interview feedback, historical hiring data, etc.

• Merge these datasets into a single cohesive dataset for analysis. Ensure that each data point is properly aligned and identifiable, such as having unique identifiers for candidates, jobs, etc.

• Handle any inconsistencies or missing values during the integration process. This might involve imputation techniques or removing incomplete records if they cannot be salvaged.

2. Label Encoding:

• Identify the target variable for your fraud detection task. In this case, it would be whether a recruitment activity is fraudulent or not (binary classification).

• Convert categorical variables into numerical representations using label encoding. This is necessary because machine learning models typically work with numerical data.

• For example, if you have a categorical variable like "job category" with values like "engineering," "marketing," and "sales," you can encode them as 0, 1, 2, etc.

• Ensure that label encoding is applied consistently across all categorical variables in the dataset.

C. Feature Extraction

Feature extraction in the context of fraud detection in job recruitment involves identifying and selecting relevant information from the integrated dataset that can be used to distinguish between fraudulent and legitimate activities. Here's how you can perform feature extraction:

1. Identify Relevant Features:

• Review the integrated dataset and identify potential features that could be indicative of fraudulent behavior in job recruitment. These features may include:

• Candidate attributes: education level, experience, skills, etc.

• Job posting details: job title, job description, required qualifications, etc.

• Recruitment process information: interview feedback, time taken to hire, number of interviews conducted, etc.

• Historical hiring patterns: frequency of job postings, typical qualifications required, etc.

• Consider both numerical and categorical features that could provide valuable insights into fraudulent activities.

2. Feature Engineering:

• Create new features from existing ones if necessary. For example:

• Calculate derived features such as the ratio of relevant skills to required skills, or the match percentage between candidate skills and job requirements.

• Extract temporal features such as the time of day or day of the week when a job posting was made, or the duration of the recruitment process.

• Feature engineering aims to enhance the predictive power of the model by providing it with more meaningful input features.

3. Dimensionality Reduction:

• If the dataset contains a large number of features, consider using dimensionality reduction techniques to reduce the complexity of the data while retaining important information.

• Techniques like Principal Component Analysis (PCA) or feature selection methods such as Recursive Feature Elimination (RFE) can help identify the most relevant subset of features for the task.

4. Normalization/Scaling:

• Ensure that numerical features are scaled or normalized to a similar range to prevent features with larger scales from dominating the model training process.

• Common scaling techniques include Min-Max scaling or Standardization.

5. Handling Textual Data:

• If your dataset contains textual data such as job descriptions or resumes, you may need to preprocess this data separately.

• Techniques like tokenization, stop-word removal, and stemming/lemmatization can be applied to convert text data into a format suitable for analysis by machine learning mode.

VI. TABLES AND RESULTS

The results of this ensemble model has been depicted with the help of graphical representations and confusion matrix.

A. Confusion Matrix:

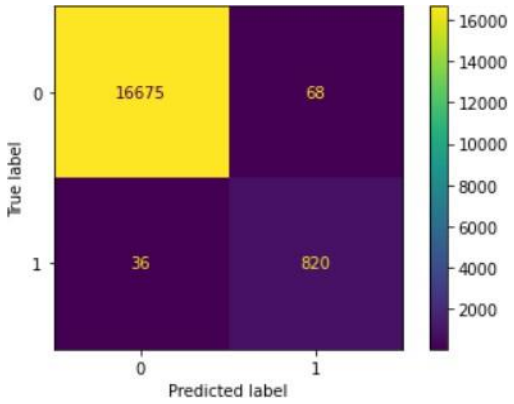

Fig. 2. Confusion Matrix for ensemble Model

Accuracy = TP+TN/TP+FP+FN+TN

Precision: Precision indicates the proportion of positive predictions that were actually correct.

Precision = TP/TP+FP

Recall: Recall quantifies the proportion of actual positive instances correctly identified by the model.

Recall = TP/TP+FN

F1 Score: F1 score is the harmonic mean of precision and recall, providing a balanced measure of model performance.

F1 Score = 2*(Recall * Precision) / (Recall + Precision) (TP= True Positive, TN= True Negative, FP= False Positive, FN= False Negative)]Accuracy: Accuracy measures the proportion of correctly classified instances out of the total predictions made by the model.

Accuracy = TP+TN/TP+FP+FN+TN

Precision: Precision indicates the proportion of positive predictions that were actually correct.

Precision = TP/TP+FP

Recall: Recall quantifies the proportion of actual positive instances correctly identified by the model.

Recall = TP/TP+FN

F1 Score: F1 score is the harmonic mean of precision and recall, providing a balanced measure of model performance.

F1 Score = 2*(Recall * Precision) / (Recall + Precision) (TP= True Positive, TN= True Negative, FP= False Positive, FN= False Negative)

B. Plot check of fakeness in posting:

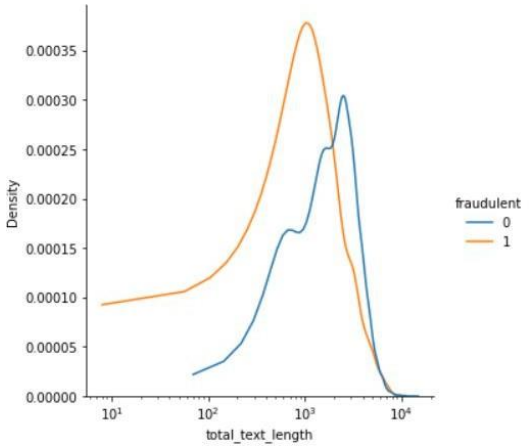

Fig. 3. Textual Contents

It is evidently clear that fake ads generally have shorter textual contents.

C. Accuracies and their Comparison:

	Linear SVM	Random Forest	XGBoost	Ensembled model
Accuracy	97	98	95	98.5
Precision	99	97	99	99
Recall	98	95	97	97
F1-Score	96	98	98	99
Support	4186	4186	4186	4186

Table-1. Accuracy comparision table

VII. CONCLUSION:

To conclude this, according to the recent studies most of the fraud jobs are detected easily with the help of several algorithms and models that are being used for the training of test data and datasets. We performed the data pre processing until we got the most desired accuracy than other existing models. We also studied the impacts and most frequent way of advertising these jobs by knowing frequent words used for doing this so. Challenges that are being faced by job seekers and students can also be reduced to an certain extent with the help of this model and study. This work shows a comparative study on the evaluation of traditional machine learning and deep learning based classifiers. We have found highest classification accuracy for Linear SVM classifier, Random Forest Classifier along with XGBoost classifier by traing and testing on the EMSCAD dataset.

REFERENCES

[1] Prashanth, C., et al. ”Reveal: Online fake job advert detection application using machine learning.” 2022 IEEE Delhi Section Conference (DELCON). IEEE, 2022.

[2] Habiba, Sultana Umme, Md Khairul Islam, and Farzana Tasnim. ”A comparative study on fake job post prediction using different data mining techniques.” 2021 2nd international conference on robotics, electrical and signal processing techniques (ICREST). IEEE, 2021.

[3] Nessa, Iffatun, et al. ”Recruitment scam detection using gated recurrent unit.” 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC). IEEE, 2022.

[4] Tabassum, Hridita, et al. ”Detecting online recruitment fraud using machine learning.” 2021 9th international conference on information and communication technology (ICoICT). IEEE, 2021.

[5] Naude´, Marcel, Kolawole John Adebayo, and Rohan Nanda. ”A machine learning approach to detecting fraudulent job types.” AI SOCIETY 38.2 (2023): 1013-1024.

[6] Anita, C. S., et al. ”Fake job detection and analysis using machine learning and deep learning algorithms.” Revista Geintec-Gestao Inovacao e Tecnologias 11.2 (2021): 642-650

[7] Swetha, K., et al. ”FAKE JOB DETECTION USING MACHINE LEARNING APPROACH.” Journal of Engineering Sciences 14.02 (2023).

[8] Mahbub, Syed, Eric Pardede, and A. S. M. Kayes. ”Online recruitment fraud detection: A study on contextual features in Australian job industries.” IEEE Access 10 (2022): 82776-82787.