# Detection of Fake Job Postings on Online Using Convolutional Neural Network

Md Istakiak Adnan Palash[1]
Arijit Diganto[1]
Osama Nazmul Fatan[1]
Kazi Abu Taher[1]
Md Jaber Al Nahian[1,2]

## Abstract

The present era focuses on every aspect of modern civilization that can be handled online, such as internet banking, teaching, safety, and employment, etc. This advancement in technology makes it easy for scammers to make money very quickly by looting people. Fake job advertisements are among the latest scams. When people apply for these fake jobs, they have topay fees and send their personal information to the fraudsters, which results in a scam and losing money. Therefore, in this paper, we have proposed a novel Convolutional Neural Network(CNN) to identify fake job postings efficiently. A publicly available dataset named EMSCAD was used to validate our proposed model. A comparison was also made between our proposed model and several state-of-the-art machine learning algorithms. In our experiments, we found that our proposed model had a greater accuracy than other machine learning algorithms. In addition, this study conducts a critical comparison of our method with the most recent existing studies.

**Keywords:** Fake Job Posting; COVID-19; Detection; Machine Learning; CNN.

## 1. Introduction

As the corona virus struck the world, the whole world took a very big blow. Everything stood still for a brief period of time, as everyone was instructed to maintain social distancing. It mostly impacted the job industry; many people lost their jobs. While maintaining the social distance, taking recruitment for companies were very difficult and time consuming. So, everything went online. Many people took the liberty of this situation to scam other people. Coronavirus 2019 (COVID-19) pandemic initially occurred in Wuhan, China, in 2019, and

[1] Department of Information and Communication Technology, Bangladesh University of Professionals, Dhaka, Bangladesh.

[2] Corresponding Author: jaber.nahian@bup.edu.bd

quickly spread throughout the world, posing a serious public threat in every sector of China to everywhere (Atangana & Araz, 2020). The COVID-19 epidemic affected many lives around the world and posed an unprecedented threat not only to public health and food systems, but also to the job sectors. According to a report of the World Bank, due to the pandemic around 68 percent of the people of Dhaka and Chittagong in Bangladesh have lost their job in urban areas (Genoni et al., 2020). According to this survey, the rate of job lost in the capital was 76 percent, while it was 59 percent in the port city. A certain group of people is abusing these situations for their own benefit by creating fake jobs to scam helpless people. Scammers are targeting a large number of people who have lost their jobs or seen their wages decrease as a result due to Covid- 19 outbreak. They are mainly scamming people by work-from-home jobs or remote jobs related scams. Many people are looking for legitimate work-from-home jobs because of lay off, pay cuts, quarantines, and stay-at-home orders. Also, some who have a job are also trying to earn more to pay off debt. Some examples of job scams are work-from-home job scams,job placement service scams, government, and postal job scams, nanny, caregiver, virtual personal assistant job scams, and mystery shopper scams (*Job Scams*, 2020). The Ministry of Liberation War Affairs of Bangladesh, had warned the public in November 2020 about a fraudulent job circular circulated by a private entity named ‘Grameen Service Bangladesh Limited’ claiming government approval, which is totally inaccurate and motivated (*Govtwarns on fake job circular by Grameen Service Bangladesh Ltd”, Govt warns on fake job circular by Grameen Service Bangladesh Ltd*, 2022). Also, high-income economic countries faced this problem. During the lockdown, millions of Americans were out of work since the fake jobs have grown (Reinicke, 2020). According to SAFER jobs, a charity that assists flexible economy workers, during the lockdown, the number of fake job circulars increased by 66% in the UK(Burke, 2020). As per the aforementioned problems, early detection of fake job posts is important. Therefore, in this paper, we have proposed a noble CNN-based classification model to detect fake job postings on different platforms. The main contribution of this work is given below:

a) We used a publicly available dataset to validate our model.

b) Some preprocessing steps have been performed before the data is fed into the proposed CNN model.

c) Finally, we have compared our proposed model’s performance with the state-of- the-art models and different machine learning algorithms. The results showed that our proposed model outperformed the existing models.

The following sections make up the remainder of this paper: Section 2 describes the existing works, and Section 3 demonstrates the methodology of the proposed model. Result analysis & discussion is presented in Section 4 and the final section of this paper concludes the whole work.

## 2. Related Works

In recent times, especially during the COVID-19 period, people's dependence on online has increased significantly. On the other hand, people are being deceived in various ways by online fraudsters who snatch important information from them. A group of fraudsters is currently cheating on the common people through fake job advertisements, stealing their personal information and money from them. Many researchers have already worked to identify online fake job advertisements. Different machine learning and deep learning algorithms have been applied for predicting fraud jobs.

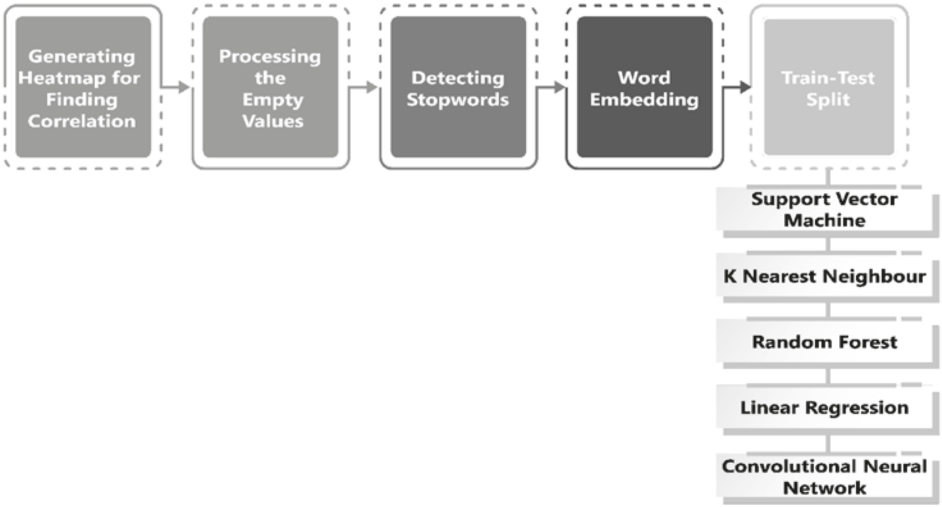


**Figure 1**: A workflow diagram of the proposed work

Vidros et al. (2017a) first introduced this online fraud. They have also introduced the Employment Scam Aegean Dataset (EMSCAD) publicly. As the dataset is imbalanced, they have taken equal samples from both fraudulent and non-fraudulent jobs. They have described the factors related to fraud and gave a brief overview of the features. They split their experiment into two steps. In the first step, they used bow model and in the second step, they transformed all the data from a balanced dataset into a vector of binary features. Lastly, they compared their two different approaches and their performance. The random Forest algorithm performed better in two cases with an accuracy of 91.22% and 90.56% respectively.

Alghamdi et al. (2019) proposed an intelligent model for the detection of fraudulent and non-fraudulent jobs. They have used SVM for feature selection and Random Forest for classification purposes and it gives 97.41% accuracy. They conducted their whole experiment with Weka tools and divided their working procedure into three different stages. They are the pre-processing stage, feature selection stage, and a classification stage. Company profile, has the company logo and Industry are the main three features that are extracted from the dataset.

Dutta and Bandyopadhyay (2020) presented a single classifier and ensemble classifier- based model for the detection of fraudulent jobs. They adjusted the parameters for different algorithms. In MLP they used five hidden layers and for K-NN they set the value of K=5. On the other hand, 500 estimators have been used for ensemble classifiers. For the evaluation of the models, they used different methods like Accuracy, F-1 score, Cohen- Kappa Score, etc. Finally, Decision Tree and Random Forest performed better for single classifier and ensemble classifier.

Habiba et al. (2021) used different data mining techniques for the classification of fake jobs. They followed two different procedures for the classification. First, they used conventional machine learning algorithms for the detection of fake job postings and in the second step, they used deep neural networks to the detection of the fake jobs. They used 10-fold cross-validation for training purposes. Their model achieved 99 % accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.To reduce the overfitting of the model, they used the dropout layer. Lastly, they showed that Deep Neural Network performs much better than conventional algorithms.

Anita et al. (2021) applied machine learning and deep learning algorithms to differentiate real jobs and fake jobs. They also emphasized on data analysis and data cleaning. The data analysis part gives an insight into the data. In the data cleaning part, they removed the columns which have very large null values. Then they removed all the stop words and combined all textual data together in one column for the further process.

On the other hand, Ranparia et al. (2020) trained their model as a Sequential Neural Network and used the Global Vector model (GloVe) algorithm. For analyzing the pattern of the data, they used Natural Language Processing. In Exploratory Data Analysis they minimize computation, remove noisy data, visualize the data, and improve model accuracy. For the testing purpose, they predict 138 actual jobs which they collectfrom LinkedIn. Their model predicts 136 job postings as non-fraudulent which indicates that 99.27% of jobs are correctly detected by the model.

Considering the class imbalance problem in mind, Vo et al. (2021) presented an oversampling technique known as Fake Job Description Detection Using Oversampling Techniques for reducing the imbalance of the dataset. They used the

Employment Scam Aegean (ESA) dataset in their experiment. Firstly, they preprocessed their dataset by removing stop words, and tokenizers. Then they extracted features using a bag of words and term frequency-inverse document frequency which converted the feature into a vector. Before training, they used SVMSMOTE which is an oversampling technique. It is applied to balance the imbalance dataset by generating new synthetic samples. After that, they trained their model. They compared both results which they found before applying over-SVMSMOTE and after applying over-SVMSMOTE where Logistic Regression gives 86.60% accuracy and over-SVMSMOTE-LR gives 92.02% accuracy respectively. In both cases, they found different results where after applying over-SVMSMOTE the model performed better, and the accuracy also increased significantly.

Shibly et al. (2021) used the Microsoft Azure Machine Learning Studio platform for their experiment. They compared the performance of two different types of algorithms. They are the Two-Class Boosted Decision Tree algorithm and Two-Class Decision Forest algorithms and evaluate the performance of the algorithms considering the Accuracy, F1 Score, Recall, and Precision.

Lal et al. (2019) proposed an ensemble learning-based model called ORF Detector. They categorized the features of the dataset into three parts: Linguistic, Contextual, and Meta-data. For building their model they used three base classifiers and three ensemble techniques. They used WEKA for the implementation of the algorithms. The proposed model gives 95.4% accuracy and performed better than the baseline classifiers.

Shree et al. (2021) divided their experiment into five modules. They are dataset collection & preprocessing, data visualization, applying classification algorithm, evaluation, result, and analysis. They also applied the feature selection technique, checked for missing values, and visualized the data. After that, they trained and tested their model using three different algorithms where Random Forest Classification performed better than others which gives 99.8% accuracy.

Mehboob and Malik (2021) proposed a fraud detection framework. They used EMSCAD dataset for their experiment. For their experiment model, they followed data preparation, and feature selection and then applied machine learning algorithms. For better investigation of the impact of the fraud, they categorized them into three different features. They are the organization features, job description features, and compensation features. In their proposed model XGBoost performed better than all other algorithms.

## 3. Methodology

In this section, we have discussed about the working procedure of our research work. The workflow diagram of our research work is shown in Figure 1. The following sub-sections will describe our used dataset, data preprocessing, and proposed CNN architecture.

### 3.1 Data collection

We used a dataset published by the University of Aegean which is called the Employment Scam Aegean Dataset (EMSCAD) (Vidros et al., 2017). This is a publicly available dataset, and this dataset is also available on Kaggle. The dataset contains 18 columns describing different types of features of a specific job These columns tell us about the location of the job, which department we are applying to, what the salary would be, company name, job description, what are the requirements for the job, the employment type, and many more. Some of the columns have numeric values such as the columns 'has company logo', 'has question', 'telecommuting', and 'fraudulent'. The values in these columns are either 1 or 0. The rest of the columns have text values. In the dataset there are two types of jobs posted, types are fraudulent and non-fraudulent. Among the total of 17800 job posted in the dataset, number of non-fraudulent jobs were 17014 and 866 were fraudulent jobs. So, the percentage of the jobs were:

Non-fraudulent jobs: (17014/17800) X 100% = 95%

Fraudulent jobs: (866/17800) X 100% = 5%

### 3.2 Data preprocessing

The dataset is in a CSV format. It has both numeric and text data. We need a tool to read data from the CSV file. For this reason, the best option available is pandas in python. Pandas is a library in the python programming language for manipulating and analyzing data. After that, we looked for NAN values in all the columns. There were no numerical missing values but for the missing text values, we replaced the NAN values with a space(' '). Then we created a variable called text. This contains the concatenated values from the following columns: company profile, description, requirements, and benefits. Whitespace was added before each concatenation point. There is a sentence formed from the above- mentioned columns for each job. We will tokenize each and every sentence.

Now comes the part of processing this data and splitting the data into a test set and train set. Before doing tokenization, removal of stop words was done with the help of NLTK library. Natural Language toolkit or nltk is a library used for preprocessing text data. Stop words in any language are those words which does not add much meaning to a sentence. Even if we discard these words from a sentence, the meaning of the sentence won't change that much.

To split the dataset into train set and test set we will use a library called sklearn. Scikit- learn or sklearn is a machine learning library for python. It has many features. Among all those features we will use train test split. This is a function in model selection for splitting the dataset into two subsets. With the help of this function, we don't have to manually split the dataset. It will split the dataset randomly. We split the dataset into 75% training data and 25% test data.

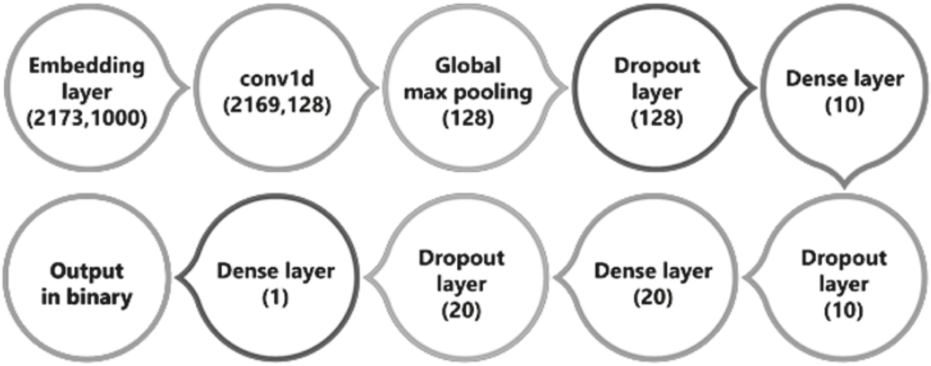


**Figure 2**: Proposed CNN architecture

After splitting the dataset, we will tokenize the sentences both in the training set and the test set. We will use the Keras tokenization function. After fitting the model, we used a method called texts to sequence both on the train set and the test set. It converts each sentence into a sequence of integers. Only words known by the tokenizer are taken into account. All the sentences aren't of the same length. So the model may emphasize the data with greater length. Hence, to avoid this problem we will add zeros in every sequence. This will make every sequence of equal length. This process was done using the pad sequences method. The maxlen parameter in this method was equal to 2173. This was the maximum length we found of a sequence. So now every sequence is of equal length.

### 3.3 Proposed CNN Model

We proposed a novel convolutional neural network to detect fake job postings. The proposed CNN is made up of several layers. The first layer is embedding layer which sizeis (2173,1000). Convolution layer is like applying a filter to the input layer to extract the required information. It has filter size 128 and kernel size 5. It passes the extracted features to the next layer, which is max pooling layer. Since the size of the input was very big, we would need much more time to process the data. So, to reduce the dimensionality we onlyextracted the most prominent features from the previous layer. Hence the output from this layer would be the most notable features of the previous feature map. The next layer is dropout layer, which has the same filter size as the con1d layer. Dropout layers are added to avoid overfitting. The next layer is dense layer, it is a layer of fully connected layer. In this layer, every neuron received input from every neuron of the previous layer. It was done to create diversity in the input type. As in to mix up the extracted features to detectany missing clues. In the dense layers we used relu as the activation function. Another dropout and dense layer were added on, and then the features were passed on to the outputlayer. The binary classification was done on the output layer, determining whether the job was a fake job or not. On the output layer, sigmoid activation function was used, since the output would be either 0 or 1. Batch size of 50 and 10 epochs were run over the dataset. Figure 2 shows the architecture of the proposed CNN model.

## 4. Result Analysis and Discussion

In this experiment, we have implemented both the classical machine learning algorithms and Convolutional Neural Network (CNN) for the classification of fake job postings. First, we have implemented machine learning-based model by using Logistic Regression (LR), Random Forest (RF), K-Nearest Neighbours (KNN), and Support Vector Machine (SVM) algorithms. From these algorithms SVM gives the highest accuracy which is 98.45%. Logistic

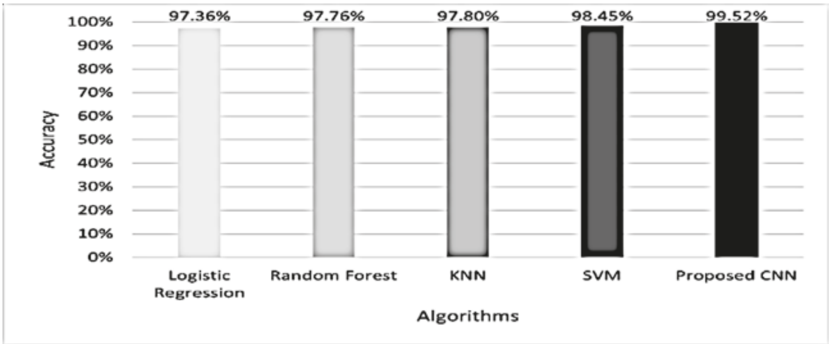
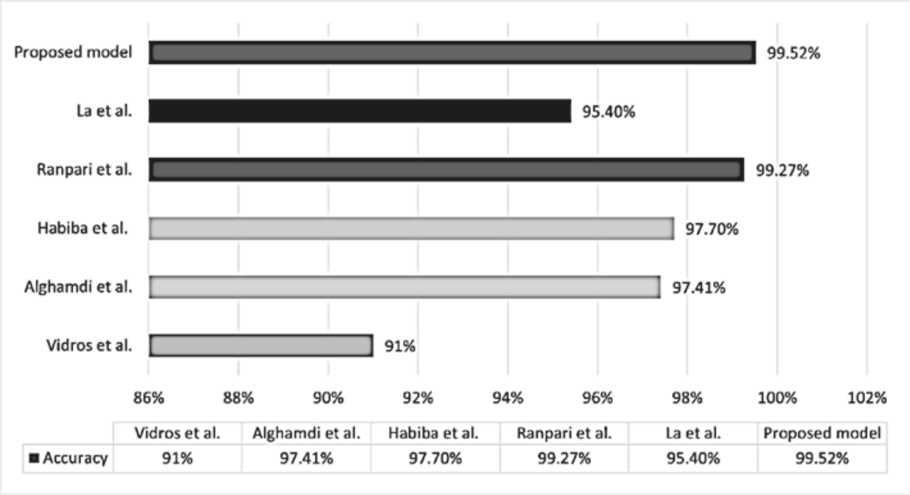


**Figure 3**: Performance comparison of our proposed CNN model with different state of the art machine learning model.Regression,

Random Forest (RF), and K-Nearest Neighbours (KNN) give an accuracy of 97.36%,97.76%, and 97.80% respectively. A performance comparison of above-mentioned machine learning models with the proposed CNN model is given in Figure 3.

After that, we implemented our proposed model using a Convolutional Neural Network (CNN) and then compared the results. Our proposed model performs better than the other machine learning model in this case with an accuracy of 99.52%. It shows an indication that our proposed CNN model can perform well to detect fake job postings. A performance comparison among existing works (Alghamdi et al., 2019; Habiba et al., 2021; Lal et al., 2019; Ranparia et al., 2020; Vidros et al., 2017a) is shown in Figure 4. The bar chart shows that our model performs better than the existing studies.



| | Vidros et al. | Alghamdi et al. | Habiba et al. | Ranpari et al. | La et al. | Proposed model |
|---|---|---|---|---|---|---|
| ■ Accuracy | 91% | 97.41% | 97.70% | 99.27% | 95.40% | 99.52% |

**Figure 4**: Performance comparison of existing works with our proposed CNN model.

## 5. Conclusion

Online recruitment is a big part of the recruitment and hiring process for a company or organization. By doing online recruitment they reach out to much more people from different places rather than staying local. During covid-19 situation, many people lost their jobs. In this post-covid situation, people are desperately looking for jobs. Therefore, some people will try to scam others by posting fake jobs online and taking money from them. We built a simple, yet very accurate fake job posts detecting model. We proposed a Convolutional Neural Network (CNN) to detect if a job post is a fake post or a genuine post and we got an accuracy of 99.52%. We worked with the Employment Scam Aegean Dataset. We had a very limited amount of data to work with. We compared our proposed model with different state of the art machine learning models such as SVM, Logistic Regression, K-Nearest Neighbours, and Random Forest. It is showed that our CNN model provided a better result than all of the above. This will help prevent people from being scammed. Also, the model can be further improved with more data from the recent following times.

## References

Alghamdi, B., Alharby, F., et al. (2019). An intelligent model for online recruitment fraud detection. Journal of Information Security, 10(03), 155.

Anita, C., Nagarajan, P., Sairam, G. A., Ganesh, P., & Deepak kumar, G. (2021). Fake job detection and analysis using machine learning and deep learning algorithms. Revista Geintec-Gestao Inovacao e Tecnologias, 11(2), 642–650.

Atangana, A., & Araz, S. İ. (2020). Mathematical model of covid-19 spread in turkeyand south africa: theory, methods, and applications. Advances in Difference Equations, 2020(1), 1–89.

Burke, J. (2020). Covid-19 scam — fake job ads set to rise, warns expert. https:// www.hrgrapevine.com/content/article/2020-09-04-fake-job-ads-set-to-rise-warns-expert. (Accessed: 2022-06-05)

Dutta, S., & Bandyopadhyay, S. K. (2020). Fake job recruitment detection using machine learning approach. International Journal of Engineering Trends and Technology, 68(4), 48–53.

Genoni, M. E., Khan, A. I., Krishnan, N., Palaniswamy, N., & Raza, W. (2020). Losing livelihoods: The labor market impacts of covid-19 in bangladesh. World Bank.

Govt warns on fake job circular by grameen service bangladesh ltd”, govt warns on fake job circular by grameen service bangladesh ltd. (2022). https://m.theindependentbd.com/post/255536. (Accessed: 2022-06-05)

Habiba, S. U., Islam, M. K., & Tasnim, F. (2021). A comparative study on fake job post prediction using different data mining techniques. In 2021 2nd international conference on robotics, electrical and signal processing techniques (icrest) (pp. 543–546).

Job scams. (2020). https://consumer.ftc.gov/articles/job-scams. (Accessed: 2022-06-05)

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). Orf detector: ensemble learning based online recruitment fraud detection. In 2019 twelfth interna- tional conference on contemporary computing (ic3) (pp. 1–5).

Mehboob, A., & Malik, M. (2021). Smart fraud detection framework for job recruitments. Arabian Journal for Science and Engineering, 46(4), 3067–3078.

Ranparia, D., Kumari, S., & Sahani, A. (2020). Fake job prediction using sequential net- work. In 2020 ieee 15th international conference on industrial and information systems (iciis) (pp. 339–343).

Reinicke, C. (2020). Job scams have increased as covid-19 put millions of americans out of work.here's how to avoid one. https://www.cnbc.com/2020/10/06/job-scams-have increased-during-the-covid-19-crisis-how-to-one.html. (Accessed:2022-06-05)

Shibly, F., Uzzal, S., & Naleer, H. (2021). Performance comparison of two class boosteddecision tree and two class decision forest algorithms in predicting fake job postings.

Shree, R. A., Nirmala, D., Sweatha, S., & Sneha, S. (2021). Ensemble modeling on job scam detection. In Journal of physics: Conference series (Vol. 1916, p. 012167).

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017a). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. Future Internet, 9(1), 6.

Vo, M. T., Vo, A. H., Nguyen, T., Sharma, R., & Le, T. (2021). Dealing with the class imbalance problem in the detection of fake job descriptions. Computers, Materials & Continua, 68(1), 521–535.