Computers, Materials & Continua
DOI:10.32604/cmc.2021.015645
Article

Tech Science Press

Dealing with the Class Imbalance Problem in the Detection of Fake Job Descriptions

Minh Thanh Vo[1], Anh H. Vo[2], Trang Nguyen[3], Rohit Sharma[4] and Tuong Le[2,5,*]

[1]Faculty of Information Technology, Ho Chi Minh City University of Technology (HUTECH), Ho Chi Minh City, Vietnam
[2]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[3]Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
[4]Department of Electronics & Communication Engineering, SRM Institute of Science and Technology, NCR Campus, Ghaziabad, India
[5]Informetrics Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[*]Corresponding Author: Tuong Le. Email: lecungtuong@tdtu.edu.vn
Received: 01 December 2020; Accepted: 02 February 2021

Abstract: In recent years, the detection of fake job descriptions has become increasingly necessary because social networking has changed the way people access burgeoning information in the internet age. Identifying fraud in job descriptions can help jobseekers to avoid many of the risks of job hunting. However, the problem of detecting fake job descriptions comes up against the problem of class imbalance when the number of genuine jobs exceeds the number of fake jobs. This causes a reduction in the predictability and performance of traditional machine learning models. We therefore present an efficient framework that uses an oversampling technique called FJD-OT (Fake Job Description Detection Using Oversampling Techniques) to improve the predictability of detecting fake job descriptions. In the proposed framework, we apply several techniques including the removal of stop words and the use of a tokenizer to preprocess the text data in the first module. We then use a bag of words in combination with the term frequency-inverse document frequency (TF-IDF) approach to extract the features from the text data to create the feature dataset in the second module. Next, our framework applies *k*-fold cross-validation, a commonly used technique to test the effectiveness of machine learning models, that splits the experimental dataset [the Employment Scam Aegean (ESA) dataset in our study] into training and test sets for evaluation. The training set is passed through the third module, an oversampling module in which the SVMSMOTE method is used to balance data before training the classifiers in the last module. The experimental results indicate that the proposed approach significantly improves the predictability

of fake job description detection on the ESA dataset based on several popular performance metrics.

Keywords: Fake job description detection; class imbalance problem; oversampling techniques

1 Introduction

Recently, machine learning and deep learning have become topics of research interest in the context of the Industrial Revolution 4.0. These techniques offer numerous applications in areas such as smart energy management [1–4], social network analysis [5–7], business intelligence [8–12], medical informatics [13–15], computer vision [16–19], software management [20,21] and structural engineering [22]. Natural language processing (NLP) [23] is a subfield of artificial intelligence that is concerned with interactions between computers and human language. The prevalence of false information in many forms of media such as social networking, news blogs, and online newspapers makes it difficult to determine the trustworthiness of content. Machine learning models are therefore needed that are able to provide valuable insights into the reliability of content on the internet. The detection of fake messages [24–26] relies on machine learning models that can automatically identify fake news using artificial intelligence and NLP techniques. In 2019, Bondielli et al. [24] investigated different methods of machine learning for the detection of fake messages and rumors. This study also provided a comprehensive analysis of the techniques mentioned earlier to provide an overview of the problem. Reis et al. [25] explored different types of features extracted from news reports and identified the benefits and importance of these features. In 2019, Gravanis et al. [26] used ensemble machine learning models with optimal feature selection to achieve high accuracy in the recognition of fake messages. This approach used linguistic cues supplemented by word embeddings. In addition, Lingam et al. [27] have conducted research on the detection of fraud in chatbots, which are becoming increasingly popular on social networks. The authors designed a deep Q-network model that included a deep Q-learning (DQL) model based on an updated Q-value function. Experiments conducted on datasets drawn from the Twitter network demonstrated the effectiveness of their proposed model. The problem of detecting fake job descriptions [28] is also an interesting task that can help users in the age of information proliferation to distinguish easily between genuine articles and fraudulent ones. This type of distinction promises to be an important one in relation to various other intelligent systems.

In 2017, Vidros et al. [28] released the Employment Scam Aegean (ESA) dataset that contains 17,880 real job descriptions and 866 scams from the time period 2012 to 2014. The authors also applied several machine learning models to detect fake job descriptions and found that logistic regression was the best machine learning model for solving this task in terms of its accuracy. The class imbalance problem that affects the ESA dataset was the reason for the lower predictability of traditional machine learning models. In addition, the metric of accuracy cannot be used to evaluate machine learning models in imbalanced datasets. In this study, we therefore focused on developing an efficient framework based on oversampling techniques to improve the predictability of traditional classifiers for the problem of the detection of fake job descriptions. The proposed framework, called FJD-OT, initially applies two modules that conduct preprocessing and feature extraction on a text dataset to convert this dataset to a feature vector dataset. Next, FJD-OT utilizes k-fold cross-validation to split the dataset into training and test sets that are used to evaluate the proposed framework. An oversampling module that utilizes the SVMSMOTE technique is then used to balance the training set in terms of the class distribution

at a specific balance ratio. Subsequently, a classification model is trained on this balanced dataset to allow it to identify fake job descriptions from the test set. The results presented in Section 4 demonstrate that the proposed framework achieves the best predictability on the experimental dataset compared with other state-of-the-art methods based on several performance metrics that are commonly used to evaluate machine learning models. These metrics include AUC (area under the receiver operating characteristic curve), balanced accuracy, accuracy, recall, g-mean, sensitivity, and specificity. The most important contributions of this study are as follows: (i) An efficient framework based on oversampling techniques is developed for the problem of detecting fake job descriptions; (ii) the SVMSMOTE technique and logistic regression are applied to improve the predictability of the proposed framework; and (iii) empirical experiments are conducted to verify the effectiveness of FJD-OT on the experimental dataset.

The remainder of this article is structured as follows. An overview of related works is given in Section 2, which focuses on studies of the data imbalance problem, solutions, and applications. The proposed framework for the detection of fake job descriptions is presented in Section 3. Three empirical experiments are described in Section 4. Finally, Section 5 summarizes the results of this study and offers several directions for future work.

2 Related Work

The problem of class imbalance often occurs in datasets in various domains. It arises under conditions where there are two types of classes: a minority and a majority class. The minority class consists of a small number of data samples while the majority class contains a very large number of samples. Traditional classification models have a strong bias toward the majority class in such datasets and this is the reason for the lower predictability. Many approaches have therefore been developed to handle the problem of class imbalance. Fernández et al. [29] grouped these approaches into four categories: algorithm-level approaches, data-level approaches, cost-sensitive learning frameworks, and ensemble-based methods. The first category adapts existing classifiers to distort the learning toward the minority class [30,31] without changing the training data. The second category of data-level approaches modifies the class distribution by resampling the data space [32–35]. There are three sub-categories of data-level approaches: undersampling, oversampling, and hybrid techniques. Undersampling approaches balance the data distribution by removing real data instances from the majority class, whereas oversampling techniques add synthetic instances to the minority class. Hybrid techniques combine undersampling and oversampling techniques. The third approach is a cost-sensitive learning framework that combines data- and algorithm-level approaches. These frameworks add costs to data instances (at the data level) and modify the learning process to accept these costs (at the algorithm level) [36]. In this group of models, the classifier is biased toward the minority class by assuming higher misclassification costs for this class and trying to minimize the total cost errors of both classes. The final group is ensemble-based methods. This group usually consists of a combination of an ensemble learning algorithm and one of the earlier mentioned techniques, in particular data-level and cost-sensitive methods [9]. The new hybrid method usually preprocesses the data before each classifier is trained by combining a data-level approach with an ensemble learning algorithm. In contrast, cost-sensitive ensembles route the cost minimization through the ensemble learning algorithm rather than modifying the base classifier to accept the costs in the learning process.

The problem of class imbalance occurs in most areas of research, including computer vision [37,38], natural language processing [39,40], and time series analysis [41,42]. Ali-Gombe and Elyan utilized a generative adversarial network (GAN) to address the class imbalance problem in

multiple image classification problems. The authors used several fake classes within this framework to ensure fine-grained production and correct classification of the samples in the minority class. Meanwhile, Vo et al. [38] used a combination of deep features and an ensemble approach to deal with the class imbalance problem for smile recognition. Their framework first uses a deep learning model to extract the deep features of the images. Next, these features are used to train the XGBoost classifier (an ensemble classifier). Their results confirmed that the proposed framework achieved better AUC results than other state-of-the-art approaches for smile detection in different scenarios involving imbalanced data. In NLP, the recognition of named entities in free texts suffers greatly from the problem of class imbalance because many samples of any free text do not belong to a particular entity. Akkasi et al. [39] therefore developed an undersampling approach for sequenced data that preserved the existing correlations between the sequenced samples making up sentences. The experimental results confirmed that this approach improved the performance of the classifiers. Recently, Barushka et al. [40] also addressed the problem of class imbalance in the detection of spam by proposing a novel cost-sensitive approach. The proposed framework was composed of two phases: first, a multi-objective evolutionary feature selection method was used to minimize both the cost of misclassification of the proposed model and the number of attributes required for spam filtering; next, cost-sensitive ensemble learning techniques were applied with regularized deep neural networks for basic learning. Experiments were performed with two benchmark datasets and the results indicated that the proposed framework outperformed other state-of-the-art models for spam filtering in social networks, including random forest, naïve Bayes, and support vector machine (SVM). Yan et al. [41] introduced a quality measure based on optimizing shapelets to address the class imbalance problem in time series classification. These authors also utilized two oversampling approaches based on shapelet features to re-balance binary and multi-class time series datasets. Their experimental results confirmed that the proposed approach achieved more competitive results in terms of statistical significance than the most advanced methods. In another approach, He et al. [42] developed an ensemble of shapelet-based classifiers for the problem of inter-class and intra-class imbalanced multivariate time series classification. In this method, a cluster-based shapelet selection method was developed to identify an optimal set of stable and robust shapelets to deal with intra-class imbalance classification. Their experimental results suggested that the proposed approach could be used to effectively predict multivariate time series data with imbalances between and within classes at an early stage.

3 FJD-OT: Fake Job Description Detection Using Oversampling Techniques

3.1 Experimental Dataset

In 2017, the Laboratory of Information and Communication Systems Security (ICSS) at the University of the Aegean, Greece, released the ESA dataset [28] for the detection of fake job descriptions. This dataset contains 17,014 real job descriptions and 866 fake ones, published within the time period 2012 to 2014. The dataset consists of both textual information and meta-information related to the jobs. The dataset can be used to train machine learning models on the job descriptions to predict fraud in the fields of human resources and recruiting services.

The class distribution and class count for the ESA dataset is shown in Fig. 1. This dataset has a class imbalance problem as only 5% of posts are fake while 95% are real. In mathematical terms, let χ be a dataset consisting of two classes c_{min} and c_{maj}, which represent the real and

fake job description (JD) classes, respectively. The balancing ratio of χ, which is denoted by br_c, is defined by the following formula:

$$br_c = \frac{|c_{min}|}{|c_{maj}|}, \tag{1}$$

where $|c_{min}|$ and $|c_{maj}|$ are the number of samples in the fake and real JD classes, respectively. The balancing ratio of the class imbalance dataset is small; the smaller this ratio, the more difficult machine learning will be. This study therefore presents an efficient framework based on an oversampling technique named SVMSMOTE to balance the class distribution, and to improve the performance on the task of detection of fake job descriptions.

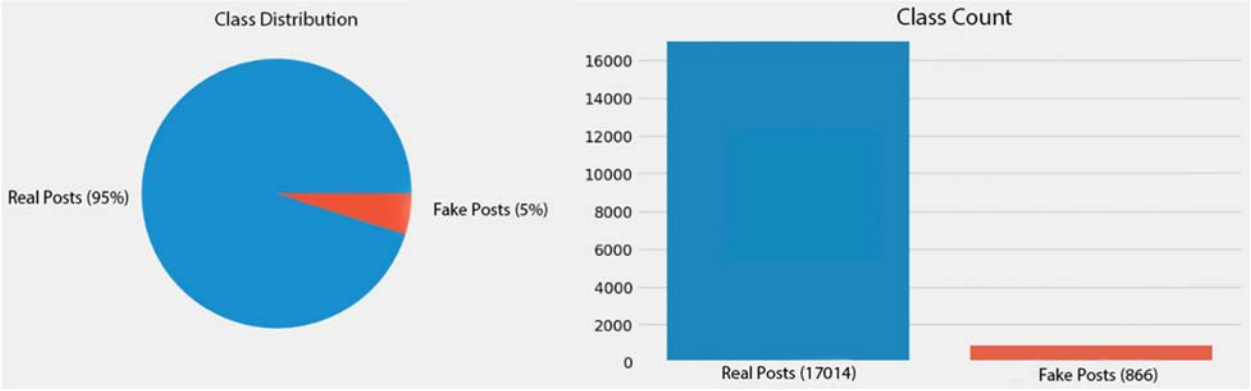

Figure 1: Class count and distribution of the ESA dataset

3.2 Proposed Framework

In this section, we introduce the overall architecture of the proposed framework, FJD-OT (Fig. 2). The framework consists of four main modules: preprocessing, feature extraction, oversampling, and classification. These are described in detail in the rest of this section.

The first module utilizes several techniques such as the removal of stop words and a tokenizer to preprocess the text data. Meanwhile, the bag of words and term frequency-inverse document frequency (TF-IDF) approaches are used to convert the text dataset to a feature vector in the second module. The bag of words identifies how often a term occurs in a document (or job description), which allows a computer to compare these documents and evaluate their similarities for several applications such as document retrieval, document classification, and topic modeling. Next, TF-IDF is applied to calculate the weight of each term in the document. Certain words (such as "and" or "the") are common to all documents and must be systematically discounted. In addition, the more documents in which a word appears, the less valuable this word is as a signal to differentiate any given document. The TF-IDF for each term is determined as follows:

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right), \tag{2}$$

where $tf_{i,j}$ is the number of occurrences of term i in job description j, df_i is the number of job descriptions containing term i, and N is the total number of job descriptions. We apply this technique to create feature vectors of the experimental dataset and feed them to a machine learning model to detect fake job descriptions.

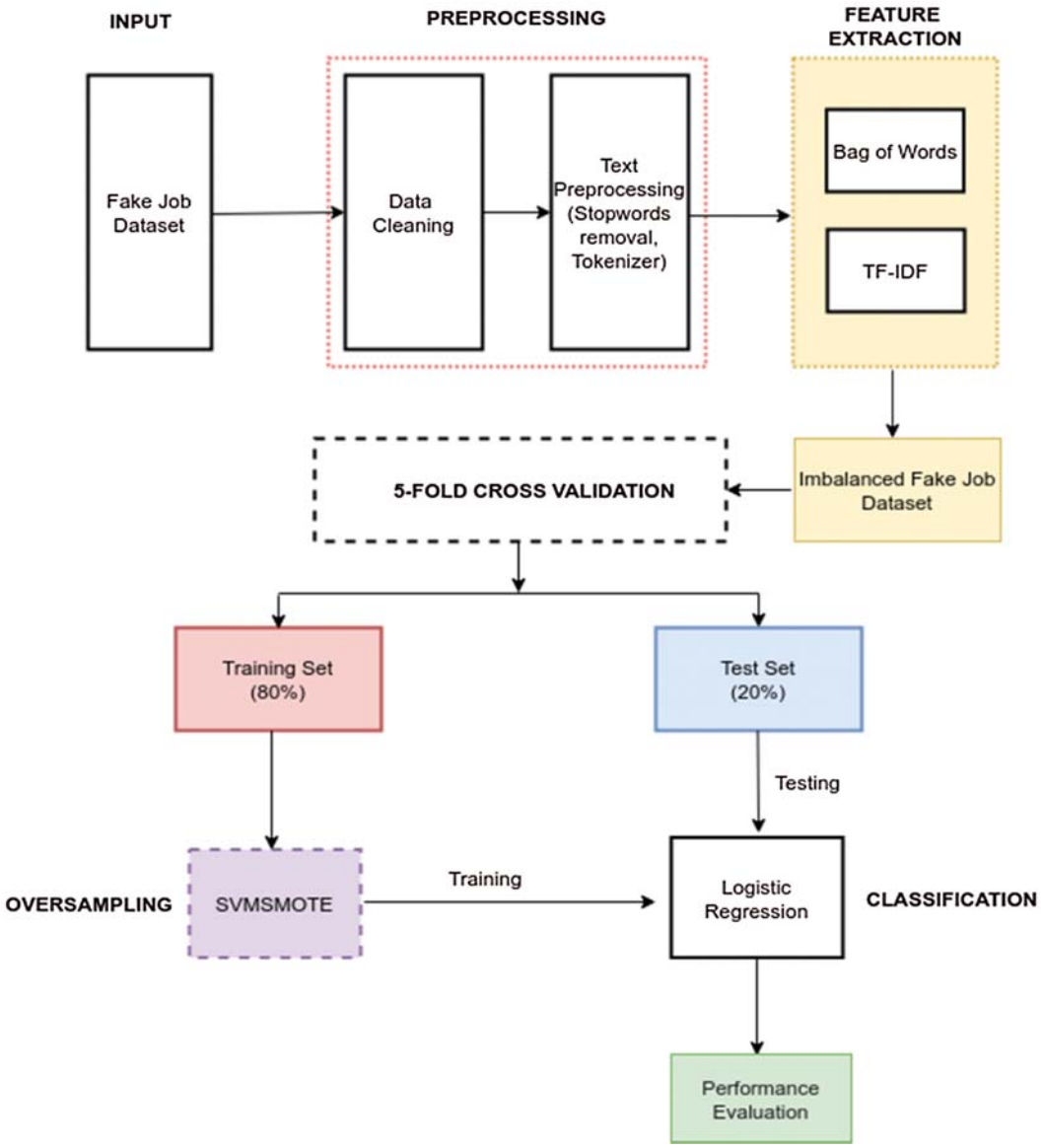

Figure 2: The proposed FJD-OT framework for the detection of fake job descriptions

Next, FJD-OT applies an oversampling technique called SVMSMOTE [35], a variant of the SMOTE algorithm, to balance the training set. Here, we first describe the SMOTE technique [32], which is used to generate new synthetic samples for the minority class to balance the class distribution, to explain in full how SVMSMOTE works. The new samples are based on the feature space similarities among the original samples. In other words, SMOTE considers the k-nearest neighbors ($\mathcal{K}_{\mathbf{x}_i}$) of sample $\mathbf{x}_i$ in the minority class. The algorithm subsequently randomly selects an element $\hat{\mathbf{x}}_i$ belonging both to $\mathcal{K}_{x_i}$ and the minority class. The feature vector of the new synthetic sample ($\mathbf{x}_{new}$) is calculated using the following formula:

$$\mathbf{x}_{new} = \mathbf{x}_i + (\hat{\mathbf{x}}_i - \mathbf{x}_i) \times \delta, \tag{3}$$

where δ is a random value in the range [0, 1]. Eq. (3)

In contrast, SVMSMOTE generates a number of new synthetic samples for the minority class near the boundary with the SVM to facilitate class differentiation. In this approach, the borderline between the two classes is first identified by an SVM trained on the original training set. The new synthetic sample is randomly generated along the line segment joining $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ using interpolation (Fig. 3a) and extrapolation (Fig. 3b) techniques, depending on the density of the majority class instances around $\mathbf{x}_i$. Interpolation creates new data points between two samples, whereas extrapolation creates new data points outside of the two samples. If samples from the majority class account for less than half of its nearest neighbors, the new synthetic sample is created with the extrapolation technique (Fig. 3b) to expand the area of the minority class toward the majority class using the following formula:

$$\mathbf{x}_{new} = \mathbf{x}_i + (\mathbf{x}_i + \hat{\mathbf{x}}_i)\mathrm{x} \times \delta. \tag{4}$$

Otherwise, $\mathbf{x}_{new}$ is created with interpolation (Fig. 3a) in a similar way to SMOTE and as shown in Eq. (3).

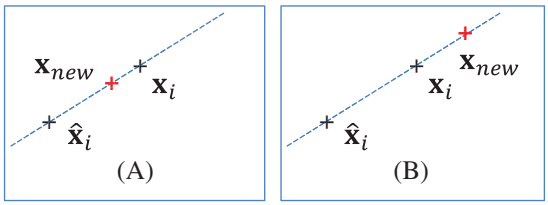

Figure 3: (a) Interpolation and (b) extrapolation

After SVMSMOTE has balanced the training data, these data are subsequently used to train the logistic regression following the suggestion made by Vidros et al. [28] for the detection of fake job descriptions. Logistic regression uses a logistic function to measure the relationship between the categorical dependent variable (target value) and one or more independent variables (input values) by estimating probabilities. The logistic function is the cumulative distribution function of the logistic distribution, which maps any real value in the range $-\infty$ to ∞ to another value in the range zero to one as follows:

$$\phi(z) = \frac{1}{1+e^{-z}}, \tag{5}$$

where $\phi(z)$ is the output between zero and one (probability estimate) and z is the input value. A graph of the sigmoid function is shown in Fig. 4.

Notice that $\phi(z)$ tends toward one when $z \to \infty$, and toward zero when $z \to -\infty$. In addition, $\phi(z)$ is always bounded by [0, 1]. The algorithm needs to make sure that $p(y=1)+p(y=0)$ is equal to one to create a probability. This requirement can be solved as follows:

$$P(y=1) = \frac{1}{1+e^{-z}}, \tag{6}$$

$$P(y=0) = \frac{e^{-z}}{1+e^{-z}}. \tag{7}$$

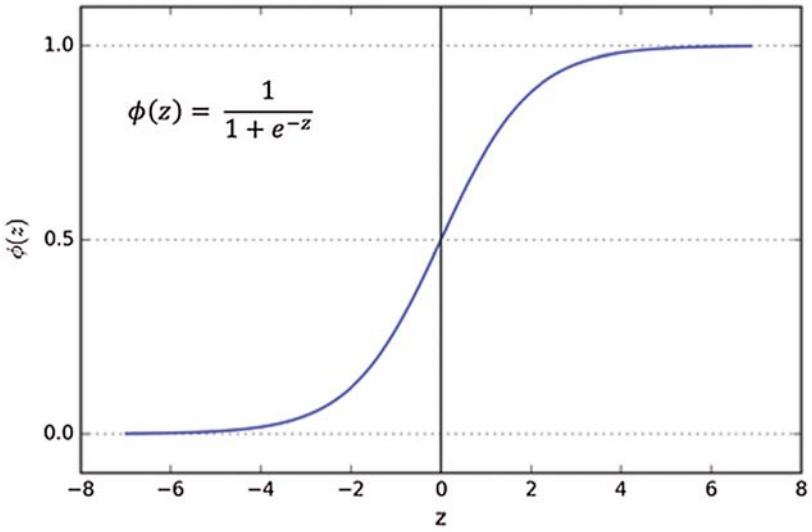

Figure 4: Graph of a sigmoid function

Next, we compute the probability $P(y=1 \mid x)$ for a test instance x. Notice that this algorithm uses 0.5 as the decision boundary as follows:

$$\hat{y} = \begin{cases} 1, & if \;\; P(y=1 \mid x) > 0.5 \\ 0, & otherwise \end{cases}. \tag{8}$$

4 Performance Evaluation

4.1 Experimental Setting

All of the experiments in this study were implemented in version 3.7 of the Python programming language. The operating system was Ubuntu 16.04 LTS, with an Intel Core i7-6820HQ (2.7 GHz × 8 cores), 16 GB of RAM and GeForce GTX 940MX graphics processor. The imbalanced-learn [43] and scikit-learn [44] packages were used to perform the experiments.

The proposed approach was evaluated in several ways. Experiments were conducted to analyze the impact of the balancing ratio on the FJD-OT framework (Section 4.2), to compare the FJD-OT framework with other state-of-the-art approaches (Section 4.3), and to conduct a type error analysis (Section 4.4) for the detection of fake job descriptions. Several common metrics were used to evaluate the classification problem, including the balanced accuracy (balanced ACC), accuracy (ACC), geometric mean (Gmean), recall, sensitivity, and specificity. These metrics were computed as follows:

$$Balanced \;\; ACC = \frac{TP+TN}{2}, \tag{9}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}, \tag{10}$$

$$Sensitivity = \frac{TP}{TP+FN}, \tag{11}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{12}$$

$$G\text{-}mean = \sqrt{Sensitivity \times Specificity}, \tag{13}$$

$$Recall = \frac{TP}{TP + FN}. \tag{14}$$

In addition, the area under the curve (AUC) [45] of the receiver operating characteristic (ROC) curve needs to be considered in the class imbalance problem. The ROC curve can be used to visualize the trade-off between the true positive rate $TP_{rate} = \frac{TP}{TP+FN}$ and the false positive rate $FP_{rate} = \frac{FP}{FP+TN}$. This curve has the property that no machine learning model can increase the true positive rate without an increment in the false positive rate. Each point on this curve indicates the performance of a model for a specific distribution. AUC is typically used as a performance metric for the evaluation of classification performance and especially for the class imbalance problem. The higher the value of AUC, the better the performance of the model.

A detailed description of all the experimental methods used in this study is shown in Tab. 1. Vidros et al. [28] found that logistic regression is the best machine learning model for the detection of fake job descriptions on the ESA dataset. We therefore implement this approach and this is shown as Method 1 in Tab. 1. Methods 2–5 are combinations of logistic regression with the four oversampling techniques of SMOTE [32], ADASYN [33], borderline SMOTE [34], and SVMSMOTE [35]. They are denoted by Over-SMOTE-LR, Over-ADASYN-LR, Over-BSMOTE-LR, and Over-SVMSMOTE-LR, respectively. The five remaining methods (6–10) are combinations of logistic regression with five undersampling techniques: repeated edited nearest neighbors, Tomek links, instance hardness threshold, near miss, and the neighborhood cleaning rule. They are denoted by Under-RENN-LR, Under-TL-LR, Under-IHT-LR, Under-NM-LR, and Under-NCR-LR, respectively.

Table 1: Descriptions of experimental methods

No.	Type of resampling	Resampling technique	Abbreviation
1	None	None	LR [28]
2	Oversampling	SMOTE [32]	Over-SMOTE-LR
3		ADASYN [33]	Over-ADASYN-LR
4		Borderline SMOTE [34]	Over-BSMOTE-LR
5		SVMSMOTE [35]	Over-SVMSMOTE-LR
6	Undersampling	Repeated edited nearest neighbors	Under-RENN-LR
7		Tomek links	Under-TL-LR
8		Instance hardness threshold	Under-IHT-LR
9		Near miss	Under-NM-LR
10		Neighborhood cleaning rule	Under-NCR-LR

4.2 Impact of the Balancing Ratio on the Proposed Framework

In the first experiment, we change the target balancing ratio from 0.1 to 1 with a step size of 0.01 to balance the experimental dataset at many different levels. This experiment is conducted to evaluate the impact of the balancing ratio on the various performance metrics used in this

study and to select the value that produces the best performance results for the experimental dataset (Fig. 5).

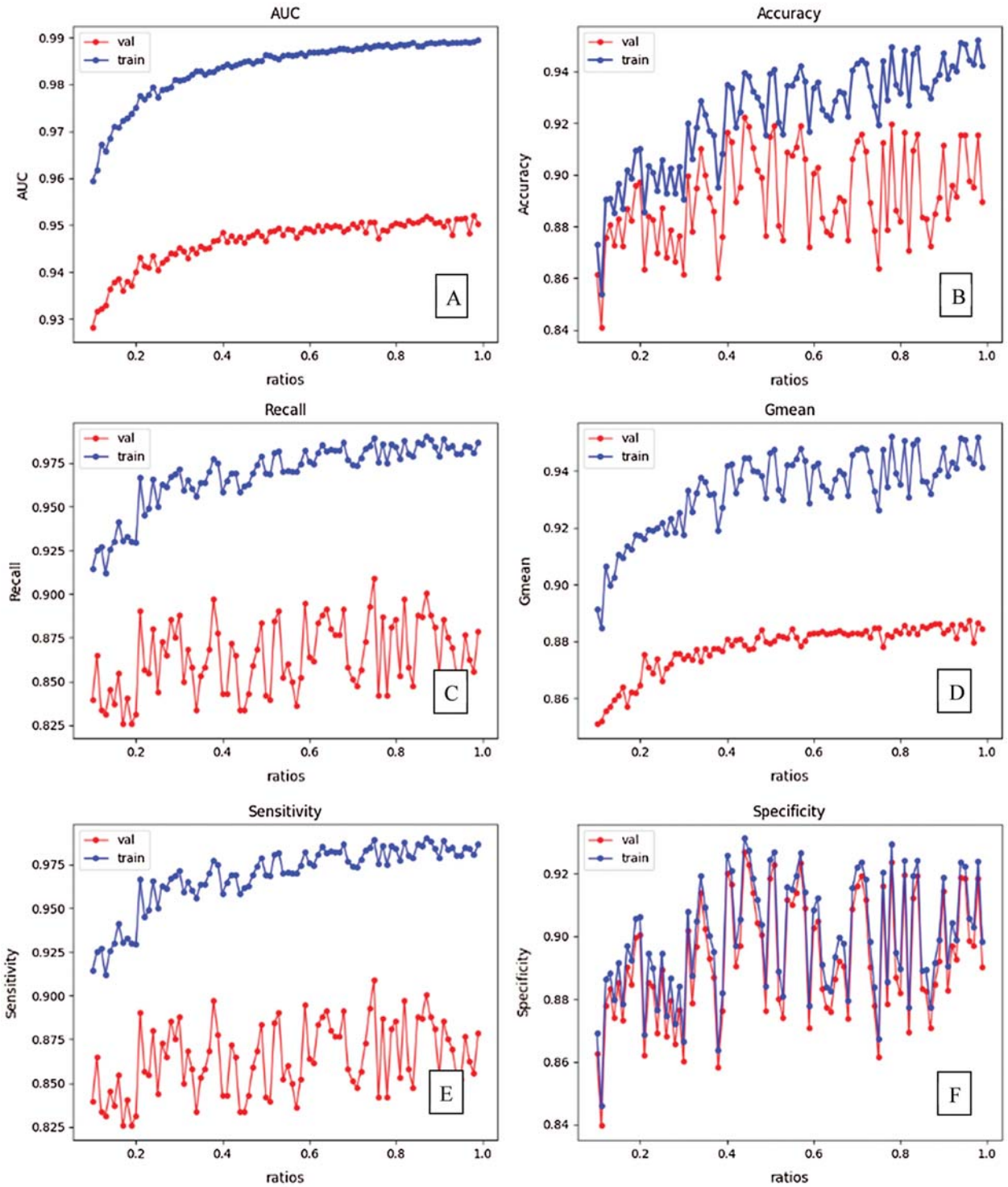

Figure 5: The impact of balancing ratio on several performance metrics: (A) AUC, (B) Accuracy, (C) Recall, (D) Gmean, (E) Sensitivity, and (F) Specificity

The performance of FJD-OT shows a general increase in the AUC score for the training and test sets as the balancing ratio is increased (Fig. 5a). Similarly, we find that with a balancing ratio of approximately 0.78, the performance of FJD-OT reaches a peak for all performance metrics as shown in Fig. 5. The performance results begin to decrease for balancing ratios greater than 0.78. We therefore selected this balancing ratio as the optimal parameter when conducting the remaining experiments.

4.3 Comparison with State-of-the-Art Approaches

This section presents a comparison of several experimental methods based on the performance metrics introduced in Section 4.1. The ROC curves for 10 experimental methods are shown in Fig. 6. The results indicate that Over-SMOTE-LR, Over-ADASYN-LR, and Over-SVMSMOTE-LR obtain the best AUC value of 0.96, which is an impressive result. The other oversampling-based approach, Over-BSMOTE-LR, yielded an AUC value of 0.95, while the LR approach achieved only 0.93. The undersampling-based methods Under-RENN-LR, Under-TL-LR, Under-IHT-LR, Under-NM-LR, and Under-NCR-LR achieved AUC values of less than 0.93. Hence, SMOTE, ADASYN, and SVMSMOTE are the three best oversampling methods for the experimental dataset in terms of AUC.

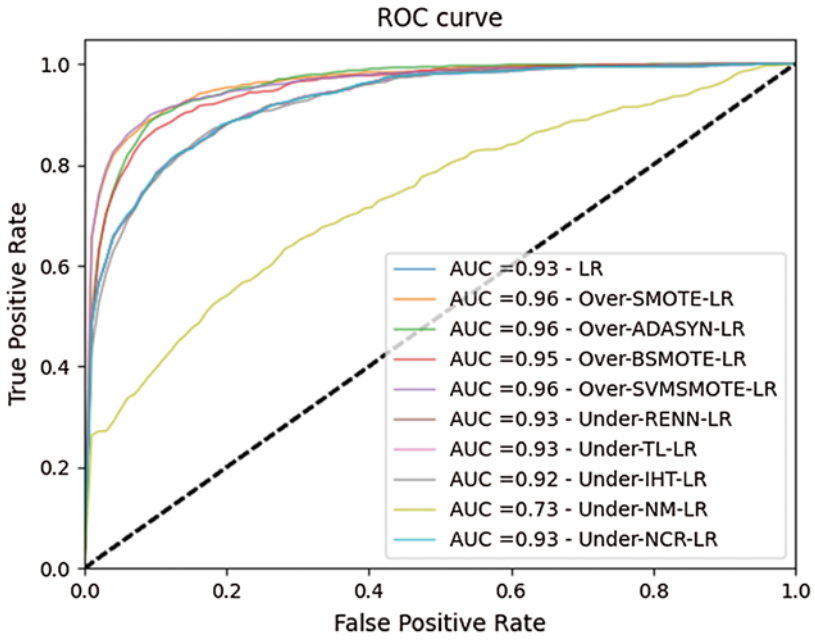

Figure 6: ROC curves for each experimental method for the ESA dataset

The performance results for all the experimental methods are shown in Tab. 2. Over-SVMSMOTE-LR outperforms the LR method without the need to apply resampling techniques to all metrics. Over-SVMSMOTE-LR is also the best of the oversampling techniques, including Over-SMOTE-LR, Over-ADASYN-LR, Over-BSMOTE-LR, and Over-SVMSMOTE-LR, in terms of balanced ACC and Gmean as shown in Tab. 2. Over-SVMSMOTE-LR achieved similar performance on the rest of the metrics as compared to other oversampling-based techniques. Over-SVMSMOTE-LR yields a similar value to Over-SMOTE-LR and Over-ADASYN-LR (0.96) for

AUC, the same value as Over-SMOTE-LR (0.92) for ACC, the same result as Over-ADASYN-LR (0.89) for recall and sensitivity, and the same value as Over-SMOTE-LR (0.92) for specificity (Tab. 2). The results for the undersampling techniques in Tab. 2 indicate that their influence on the performance is not more than the first method. Overall, our empirical analysis also demonstrates that the oversampling techniques help to improve the effectiveness of the performance for the problem of detecting fake job descriptions. The Over-SVMSMOTE-LR method is much more stable than the other methods based on oversampling techniques, including Over-SMOTE-LR, Over-ADASYN-LR, and Over-BSMOTE-LR.

Table 2: Experimental results for each experimental method based on several performance metrics for the ESA dataset

No.	Approach	Balanced ACC	ACC	Recall	G-mean	Sensitivity	Specificity	AUC
1	LR	0.85 ± 0.006	0.87 ± 0.029	0.83 ± 0.039	0.85 ± 0.006	0.83 ± 0.039	0.87 ± 0.033	0.93 ± 0.01
2	Over-SMOTE-LR	0.90 ± 0.010	**0.92 ± 0.026**	0.88 ± 0.040	0.90 ± 0.009	0.88 ± 0.040	**0.92 ± 0.029**	**0.96 ± 0.01**
3	Over-ADASYN-LR	0.90 ± 0.09	0.89 ± 0.016	**0.89 ± 0.022**	0.89 ± 0.008	**0.89 ± 0.022**	0.89 ± 0.018	**0.96 ± 0.00**
4	Over-BSMOTE-LR	0.89 ± 0.010	0.90 ± 0.020	0.87 ± 0.030	0.88 ± 0.011	0.87 ± 0.030	0.90 ± 0.022	0.95 ± 0.01
5	Over-SVMSMOTE-LR	**0.91 ± 0.09**	**0.92 ± 0.019**	**0.89 ± 0.034**	**0.91 ± 0.010**	**0.89 ± 0.034**	**0.92 ± 0.021**	**0.96 ± 0.01**
6	Under-RENN-LR	0.85 ± 0.06	0.86 ± 0.029	0.84 ± 0.035	0.85 ± 0.006	0.84 ± 0.035	0.86 ± 0.030	0.93 ± 0.01
7	Under-TL-LR	0.85 ± 0.06	0.86 ± 0.030	0.84 ± 0.035	0.85 ± 0.006	0.84 ± 0.034	0.86 ± 0.033	0.93 ± 0.01
8	Under-IHT-LR	0.85 ± 0.03	0.86 ± 0.032	0.84 ± 0.040	0.85 ± 0.004	0.84 ± 0.040	0.86 ± 0.035	0.92 ± 0.01
9	Under-NM-LR	0.68 ± 0.012	0.73 ± 0.045	0.62 ± 0.045	0.67 ± 0.011	0.62 ± 0.045	0.73 ± 0.049	0.73 ± 0.02
10	Under-NCR-LR	0.85 ± 0.06	0.88 ± 0.027	0.83 ± 0.038	0.85 ± 0.007	0.83 ± 0.038	0.87 ± 0.030	0.93 ± 0.01

4.4 Type Error Analysis

We performed a type error analysis to investigate the impact of Over-SVMSMOTE-LR on the reliability of the predictive model compared with LR [27] due to the evidence we described in Section 4.3. The issue of type II errors is one of the challenges that arise in imbalanced data, where fake job descriptions are predicted to be real job descriptions. Type I errors occur when samples in the majority class (Real JD) are misclassified into the minority class (Fake JD). LR gives a type I error rate of 13.25% (Fig. 7), as it predicts 450 wrong samples over 3,403 real job description samples (major class). The type II error rate obtained by LR was 16.76%, as it predicts 29 wrong samples versus 173 fake-post samples (minor class). The type I error rate is only 7.85% (267 wrong samples) for the Over-SVMSMOTE-LR method. Clearly, it is better than LR with 163 samples (5.4%), which are predicted successful samples for the Real-Post class. Likewise, Over-SVMSMOTE-LR has a reduced rate of type II errors compared with the LR method, with only 19 wrong samples (11%), and it improves 5.76% rather than the type II error value (16.76%) of the non-sampling method (LR). Over-SVMSMOTE-LR therefore significantly improves the performance of the prediction model in terms of both type I and type II errors.

Confusion matrix of LR

		Predicted	
		Real-Post	Fake-Post
Actual	Real-Post	2952 (86.75%)	450 (13.25%)
	Fake-Post	29 (16.76%)	144 (83.24%)

Confusion matrix of Over-SVMSMOTE-LR

		Predicted	
		Real-Post	Fake-Post
Actual	Real-Post	3136 (92.15%)	267 (7.85%)
	Fake-Post	19 (11%)	154 (89%)

Figure 7: Confusion matrices for LR and Over-SVMSMOTE-LR for the ESA dataset. (a) Confusion matrix of LR (b) confusion matrix of over-SVMSMOTE-LR

5 Conclusions and Future Work

This study presents an efficient framework based on an oversampling technique called FJD-OT to improve predictability in the detection of forged job descriptions. The bag of words, TF-IDF methods, and several preprocessing techniques are applied to preprocess and extract features from textual data to create a feature vector dataset. We utilized k-fold cross validation to split the dataset into training and test sets to evaluate our method experimentally. The training set was then passed through an oversampling module in FJD-OT, which is based on the SVMSMOTE method, to balance the data before training the machine learning models. Three empirical experiments were conducted to investigate the impact of the balancing ratio on the FJD-OT framework. This allowed us to compare our method with other state-of-the-art approaches and to conduct a type error analysis to confirm the effectiveness of the proposed approach. The results confirm that our proposed framework based on the oversampling method SVMSMOTE significantly outperforms the other experimental approaches for the ESA dataset when several popular performance metrics are considered.

In the future, we will develop a new oversampling technique that is more suited to the ESA dataset. This will help to improve the predictive ability of our approach for the detection of fake job descriptions. Several feature extraction techniques will be studied to enhance the performance of our technique for this task. In addition, further research will be needed for the online learning model due to the rapid growth in data.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] T. Le, M. T. Vo, T. Kieu, E. Hwang, S. Rho *et al.,* "Multiple electric energy consumption forecasting using a cluster-based strategy for transfer learning in smart building," *Sensors*, vol. 20, no. 9, pp. 2668, 2020.

[2] Z. A. Khan, T. Hussain, A. Ullah, S. Rho, M. Y. Lee *et al.,* "Towards efficient electricity forecasting in residential and commercial buildings: A novel hybrid CNN with a LSTM-AE based framework," *Sensors*, vol. 20, no. 5, pp. 1399, 2020.

[3] D. Sembroiz-Ausejo, D. Careglio, S. Ricciardi and U. Fiore, "Planning and operational energy optimization solutions for smart buildings," *Information Sciences*, vol. 476, no. 10, pp. 439–452, 2019.

[4] T. Le, M. T. Vo, B. Vo, E. Hwang, S. Rho *et al.,* "Improving electric energy consumption prediction using CNN and Bi-LSTM," *Applied Sciences*, vol. 9, no. 20, pp. 4237, 2019.

[5] R. Ren, M. Tang and H. Liao, "Managing minority opinions in micro-grid planning by a social network analysis-based large scale group decision making method with hesitant fuzzy linguistic information," *Knowledge-Based Systems*, vol. 189, no. 19, pp. 105060, 2020.

[6] T. Vo, R. Sharma, R. Kumar, H. S. Le and B. T. Pham, "Crime rate detection using social media of different crime locations and twitter part-of-speech tagger with brown clustering," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 4, pp. 4287–4299, 2020.

[7] G. Lingam, R. R. Rout and D. Somayajulu, "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks," *Applied Intelligence*, vol. 49, no. 11, pp. 3947–3964, 2019.

[8] V. L. Hoang, H. S. Le, M. Khari, K. Arora, S. Chopra *et al.,* "A new approach for construction of geo-demographic segmentation model and prediction analysis," *Computational Intelligence and Neuroscience*, vol. 2019, no. 1, pp. 1–10, 2019.

[9] T. Le, B. Vo, H. Fujita, N. T. Nguyen and S. W. Baik, "A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting," *Information Sciences*, vol. 494, no. 1, pp. 294–310, 2019.

[10] T. Le, M. Y. Lee, J. R. Park and S. W. Baik, "Oversampling techniques for bankruptcy prediction: Novel features from a transaction dataset," *Symmetry*, vol. 10, no. 4, pp. 79, 2018.

[11] B. Tanuwijaya, G. Selvachandran, H. S. Le, M. Abdel-Basset, X. H. Huynh *et al.,* "A novel single valued neutrosophic hesitant fuzzy time series model: Applications in Indonesian and Argentinian stock index forecasting," *IEEE Access*, vol. 8, pp. 60126–60141, 2020.

[12] A. H. Vo, T. Nguyen and T. Le, "Brent oil price prediction using Bi-LSTM network," *Intelligent Automation & Soft Computing*, vol. 26, no. 6, pp. 1307–1317, 2020.

[13] R. Sharma and N. Hooda, "Optimized ensemble machine learning framework for high dimensional imbalanced bio assays," *Revue d'Intelligence Artificielle*, vol. 33, no. 5, pp. 387–392, 2019.

[14] N. Shanavas, H. Wang, Z. Lin and G. I. Hawe, "Ontology-based enriched concept graphs for medical document classification," *Information Sciences*, vol. 525, no. 1, pp. 172–181, 2020.

[15] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan *et al.,* "Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals," *Applied Intelligence*, vol. 49, no. 1, pp. 16–27, 2019.

[16] H. A. Vo, H. S. Le, M. T. Vo and T. Le, "A novel framework for trash classification using deep transfer learning," *IEEE Access*, vol. 7, no. 1, pp. 178631–178639, 2019.

[17] M. T. Vo, T. Nguyen and T. Le, "Robust head pose estimation using extreme gradient boosting machine on stacked autoencoders neural network," *IEEE Access*, vol. 8, no. 1, pp. 3687–3694, 2020.

[18] T. N. Pham, J. W. Lee, G. R. Kwon and C. S. Park, "Efficient image splicing detection algorithm based on markov features," *Multimedia Tools and Applications*, vol. 78, no. 9, pp. 12405–12419, 2019.

[19] V. H. Pham, H. K. Jo and V. D. Hoang, "Scalable local features and hybrid classifiers for improving action recognition," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 4, pp. 3357–3372, 2019.

[20] S. Jha, R. Kumar, H. S. Le, M. Abdel-Basset, I. Priyadarshini *et al.,* "Deep learning approach for software maintainability metrics prediction," *IEEE Access*, vol. 7, pp. 61840–61855, 2019.

[21] H. Wei, C. Hu, S. Chen, Y. Xue and Q. Zhang, "Establishing a software defect prediction model via effective dimension reduction," *Information Sciences*, vol. 477, no. 3, pp. 399–409, 2019.

[22] Q. H. Doan, T. Le and D. K. Thai, "Optimization strategies of neural networks for impact damage classification of RC panels in a small dataset," *Applied Soft Computing*, vol. 102, pp. 107100, 2021.

[23] P. Tran, D. Dinh, T. Le and L. Nguyen, "Linguistic-relationships-based approach for improving word alignment," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 1, pp. 1–16, 2017.

[24] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Information Sciences*, vol. 497, no. 4, pp. 38–55, 2019.

[25] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, F. Benevenuto *et al.,* "Supervised learning for fake news detection," *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76–81, 2019.

[26] G. Gravanis, A. Vakali, K. I. Diamantaras and P. Karadais, "Behind the cues: A benchmarking study for fake news detection," *Expert Systems with Applications*, vol. 128, no. 1, pp. 201–213, 2019.

[27] G. Lingam, R. R. Rout and D. Somayajulu, "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks," *Applied Intelligence*, vol. 49, no. 11, pp. 3947–3964, 2019.

[28] 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, pp. 6, 2017.

[29] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk *et al.,* in *Learning from Imbalanced Data Sets*. Berlin, Germany: Springer, 2018.

[30] Y. Lin, Y. Lee and G. Wahba, "Support vector machines for classification in nonstandard situations," *Machine Learning*, vol. 46, no. 1–3, pp. 191–202, 2002.

[31] B. Liu, Y. Ma and C. Wong, "Improving an association rule-based classifier," in *Proc. PKDD*, Lyon, France, pp. 293–317, 2000.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[33] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IJCNN*, Hong Kong, China, pp. 1322–1328, 2008.

[34] H. Han, W. Wen-Yuan and M. Bing-Huan, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. ICIC*, Hefei, China, pp. 878–887, 2005.

[35] H. M. Nguyen, E. W. Cooper and K. Kamei, "Borderline over-sampling for imbalanced data classification," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2009.

[36] T. Le, M. T. Vo, B. Vo, M. Y. Lee and S. W. Baik, "A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction," *Complexity*, vol. 2019, no. 2, pp. 1–12, 2019.

[37] A. Ali-Gombe and E. Elyan, "MFC-GAN: Class-imbalanced dataset classification using multiple fake class generative adversarial network," *Neurocomputing*, vol. 361, no. 4, pp. 212–221, 2019.

[38] T. Vo, T. Nguyen and T. Le, "A hybrid framework for smile detection in class imbalance scenarios," *Neural Computing and Applications*, vol. 31, no. 12, pp. 8583–8592, 2019.

[39] A. Akkasi, E. Varoglu and N. Dimililer, "Balanced undersampling: A novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text," *Applied Intelligence*, vol. 48, no. 8, pp. 1965–1978, 2018.

[40] A. Barushka and P. Hájek, "Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks," *Neural Computing and Applications*, vol. 32, no. 9, pp. 4239–4257, 2020.

[41] Q. Yan and Y. Cao, "Optimizing shapelets quality measure for imbalanced time series classification," *Applied Intelligence*, vol. 50, no. 2, pp. 519–536, 2020.

[42] G. He, W. Zhao, X. Xia, R. Peng and X. Wu, "An ensemble of shapelet-based classifiers on inter-class and intra-class imbalanced multivariate time series at the early stage," *Soft Computing*, vol. 23, no. 15, pp. 6097–6114, 2019.

[43] G. Lemaitre, F. Nogueira and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, pp. 17:1–17:5, 2017.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.,* "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[45] T. Fawcett, "An Introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.