Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8853468, 9 pages
https://doi.org/10.1155/2020/8853468

Research Article

Automated Fraudulent Phone Call Recognition through Deep Learning

Jian Xing,[1,2,3] Miao Yu ,[1,2] Shupeng Wang,[1,2] Yaru Zhang,[1,2] and Yu Ding[1,2]

[1]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China*
[2]*School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China*
[3]*National Computer Network Emergency Response Technical Team/Coordination Center of China Xinjiang Branch, Urumqi, China*

Correspondence should be addressed to Miao Yu; yumiao@iie.ac.cn

Received 21 June 2020; Revised 18 July 2020; Accepted 7 August 2020; Published 28 August 2020

Academic Editor: Ashok Kumar Das

Several studies have shown that the phone number and call behavior generated by a phone call reveal the type of phone call. By analyzing the phone number rules and call behavior patterns, we can recognize the fraudulent phone call. The success of this recognition heavily depends on the particular set of features that are used to construct the classifier. Since these features are human-labor engineered, any change introduced to the telephone fraud can render these carefully constructed features ineffective. In this paper, we show that we can automate the feature engineering process and, thus, automatically recognize the fraudulent phone call by applying our proposed novel approach based on deep learning. We design and construct a new classifier based on Call Detail Records (CDR) for fraudulent phone call recognition and find that the performance achieved by our deep learning-based approach outperforms competing methods. Experimental results demonstrate the effectiveness of the proposed approach. Specifically, in our accuracy evaluation, the obtained accuracy exceeds 99%, and the most performant deep learning model is 4.7% more accurate than the state-of-the-art recognition model on average. Furthermore, we show that our deep learning approach is very stable in real-world environments, and the implicit features automatically learned by our approach are far more resilient to dynamic changes of a fraudulent phone number and its call behavior over time. We conclude that the ability to automatically construct the most relevant phone number features and call behavior features and perform accurate fraudulent phone call recognition makes our deep learning-based approach a precise, efficient, and robust technique for fraudulent phone call recognition.

1. Introduction

Fraudulent phone call recognition represents an essential task for both preventing and curbing fraud effectively [1]. In recent years, with the continuous transfer of telephone fraud to overseas countries and the widespread use of VoIP and phone number modification software, fraudulent phone number is constantly changing and becoming more covert [2]. The traditional crowdsourcing model based on a black list is no longer effective as a result of these changes. Meanwhile, in order to avoid investigation, the fraudulent phone call behavior is constantly upgrading and changing, and the opposability is increasing [2]. The difficulty with this recognition task is randomness of fraudulent phone number and opposability of its call behavior.

“Scam Call Activity Regularity and Behavior Features Analysis Report 2016” [3] released by the 360 Internet Security Center and some other previous researches indicates that there are some differences between fraudulent phone calls and normal phone calls in call frequency, call time, long-distance call rate, and other behavior features [4]. At the same time, although a fraudulent phone number has randomness and variability, the phone number itself also has certain regularity, such as a nonstandard number, international number, short number, or fake number [5]. With the above features, many traditional machine learning approaches are proposed on the field of fraudulent phone call recognition [6–8].

In the related works, fraudulent phone call recognition is treated as a classification problem. This problem is solved by, first, manually engineering features of a fraudulent phone call and then classifying these features with state-of-practice machine learning algorithms. An essential step of traditional machine learning is feature engineering. Feature engineering is a manual process, based on intuition and expert knowledge, to find a representation of raw data that conveys characteristics that are most relevant to the learning problem. Proposed approaches [8] have shown that finding distinctive features is essential for accurate recognition of a fraudulent phone call. Moreover, the cost of these tasks is expensive as fraudulent phone numbers and corresponding call behaviors are dynamic. So far, the research community has not determined whether it can successfully automate the feature extraction step for classification. Therefore, the automatic and accurate recognition of fraudulent phone calls has become a challenge, and this is the key problem that we address in this work.

In this paper, we propose a novel approach for fraudulent phone call recognition based on deep learning (DL) [9]. Our approach can incorporate automatic feature learning, and thus, it is not defined by a particular feature set. This may be a game-changer in the arms race between fraud and antifraud, because the deep learning-based antifraud is designed to be adaptive to any perturbations in the features introduced by fraud. The approach we present in this work is the first automated fraudulent phone call recognition approach, and it outperforms the state-of-the-art approaches.

The key contributions of our work are summarized as follows:

(1) We design and construct a classifier based on Calling Detail Records (CDR) for fraudulent phone call recognition. The classifier only uses the CDR as input data, so it can be constructed easily, quickly, and efficiently. It provides a basic framework for recognition task and defines the main steps of the task

(2) Our study provides the first systematic exploration of state-of-the-art deep learning algorithms applied to fraudulent phone call recognition, namely, convolutional, recurrent, and feedforward deep neural networks. We design, tune, and evaluate three models—the Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and Stacked Denoising Autoencoder (SDAE). Our DL models are capable of *automatically* learning phone number features and call behavior features for fraudulent phone call recognition. We demonstrate that our DL-based approach achieves a higher accuracy rate than the state-of-the-art approaches

(3) We reevaluate previous work on our new real-world datasets. As a result of a systematic comparison of our novel DL-based approach to previous fraudulent phone call recognition approaches, we demonstrate comparable recognition results with slight improvements of up to 3.0%-4.7% on average. Furthermore, our DL models reveal more general and stable phone number features and call behavior features of fraudulent phone calls than the state-of-the-art approaches, which make them more robust to concept drift caused by a highly dynamic fraudulent phone number and its call behavior

(4) We make the generated dataset publicly available, allowing researchers to replicate our results and systematically evaluate new approaches to fraudulent phone call recognition

The rest of this paper is structured as follows. Section 2 describes the related work. Section 3 formally defines the fraudulent phone call recognition problem. Section 4 presents the proposed approach in details. Section 5 outlines the dataset we collected. Section 6 displays the experimental results. Finally, Section 7 concludes with discussion.

2. Related Work

This section reviews recent related work on fraudulent phone call recognition relying on traditional machine learning algorithms and the application of deep learning.

The past decade has witnessed remarkable progress in machine learning on various practical applications [10–17], especially in fraudulent phone call recognition. Previous studies have shown that fraudulent phone calls can be effectively recognized through cognitive learning of the phone number features and call behavior features. Among them, Zhou et al. [4] made a statistical analysis of a user's call behavior and found that the call time frequency, call time interval, call frequency of the same object, call cycle, and call interval had obvious regularity. However, due to the limited number of samples, it failed to extract the call behavior features of a fraudulent phone call. Wang and Wang [18] proposed a recognition approach of nuisance calls based on Random Forest. It preliminary found that phone numbers had features that could be used to identify them. However, the accuracy of the algorithm was only 84.30%. Ji et al. [6] proposed a recognition approach of fraudulent phone calls based on SVM. It only constructed a classifier for the call behavior features of a fraudulent phone call, but did not analyze the phone number features of the fraudulent phone call, and the accuracy of the algorithm was only 76%. Other researchers like [7, 8, 19, 20] chose to use Decision Tree, Naive Bayesian models, graph mining, and other approaches [21–23] to classify and analyze call behavior features.

Almost all of these studies selected features based on expertise and their knowledge on phone number rules and call behavior patterns of fraudulent phone calls. It is a result of manual feature engineering and standard feature selection. It is still unknown whether the fraudulent phone call can be successfully recognized by automatic feature engineering. To the best of our knowledge, the only research that successfully applies deep learning to the phone scam detection problem was made by Huang et al. [24, 25]. However, the accuracy of their deep learning approach only reached 83.83%. Moreover, the work does not assess applicability of other deep learning algorithms to the problem. There is still much room for the deep learning application of fraudulent phone call recognition.

The motivation of leveraging deep learning into this problem is as follows: to overcome the defects of manual engineering features through automatic feature engineering, to improve the recognition accuracy of fraudulent phone calls, and to improve timeliness and make the recognition task easy to accomplish.

In this paper, we construct a classifier for fraudulent phone call recognition, explore three deep learning models when trained on sufficient amounts of data, and evaluate the context of dynamic changes of a fraudulent phone number and its call behavior over time. We provide a basic tuning of the DL-based approach and finally achieve a higher accuracy rate than the state-of-the-art approaches.

3. Problem Definition

In this section, we introduce the mathematical notations and formally define the fraudulent phone call recognition problem. In our proposed approach, we follow previous work and formulate fraudulent phone call recognition as a binary classification problem. Namely, we perform a supervised binary classification, where we train a classifier on a set of labeled instances and test it by assigning a label to each unlabeled instance. A phone call t can be expressed in the form $(\mathrm{C}_t, \mathrm{L}_t)$, where C_t is a raw representation of a phone call and L_t is the class label corresponding to it. C_t is a length-176 array, which can be interpreted by a neural network. Assume that the type of phone call is 2, label L_t belongs to the set $\{0, 1\}$. As such, we state the fraud phone call recognition problem as follows:

Given the raw representation of a phone call C_t and its corresponding label L_t, we aim to learn the model $\mathscr{M}$ mapping C_t to L_t, which can automatically construct the most relevant phone number features and call behavior features and perform accurate fraudulent phone call recognition.

4. Proposed Approach

4.1. The Classifier We Constructed. Figure 1 shows the overview of our constructed classifier. It consists of a data extraction and data preprocessing phase and a training and evaluation phase. In the first phase, we extract nonstatistical metadata and statistical metadata from CDR [26] and then preprocess the above data for the next stage.

In the second phase, we use special algorithms to train the model and complete the evaluation task.

4.1.1. Data Extraction and Data Preprocessing. Seven nonstatistical and statistical metadata are extracted from six fields of CDR, which result in 176 dimensions. The six fields are START_TIME, END_TIME, CALLING_NUMBER, CALLED_NUMBER, CALL_DURATION, and CALLED_LO CATION.

4.1.2. Nonstatistical Metadata. CALLING_NUMBER is extracted from CDR as nonstatistical metadata. Meanwhile, duplicated data in one day are removed. The main operation of data preprocessing is to complete the length of the CALLING_NUMBER to 17 digits with zero and then use One-Hot Encoding for digital conversion. Finally, a length-170 array is constructed, which represents the nonstatistical metadata, namely, CALLING_NUMBER.

4.1.3. Statistical Metadata. Based on the above CALLING_NUMBER, we extract six statistical metadata from CDR. They are the number of CALLED_NUMBER, the number of CALLED_NUMBER (deduplication), the maximum similarity of CALLED_NUMBER, the average similarity of CALLED_NUMBER, the average CALL_DURATION, and the number of CALLED_LOCATION. The statistical period is one day. The main operation of data preprocessing is Min–Max Normalization, which converts all statistical metadata to interval $[0, 1]$. Finally, a length-6 array is constructed, which represents the six statistical metadata.

The classifiers used in related work were designed by carefully constructing feature vectors, as described in Section 2. Our constructed classifier integrates feature learning within the training process, enabling it to classify a phone call simply based on its initial representation. One-Hot Encoding is used to convert nonstatistical metadata; because there is no rescaling or normalization, the possible loss of information associated with the preprocessing steps is avoided. Min–Max Normalization is used to convert six statistical metadata, so all of them are converted to interval $[0, 1]$. All the above operations conform to the properties of mathematical operations performed by neural networks.

4.1.4. Training and Evaluation. Two types of metadata are simply concatenated together and used as input to this layer. Multiple traditional supervised machine learning algorithms, such as k-Nearest Neighbors (K-NN), Random Forest (RF), SVM, and DL algorithms, are used to train the fraudulent phone call recognition model and evaluate it.

4.2. Our DL-Based Methodology. In this section, we provide a detailed overview of our DL-based methodology. DL provides a series of powerful machine learning techniques with deep architectures. Deep neural networks (DNNs) are the basis of DL and utilize a multilayer of nonlinear mathematical data transformations to achieve automatic hierarchical feature extraction and selection. DNN demonstrates the superiority of feature learning in solving various tasks. In this study, we apply three major types of DNNs for fraudulent phone call recognition: a convolutional CNN, a recurrent LSTM, and a feedforward SDAE. We choose to apply the models that provide the capabilities and architectural characteristics to perform the task of automated feature extraction and to benefit from the nature of our input data. These DL algorithms are conceptually the most well-suited for the recognition task at hand.

4.2.1. Three Types of DNNs. In the existing types of DNNs and corresponding DL algorithms, we evaluate three major types of neural networks: convolutional, recurrent, and feedforward.

Firstly, we propose a DNN called CNN. It is an extension of the traditional multilayer perception, based on local receive fields, shared weights, and spatial or temporal subsampling. It is a classifier built on a series of convolutional layers. Convolutional layers are used for feature extraction,

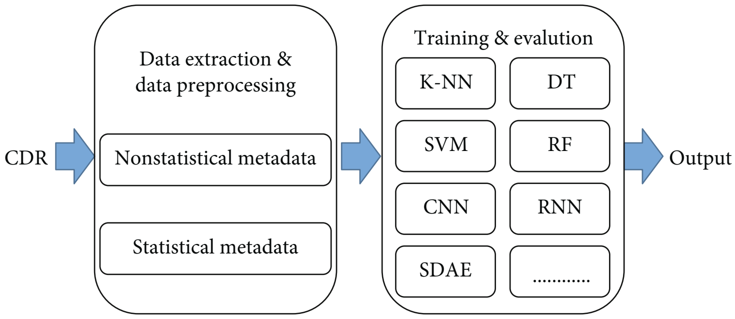

Figure 1: Overview of our constructed classifier.

starting with low-level features at the first layer and building up to more abstract concepts going deeper into the network. Convolutional layers learn numerous filters that reveal regions in the input data containing specific characteristics. These input instances are then downsampled to preserve the special regions. CNN searches for the most important features in this way as the basis for classification.

Then, we are going to introduce another DNN called LSTM. It is a recurrent neural network constructed by three internal gates, which are designed to allow the whole model to use back propagation to train the model and avoid gradient vanishing. Its design allows for learning long-term dependencies in data, enabling the model to interpret time series. In essence, the input phone number is a time series of number, and the temporal dynamics in these sequences are expected to highly reveal the corresponding phone number attributes.

The final DNN worth mentioning is SDAE. It is a deep architecture formed by stacking multiple DAE together and a feedforward network for feature learning through dimension reduction. It can extract the most prominent features in the input data hierarchically and classify them according to the derived features.

4.2.2. Hyperparameter Tuning. Traditional machine learning methods also need hyperparameter tuning but on a smaller scale than DL. Because of the parallelism of the DL algorithm, it is more feasible to tune the parameters of the DL model compared with the traditional model. As learning algorithms of neural networks are inherently parallel, graphical processing units (GPUs) can take advantage of this characteristic. Performing hyperparameter tuning on GPUs which compromises for intense computational requirements allows for rapid feedback of the model. For our DL experiments, we use one Nvidia RTX2080Ti GPU with 11 GB memory to accommodate parallelized training of the DNNs. Table 1 show the main values of the hyperparameters that we finally selected.

5. Datasets

One of the prerequisites for deep learning is the need for large amounts of training data to learn the underlying features. By processing enough representative data, the deep neural network can not only reveal the recognition features accurately but also better extend to undiscovered test cases. In the previous work, the collected datasets are relatively limited in size, which is specifically reflected in the total amount of data and the time span of collection.

In total, we evaluate our deep learning approach in comparison with traditional methods on two real-world datasets. Here, we detail the datasets collected in this article.

5.1. Six-Month Dataset. We collected all CDR from September 2018 to February 2019. Our six-month dataset contains more than 8.2 million normal phone call samples and 8284 fraudulent phone call samples. In real-world environments, the proportion of normal phone call samples is larger than that of fraudulent phone call samples. We combined the collected data into seven different datasets, which are summarized in Table 2. The normal phone call sample is regarded as the positive sample, and the fraudulent phone call sample is regarded as the negative sample. All negative samples were randomly divided into three parts: training set, validation set, and test set, respectively, 75%, 12.5%, and 12.5% of the negative samples. Meanwhile, in each part, the number of positive samples is 1, 10, 100, 200, or 1000 times the number of negative samples.

First, the training set consists of 6000 normal phone call samples and 6000 fraudulent phone call samples. The validation set and test set are composed of 1000 normal phone call samples and 1000 fraudulent phone call samples. In the training set, the proportion of normal phone call samples to fraudulent phone call samples is 1 : 1. In the remainder of the text, we refer to this dataset as SC_1.

Similarly, in the training set, the datasets of the proportion of normal phone call samples to fraudulent phone call samples of 10 : 1, 100 : 1, and 200 : 1 are referred to as SC_{10}, SC_{100}, and SC_{200} accordingly.

Second, the training set consists of 60000 normal phone call samples and 6000 fraudulent phone call samples. The validation set consists of 1000 normal phone call samples and 1000 fraudulent phone call samples. The test set consists of 10000 normal phone call samples and 1000 fraudulent phone call samples. In the test set, the proportion of normal phone call samples to fraudulent phone call samples is 10 : 1. In the remainder of the text, we refer to this dataset as TC_{10}. Similarly, in the test set, the datasets of the proportion of normal phone call samples to fraudulent phone call samples of

Table 1: Tuned hyperparameters of the selected DL models.

Hyperparameter	CNN	LSTM	SDAE
Optimizer	RMSProp	Adam	RMSProp
Batch size	512	512	512
Training epochs	10	300	500
Number of layers	8	2	5
Input units	176	176	176
Dropout	0.25	0.2	0.1
Activation	ReLU	Sigmoid	Sigmoid
Kernels	64/128	—	—
Kernel size	3	—	—
Pool size	2	—	—

Table 2: The information of the six-month dataset.

	Dataset	Number of normal phone call samples	Number of fraudulent phone call samples
SC_1	Training set	6000	6000
	Validation set	1000	1000
	Test set	1000	1000
SC_{10}	Training set	60000	6000
	Validation set	1000	1000
	Test set	1000	1000
SC_{100}	Training set	600000	6000
	Validation set	1000	1000
	Test set	1000	1000
SC_{200}	Training set	1200000	6000
	Validation set	1000	1000
	Test set	1000	1000
TC_{10}	Training set	60000	6000
	Validation set	1000	1000
	Test set	10000	1000
TC_{100}	Training set	60000	6000
	Validation set	1000	1000
	Test set	100000	1000
TC_{1000}	Training set	60000	6000
	Validation set	1000	1000
	Test set	1000000	1000

100 : 1 and 1000 : 1 are referred to as TC_{100} and TC_{1000} accordingly.

5.2. Re-Collection over Time Dataset. We collected all CDR from March 2019 to August 2019 for another six months. That is 1 to 6 months of data after the last data collection. Our re-collection over time dataset contains more than 7.9 million normal phone call samples and 2927 fraudulent phone call samples. We divided the collected data into six datasets by month, which are referred to as RC_1-RC_6 and are shown in Table 3.

The purpose of different test/train splits and amount of dataset used as different datasets is to evaluate the relationship between sample count and performance. The following experiments will elaborate on the relationship between this performance and implementation.

6. Experiments

In this section, we aim to enable a systematic comparison between our DL models and the models mentioned above, not only to evaluate the classification accuracy of the model on our new dataset but also to analyze the stability of generalization ability in real-world environments and the resilience of trained models to concept drift with a growing time gap between training and testing.

6.1. Reevaluation of the State of the Art. The models mentioned above in the literature have been proven to be suitable for this recognition problem and outperform other models; for this reason, we have selected them to compare with our DL-based models.

The first experiment achieves two objectives. The first objective is to confirm whether we can reproduce the prior work. The second objective is to assess whether we can obtain good classification results on our four new datasets, namely, SC_1, SC_{10}, SC_{100}, and SC_{200}, which are different in the training set. The criterion of evaluation is accuracy. To ensure the reliability of our experiments, we estimate the models' performance by conducting a 10-fold crossvalidation on each dataset.

The following results were obtained on a server with an Intel i9-9900k, 64 GB DDR4 memory and one Nvidia RTX2080Ti GPU. Table 4 shows the classification accuracy obtained through crossfold validation for the four algorithms on four datasets. All algorithms achieve better accuracy in the first two datasets. With the change of sample equilibrium in the training set, K-NN, SVM (RBF kernel), and RF are getting less accurate but still effective in the last two datasets. However, the accuracy of SVM (linear kernel) has decreased dramatically to about 50%. For binary classification, this means that the algorithm fails. One possible reason for the performance drop is that the classifier trained and evaluated in small data size might learn the partial or temporary features instead. Another interesting observation is that the classification accuracy is not more than 88% on SC_{200}. RF has the highest accuracy, and its effect remains stable. It achieves the highest accuracy of 98.55% on SC_{100}. The main conclusion here is that the RF-based classifier works very well and

Table 3: The information of re-collection over time dataset.

Dataset	Number of normal phone call samples	Number of fraudulent phone call samples	Date
RC_1	1521049	425	Mar 2019
RC_2	1150548	465	Apr 2019
RC_3	1269241	559	May 2019
RC_4	1417366	486	June 2019
RC_5	1438665	599	July 2019
RC_6	1115705	393	Aug 2019

Table 4: Accuracy of four traditional models on our four new datasets.

Dataset	K-NN	SVM (linear kernel)	SVM (RBF kernel)	RF
SC_1	95.30%	91.08%	93.95%	95.21%
SC_{10}	94.97%	84.18%	91.10%	97.77%
SC_{100}	89.48%	50.23%	84.40%	98.55%
SC_{200}	87.00%	50.03%	78.48%	87.87%

outperforms the other competing methods. As a result, we choose RF as the reference point for comparing our proposed approach with the state of the art. This decision is driven by the fact that RF performed the best on our four new datasets and proved to be more practically feasible. Therefore, we further evaluate our DL-based approach in comparison to RF.

6.2. Deep Learning for Fraudulent Phone Call Recognition. Here, we further introduce the experimental results of fraud phone call recognition based on DL. We evaluate three selected DNNs on our new dataset and assess the stability of their generalization capabilities. We assess their forecasting ability over time by testing their resilience to concept drift on data re-collected after training. Furthermore, we compare the results with RF, which is the most accurate traditional recognition method. All models use the hyperparameter selected in Table 1. All results reported in this section are computed via 10-fold crossvalidation.

6.2.1. Accuracy Evaluation. In this study, we evaluate the CNN, LSTM, and SDAE networks on our four new datasets, namely, SC_1, SC_{10}, SC_{100}, and SC_{200}. The criterion of evaluation is accuracy. The results are presented in Table 5.

First, according to these results, we can confirm the feasibility of fraudulent phone call recognition based on a DL approach with automatic feature learning. The highest success rate of the CNN, LSTM, and SDAE models is 99.70%, 99.00%, and 99.65%, respectively. These results are better than those of the traditional approaches in Section 6.1.

Table 5: Accuracy of the DL models on our four new datasets.

Dataset	CNN	LSTM	SDAE
SC_1	99.60%	99.00%	99.65%
SC_{10}	99.70%	97.20%	99.55%
SC_{100}	99.55%	97.80%	98.00%
SC_{200}	99.35%	97.50%	97.00%

Second, if we compare the three DNNs with each other, we observe that the CNN and SDAE models perform better than the LSTM model in terms of classification accuracy, with the CNN model being the most performant, especially on SC_{10}. Our interpretation is that even a small amount of the negative sample is sufficient for fraudulent phone call recognition up to 99% accuracy when deploying our model. Notably, LSTM performs much poorer; one possible reason is that it needs more fraudulent phone call samples to learn the sequence relationship among each dimension. We observe that as the positive sample increases in the training set, the performance of the three DL models gradually decreases following a similar trend, but it remains at a high level—the accuracy of the three DL models is still higher than 97%.

Third, Figure 2 compares the DL-based approach to RF. The evaluation results are better than RF's results presented in Table 4 in the previous subsection. This comparison illustrates that our DL-based approach can indeed successfully learn the features of the fraudulent phone call in an automated manner, and their generalization capabilities are obviously better than RF's, especially on SC_{200}.

6.2.2. Stability Evaluation. In this study, we evaluate the three DNNs and RF on our three new datasets, namely, TC_{10}, TC_{100}, and TC_{1000}, for assessing the stability of their generalization capabilities. In these datasets, we change the sample equilibrium in the test set to simulate the class imbalance in real-world environments. The criterion of evaluation is the AUC value, TPR, and FPR. We only select three datasets for evaluation and do not cover all sample distribution, so the AUC value is more convenient than accuracy to prove which model works better. The results are presented in Table 6.

The results show that the generalization capabilities of DL models are stable in the case of class imbalance. All the DL models are better than RF in terms of the AUC value. Some of the DL models are better than RF in terms of TPR or FPR. Meanwhile, when we compare the three DNNs with each other, we observe that they have achieved almost perfect performance in terms of the AUC value; all reached 0.99. The CNN and SDAE models are neck-and-neck for all datasets and consistently perform better than LSTM in terms of TPR. However, the LSTM model performs little better than the others in terms of FPR. This shows that the CNN and SDAE models are better at recognizing fraudulent phone call and the LSTM model is better at recognizing normal phone calls.

6.2.3. Concept Drift Evaluation. The presented experiments reflect the model's ability to recognize a fraudulent phone call. However, the results we obtained could not certainly infer whether the DNN reveals the actual features for

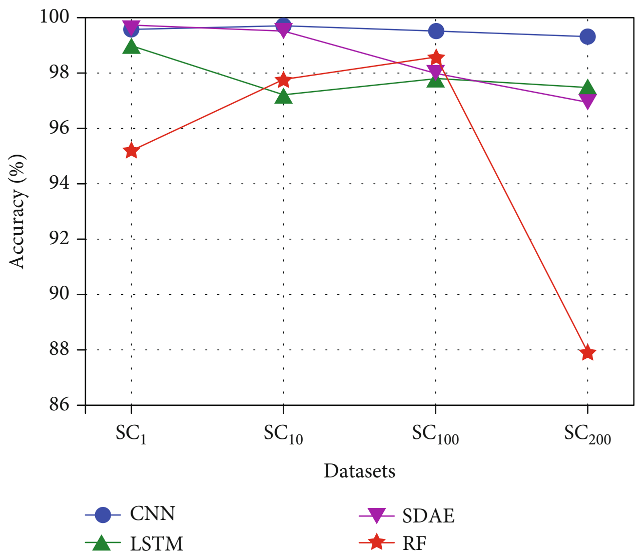

Figure 2: DL (CNN, LSTM, and SDAE) vs. RF on our four new datasets.

recognition or also learns occasional dynamics in the data instead which just happen to enable recognition. This experiment is intended to reveal how well our DNNs are able to extract features and generalize to new data.

In general, we call this phenomenon *concept drift*: a change over time in the properties of the class that the model is trying to predict. It is caused by highly dynamic changes of a fraudulent phone number and its call behavior over time. Therefore, the recognition might become less accurate over time. To reveal if our DNNs detect the actual features and assess how well they perform in case of dynamic changes, we train the models on a six-month dataset and test them on a re-collection over time dataset. The criteria of evaluation are accuracy, TPR, FPR, and precision. The results are depicted in Figure 3 for DL and RF. The plot indicates the recognition performance of various models trained on SC_{10} and evaluated on RC_1 to RC_6.

The figure demonstrates how the classification accuracy of our DNNs remains stable over time. The accuracy of all the models is higher than 99.5%. Furthermore, the SDAE model performs better than others in terms of TPR, reaching 74.69% on average, with the LSTM model being the worst performance—the TPR is only 3% on average. The CNN and RF models are neck-and-neck on all six datasets; the average is about 35%. Notably, all of them do well in terms of FPR. The FPR of all the models is less than 0.5%. However, the precision of all models is quite low. The RF model performs best, but the precision of it is only 9.4% on average. The most likely reason is the imbalance of positive and negative samples in real-world environments. Considering the extreme imbalance of the sample proportion in RC1 to RC6 (about 1 : 2700), the performance of the SDAE and RF models in terms of precision is acceptable. These results illustrate the high resilience of all the evaluated models, despite significant intervals of 1 to 6 months between the moment of training and the last evaluation. As such, these comparisons not only show that our DL-based approach indeed automates the feature engineering but also learn implicit features (hidden in the neural network), which are more robust against highly dynamic changes of a fraudulent phone number and its call behavior over time.

The main conclusion here is that the DL-based models are capable of extracting stable identifying information from our new dataset which allows for its recognition with a high accuracy, even several months after training. However, for fraudulent phone call recognition, the more important evaluation criterions of the prediction task are TPR and precision, which, respectively, represent the recall rate of the fraudulent phone call and the precision of the data identified as a fraudulent phone call. These targets closely relate to the feasibility and efficiency of investigation for fraudulent phone calls. From the above experiments, we can see that there is still much room for improvement in these two targets. One possible solution is to increase the number of negative samples in the training set so as to improve the recognition ability of the model for fraudulent phone calls. Especially for the LSTM model, it needs more fraudulent phone call samples to learn the sequence relationship among each dimension; in addition, one possible reason for LSTM's lower performance is that the structure of data is not all time series.

In the previous subsections, we have shown the relative performance of various DL models in comparison with each other and with the traditional RF classifier. In certain experimental settings, we improved beyond the state of the art, e.g., in classification accuracy and in stability of generalization capability on our new dataset. For the evaluations performed in this paper, we used the resources available at our institution, but we acknowledge that the model can be further improved by using more resources for data collection, model selection, and training.

6.2.4. Repeatability. Our source code together with the datasets is available at https://github.com/xingjian215/DLFPCC. We used Keras [27] with TensorFlow [28] backend for the implementation of the DNN classifiers.

7. Conclusion

In this paper, we proposed a new deep learning-based approach, which can effectively solve the problem of automated fraudulent phone call recognition. The main objective was to assess the feasibility of fraudulent phone call recognition through automated feature learning. We show that deep neural networks have the ability of learning phone number features and call behavior features of a fraudulent phone call automatically and outperform other competing methods among numerous research efforts in recent years on the real-world dataset. The three DNNs we investigated have shown their strengths and weaknesses in the context of fraudulent phone call recognition:

(1) CNN performed well in the accuracy and stability evaluations. However, this DNN has a higher risk of overfitting, which was revealed by the concept drift evaluation

Table 6: AUC value, TPR, and FPR of the DL models and RF on our three new datasets.

	CNN			LSTM			SDAE			RF		
Dataset	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR	AUC	TPR	FPR
TC_{10}	0.99	99.50%	0.10%	0.99	95.10%	0.00%	0.99	99.60%	0.07%	0.93	99.70%	0.00%
TC_{100}	0.99	99.90%	0.13%	0.99	95.80%	0.03%	0.99	99.90%	0.08%	0.93	99.70%	0.04%
TC_{100}	0.99	100.00%	0.12%	0.99	95.20%	0.02%	0.99	100.00%	0.08%	0.93	99.60%	0.06%

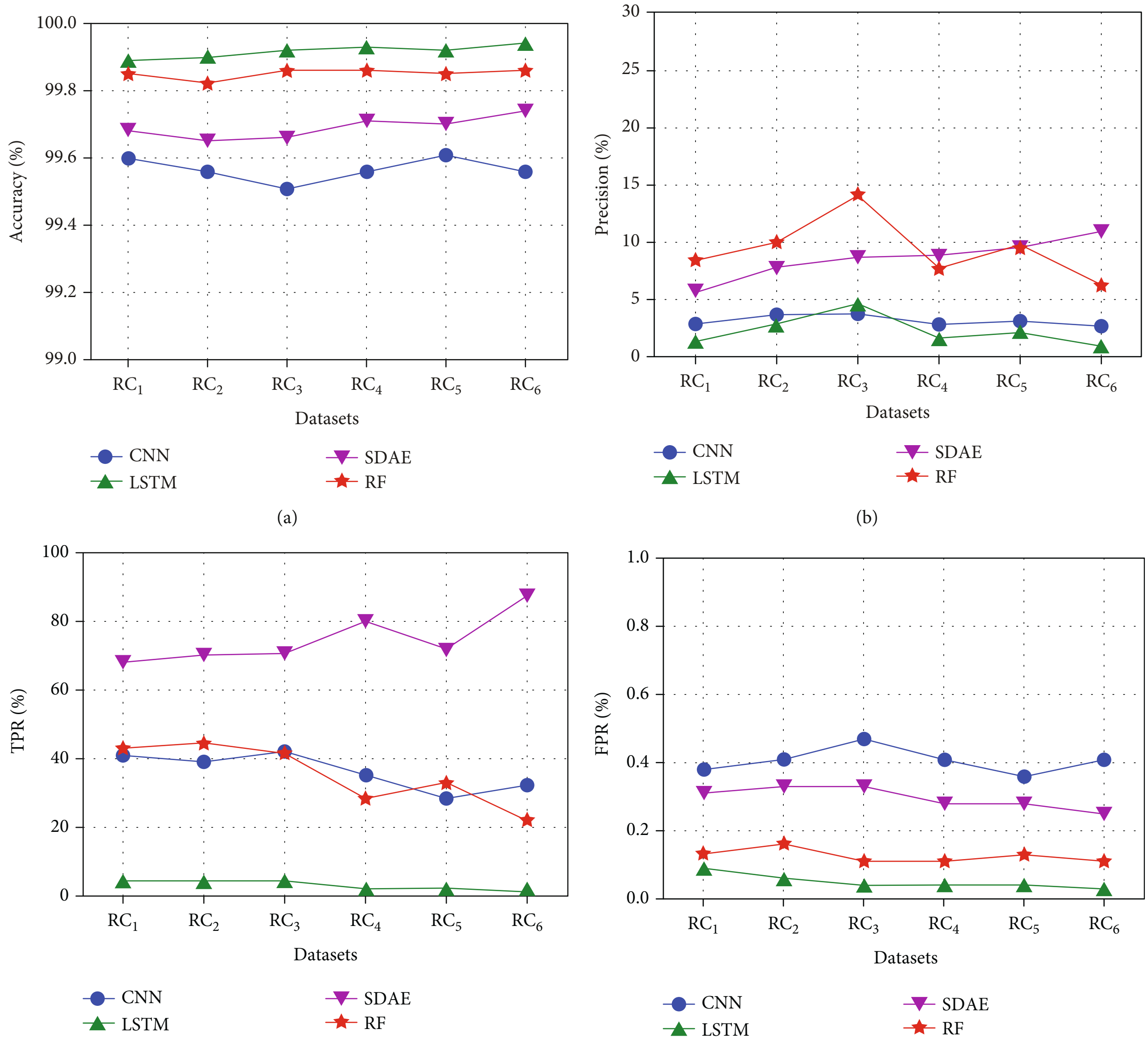

Figure 3: DL (CNN, LSTM, and SDAE) vs. RF resilience to concept drift: evaluation of RC_1 to RC_6 over time. (a) Accuracy of the DL models and RF. (b) Precision of the DL models and RF. (c) TPR of the DL models and RF. (d) FPR of the DL models and RF.

(2) LSTM performed the worst in the three selected DNNs, but it has its own characteristics in stability evaluation

(3) SDAE performed well overall in all evaluation and proved to be the best DNN in general. Especially in the concept drift evaluation, it was more robust than the other models

(4) All three DL models performed better than RF in the accuracy and stability evaluations, and SDAE proved to be more robust against a fraudulent phone number and its call behavior changes than RF

In conclusion, the application of deep learning makes fraudulent phone call recognition more accurate, effective, and robust.

Data Availability

The CDR data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 61931019 and U1803263).

References

[1] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad, "SoK: fraud in telephony networks," in *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 235–250, Paris, France, 2017.

[2] 360 Internet Security Center, "Telecomm fraud activity pattern and behavioral characteristics report 2016," 2019, http://zt.360.cn/1101061855.php?dtid=1101062366&did=490106344.

[3] 360 Internet Security Center, "China telecom fraud situation analysis report 2016," 2019, http://zt.360.cn/1101061855.php?did=490024605&dtid=1101061451.

[4] G. Zhou, G. Chen, and Y. Zhou, "User behavior in telecommunication fraud based on CDR analysis," in *Information Security and Communications Privacy*, pp. 114–118, The 30th Research Institute of China Electronics Technology Group Corporation, 2015.

[5] L. Li, Z. Ma, and Q. Chen, *Research of technology solutions and operation countermeasures to telephone fraud prevention and control*, Telecom science, 2014.

[6] Z. Ji, Y. Ma, and S. Li, *SVM based telecom fraud behavior identification method*, Computer Engineering & Software, 2017.

[7] T. Xu, *The design and implementation of visualization character relationship analysis system based on mining of call records*, Harbin Institute of Technology, 2014.

[8] X. Zhang, *Data mining techniques applied to a telecommunication anti-fraud system*, China University of Petroleum, 2006.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] X.-Y. Zhang, H. Shi, X. Zhu, and P. Li, "Active semi-supervised learning based on self-expressive correlation with generative adversarial networks," *Neurocomputing*, vol. 345, pp. 103–113, 2019.

[11] X.-Y. Zhang, S. Wang, and X. Yun, "Bidirectional active learning: a two-way exploration into unlabeled and labeled data set," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 12, pp. 3034–3044, 2015.

[12] X. Zhang, C. Xu, J. Cheng, H. Lu, and S. Ma, "Effective annotation and search for video blogs with integration of context and content analysis," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 272–285, 2009.

[13] X.-Y. Zhang, H. Shi, C. Li, K. Zheng, X. Zhu, and L. Duan, "Learning transferable self-attentive representations for action recognition in untrimmed videos with weak supervision," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1–8, Honolulu, Hawaii, USA, 2019.

[14] C. Wang, D. Wang, Y. Tu, G. Xu, and H. Wang, "Understanding node capture attacks in user authentication schemes for wireless sensor networks," *IEEE Transactions on Dependable and Secure Computing*, p. 1, 2020.

[15] D. Wang, W. Li, and P. Wang, "Measuring two-factor authentication schemes for real-time data access in industrial wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 9, pp. 4081–4092, 2018.

[16] D. Wang, H. Cheng, P. Wang, X. Huang, and G. Jian, "Zipf's law in passwords," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 11, pp. 2776–2791, 2017.

[17] R. Vera, P. Davy, and J. Marc, *Automated website fingerprinting through deep learning*, Network and Distributed Systems Security, San Diego, CA, 2018.

[18] Y. Wang and H. Wang, *Research on a combining algorithm for harassing calls to identify*, Telecom science, 2017.

[19] V. S. Tseng, J. C. Ying, and C. W. Huang, *FrauDetector: a graph-mining-based framework for fraudulent phone call detection*, ACM, KDD, 2015.

[20] J. J.-C. Ying, J. Zhang, C.-W. Huang, K.-T. Chen, and V. S. Tseng, "PFrauDetector: a parallelized graph mining approach for efficient fraudulent phone call detection," in *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 1059–1066, Wuhan, China, 2016.

[21] R. Li, Y. Zhang, Y. Tuo, and P. Chang, "A novel method for detecting telecom fraud user," in *2018 3rd International Conference on Information Systems Engineering (ICISE)*, Shanghai, China, 2018.

[22] J. C. Yang, J. C. Xu, and Q. Y. Yue, *Research on SMS fraud user identification based on spark and random forest*, Computer engineering and Science, 2019.

[23] J. M. Zhu, F. Chen, and Y. F. Huang, "The telephone harassment fraud prevention model based on block chain," *Journal of Applied Science*, vol. 37, no. 2, 2019.

[24] T. T. H.-D. Huang, C.-M. Yu, and H.-Y. Kao, "Data-driven and deep learning methodology for deceptive advertising and phone scams detection," in *Conference on Technologies and Applications of Artificial Intelligence*, pp. 166–171, Taipei, Taiwan, 2017.

[25] H. D. Huang and C. M. Yu, *Poster: adaptive data-driven and region-aware detection for deceptive advertising*, IEEE Symposium on Security and Privacy, San Jose, CA, 2016.

[26] M. Sanver and A. Karahoca, "Fraud detection using an adaptive neuro-fuzzy inference system in mobile telecommunication networks," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 15, no. 2, pp. 155–179, 2016.

[27] F. Chollet, "Keras," 2019, https://github.com/fchollet/keras.

[28] M. Abadi, "TensorFlow: large-scale machine learning on heterogeneous systems," 2019, https://www.tensorflow.org/.