Auto insurance fraud identification based on a CNN-LSTM fusion deep learning model

Huosong Xia

School of Management,
Wuhan Textile University,
Wuhan, China
and
Research Center of Enterprise Decision Support,
Key Research Institute of Humanities and Social Sciences,
Universities of Hubei Province,
Wuhan, China
Email: bxxhs@sina.com

Yanjun Zhou

School of Management,
Wuhan Textile University,
Wuhan, China
Email: 284661585@qq.com

Zuopeng Zhang*

Coggin College of Business,
University of North Florida,
Jacksonville, FL 32224, USA
Email: justin.zhang@unf.edu
*Corresponding author

Abstract: The traditional auto insurance fraud identification method relies heavily on feature engineering and domain knowledge, making it difficult to accurately and efficiently identify fraud when the amount of claim data is large and the data dimension is high. Deep learning models have strong generalisation abilities and can automatically complete feature extraction. This paper proposes a deep learning model for auto insurance fraud identification by combining convolutional neural network (CNN), long- and short-term memory (LSTM), and deep neural network (DNN). Our proposed method can extract more abstract features and help avoid the complex feature extraction process that is highly dependent on domain experts in traditional machine learning algorithms. Experiments demonstrate that our method can effectively improve the accuracy of auto risk fraud identification.

Keywords: auto insurance fraud; deep learning; CNN-LSTM.

Reference to this paper should be made as follows: Xia, H., Zhou, Y. and Zhang, Z. (2022) ‘Auto insurance fraud identification based on a CNN-LSTM fusion deep learning model’, *Int. J. Ad Hoc and Ubiquitous Computing*, Vol. 39, Nos. 1/2, pp.37–45.

Biographical notes: Huosong Xia is a Professor in the School of Management at Wuhan Textile University. He graduated from Huazhong University of Science and Technology in China and was a Visiting Scholar at Eller College of Management of the University of Arizona, USA from 2006 to 2007. His main research interests are knowledge management, data mining, e-commerce, and logistics information system. His publications have appeared in over 100 refereed papers in journals, book chapters, and conferences. He has obtained research funding from four projects with National Social Science Foundation of China and National Science Foundation of China.

Yanjun Zhou is a Master candidate in the School of Management at Wuhan Textile University. His main research interests are knowledge management, data mining, e-commerce, and logistics information system.

Zuopeng (Justin) Zhang is a faculty member in the Department of Management at University of North Florida. He received his PhD in Business Administration with a concentration in management science and information systems from Pennsylvania State University, University Park. His research interests include economics of information systems, knowledge management, electronic business, business process management, information security, and social networking. He has published research articles in various scholarly journals, books, and conference proceedings. He is the Editor-in-Chief of the *Journal of Global Information Management*. He also serves as an associate editor and an editorial board member for several other journals.

1 Introduction

As one of the four types of financial fraud, insurance fraud has become an increasingly serious social problem that causes social and insurance institutions to produce immeasurable substantial losses (Dai et al., 2017; Ngai et al., 2011; Wang and Xu, 2018). For example, according to the statistics from the US Anti-Insurance Fraud Coalition, more than $80 billion of fraudulent claims are being processed each year, while auto insurance fraud is one of the highest fraud types (Van Vlasselaer et al., 2017). The Insurance Institute of China and Oneconnect Financial Technology jointly released a white paper on intelligent risk control of China's insurance industry in 2019. According to the white paper, fraud and leakage in China's automobile insurance industry accounts for 20% of the total amount of claims, and the corresponding annual loss exceeds 20 billion yuan. Therefore, it is crucial to develop effective models to identify auto insurance fraud (Zhang and Zhu, 2020).

Auto insurance fraud identification is essentially a two-category problem, which is to identify and predict a policyholder's 'normal' or 'bullying fraud' claims through auto insurance claim data. However, insurance industry claims data has the characteristics of high dimensions, diverse data types, and large-scale, so the risk control and fraud detection capabilities of auto insurance companies need to be further strengthened (Abdallah et al., 2016). To avoid the losses resulted from auto insurance fraud, a considerable number of practitioners and scholars use machine learning algorithms to construct car insurance fraud warning models. For example, Sundarkumar et al. (2015) developed a type of support vector machine (SVM) with under-sampling to investigate auto insurance fraud. Hassan and Abraham (2016) proposed an insurance fraud detection model based on artificial neural network (ANN), decision tree (DT), and SVM, and manually selected 24 predictors from the original dataset to create the final dataset. Wang et al. (2020) proposed a modular neural network subnet training method to analyse the adjacent data sets while learning the target dataset. Yan et al. (2017) combined expert experience to manually select 27 fraud features related to auto insurance, using random forest and ant colony optimisation algorithms for fraud identification. However, traditional machine learning methods rely too much on feature engineering, which has the following limitations. First, the feature construction process is complex, time-consuming, and labour-intensive, requiring professional domain knowledge to determine important features, which is challenging. Second, feature selection is subjective, and it is difficult to obtain the hidden attributes from the data. The selection of different features has a great influence on the efficacy of the algorithm, so all the features of the data set cannot be fully reflected. Therefore, the traditional feature-based approach has certain limitations in auto insurance fraud detection.

Deep learning has been actively employed to in fraud detection and anomaly identification research. The combination of nonlinear transformation methods fits more abstract functions in large or complex data; by using this method, the shortcomings of feature engineering can be compensated by the automatic extraction of useful features (Chen and Lin, 2014). In the deep learning models, convolutional neural networks (CNN) and recurrent neural networks (RNN) extract hidden features by calculating the correlation between adjacent regions. By adding convolution operations between neurons in adjacent layers, using local visual field, and sharing convolution kernel weight, the CNN model can generate good feature expression ability and can automatically extract essential features of different levels from high-dimensional features. long- and short-term memory (LSTM) model, as a variant of RNN, finds specific sequences and previous states by calculating the correlation between current states and generates new features, effectively enhancing the interdependence between features. In the fraud identification field, Wang and Xu (2018) used the LDA algorithm in text mining to extract text features, and then input the deep neural network (DNN) model for car insurance fraud identification. For natural language processing, Wang et al. (2016) used the CNN-LSTM model for conducting fine-grained sentiment analysis. Oh et al. (2018) designed an automated system in the medical field by combining CNN and LSTM to diagnose whether the heart rhythm is abnormal. In the field of computer networks, Kim and Cho (2018) used the C-LSTM model combined with CNN, LSTM, and DNN to effectively discover anomalies in network traffic data and demonstrated good performance. In the field of fault detection, Oehmcke et al. (2018) achieved sensor fault prediction by combining a convolutional module with an improved deep learning architecture of the LSTM layer.

Table 1 Related works on auto insurance fraud

Category	*Author*	*Method*	*Description*	*Performance index*	*Value*
Statistical-based modelling	Weisberg and Derrig (1998)	Multiple linear regression	Obtain key claim fraud characteristics based on modelling	No index	No value
	Pinquet et al. (2007)	Bivariate probability model	Offset selection bias without using a random audit strategy	Accuracy	62.0%
	Bermúdez et al. (2008)	Bayesian skew logit model	Fraud two classification using asymmetric link functions	Accuracy recall precision	99.5% 99.6% 99.8%
	Jin et al. (2005)	Binary selection model	Predicting the probability of insurance fraud through logit and probit models	Predictive probability score and Silva's score test	---
Traditional machine learning-based modelling	Li et al (2018)	Random forest	A new random forest classification model based on principal component analysis and potential nearest neighbours	Accuracy	87.0%
	Sundarkumar et al. (2015)	Support vector machines	OCSVM-based undersampling method for fraud identification	Accuracy recall precision	60.61% 90.74% 58.69%
	Subudhi and Panigrahi (2020)	Decision tree, support vector machine, cluster data mathematics, multi-layer perceptron	After extracting features by clustering algorithm, input four classifiers for fraud identification	Accuracy recall specificity	84.34% 66.67% 86.95%
Unsupervised modelling	Šubelj et al. (2011)	Social network analysis	Explore different physical relationships in social networks to discover fraudulent entities.	Accuracy recall specificity	87.20% 89.13% 86.67%
	Nian et al. (2016)	Spectral sorting method anomaly detection	Detecting kernel similarity recognition between feature variables	AUC recall	74.00% 91.00%
	Yu et al. (2017)	Gang characteristics microscopic modelling	Matrix-based similarity calculation, rank sorting, and transformation algorithm detection	No index	No value

In this paper, we propose a deep learning model combining CNN and LSTM for automatic feature extraction and prediction of auto insurance fraud. The features of spatial structure from data are extracted by CNN and then used as the input for LSTM which will enhance interdependence between features. Finally, the features produced by the LSTM layer become the input into the DNN layer which will complete the accurate identification for auto insurance fraud by function fitting.

The remainder of the paper proceeds as follows. The next section reviews prior literature related to auto insurance fraud identification. Section 3 outlines our proposed hybrid deep learning model. Section 4 presents the experiment results and shows that our proposed model performs better than the traditional algorithms of machine learning. The last section concludes the paper by highlighting our findings and research limitations.

2 Literature review

There exists abundant research related to auto insurance fraud identification. Fraud identification can be broadly classified into three categories: statistical-based modelling, modelling based on traditional machine learning, and unsupervised modelling, shown in Table 1. We briefly summarised some technologies described in the literature, including performance index and values in accuracy, recall, and precision, and discussed the definition and meaning of performance index in Section 4.2.

Statistical-based method modelling is a common method for the early use of computer technology for auto risk fraud identification. For example, Weisberg and Derrig (1998) used multiple linear regression models to analyse the important characteristics of auto insurance fraud. Pinquet et al. (2007) created a statistical bivariate probability model as an audit strategy for analysing suspicious claims in insurance records. Bermúdez et al. (2008) suggested using a skewed Bayesian binary logit model to identify malicious claims found in the auto insurance market. Jin et al. (2005) considered the logit model to be more appropriate for fraud identification through scoring tests. While statistical-based methods have shown efficacy in auto fraud detection, the usefulness of statistical models are limited by the pre-determined functional form of statistical methods and fraudulent data in the same distribution as normal data (Viaene et al., 2005). Researchers such as Li et al. (2018), Sundarkumar et al. (2015) and Subudhi and Panigrahi (2020) proposed a targeted data sampling method to reduce the category imbalance in automobile insurance data and then used the traditional machine learning model for binary classification. Šubelj et al. (2011) suggested that insurance claim data can be characterised by a network, where vertices represent entities, such as cars or drivers, and edges represent relationships among entities. They classify each attribute as either an intrinsic attribute or a relationship attribute and use the relationship attributes to display the relationship between entities. Finally, they identify fraudulent claims by using an iterative evaluation algorithm (IAA). Unsupervised modelling does not rely on data tags but uses unlabelled samples for anomaly detection, but it also brings the drawbacks of complex model evaluation criteria, which are difficult to assess model performance without labels (Rayana and Akoglu, 2016).

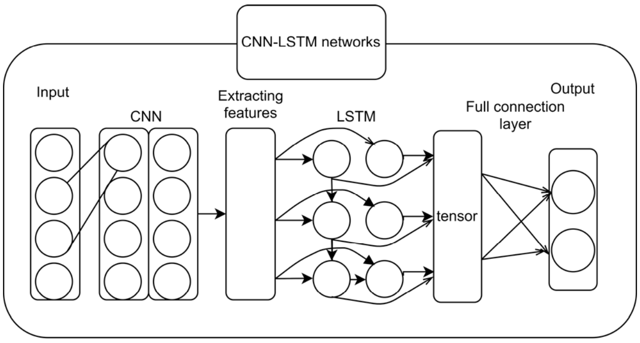
In summary, the research of the above algorithms has been used in fraudulent identification of some important variables that are clearly defined in the research of auto insurance claims. However, prior research has not been able to extract and utilise higher-level and more abstract essential features from the high-dimensional features of data. Therefore, this paper chooses to combine the RNN with strong memory ability and the CNN with strong feature expression ability in deep learning to fully exploit the data for vehicle risk fraud identification.

3 Proposed model

3.1 Mixed deep learning model

The proposed hybrid deep learning model includes CNN, LSTM, and DNN layers. Figure 1 displays the overall framework of the suggested car insurance fraud identification. Unlike traditional machine learning classifiers, deep learning algorithms do not rely on any feature engineering. Instead, they use the raw data itself to obtain the best feature through nonlinear transformations of multiple hidden layers. The model uses pre-processed data as the input, automatically extracting local common features through CNN's good feature expression, and then passing the learned features to the LSTM layer. Such a process helps to obtain inter-sequence related information while removing noise, invisibly simulates the collinear dependence between data attributes, and passes the feature to the fully connected DNN layer for fraud classification through the Relu nonlinear activation function.

Figure 1 The CNN-LSTM fusion neural network structure



The CNN layer consists of several convolutional layers, each of them containing multiple convolution kernels; different convolution kernel operations represent features extracted from different perspectives. By inputting various variables related to auto insurance fraud such as claimant related characteristics, accident occurrence characteristics, and vehicle information characteristics into the original data of the CNN layer, low-level features will be extracted from the first convolutional layer and then be passed to the next convolutional layer to extract high-level features. The specific calculation of convolutional layer is shown in equation (1) as follow:

$$x_j^{l+1} = \sigma\left(\sum_{i\in M_j} x_i^j * w_{ij}^{l+1} + b_j^{l+1}\right), \tag{1}$$

where x_j^{l+1} denotes the output value of the *j* unit from the $l + 1$ layer. σ is represented by an activation function. M_j indicates the position where the convolution kernel currently corresponds to the output feature. x_i^l represents an input characteristic of the *i* unit from the *l* layer. w_{ij}^{l+1} indicates the weight between the *i* unit from the *l* layer and the unit from the *j* layer. b_j^{l+1} denotes the bias of the *j* unit from the $l + 1$ layer. The CNN in this paper uses two layers of convolution. Each layer of convolution applies 128 convolution kernels of 3 × 3 size, with a step size of 1, and a padding of 1. The nonlinear activation function uses the ReLU function.

In the LSTM layer, the association among features can be better correlated by processing abstract features extracted from the CNN layer. LSTM is an improved form of the RNN. The introduction of gating mechanisms can be used to update information and address the problem of gradient disappearance and gradient explosion resulted from traditional RNN (Hochreiter and Schmidhuber, 1997). The forgetting gate, the external input gate, and the output gate together constitute a gating unit system, which specifies the mechanism to determine the state of each individual memory cell based on multiplication and controlling the update of each gate by a continuous value between 0 and 1. Formulas of the forgetting gate, the external input gate, and the output gate are shown in equations (2), (3), and (4) as follows:

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right), \quad (2)$$

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right), \quad (3)$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right) \quad (4)$$

where t represents a specific state, i denotes a cell unit, σ is a nonlinear activation function, $x^{(t)}$ is an input vector, $h^{(t)}$ represents the current hidden layer vector that contains the output of all LSTM cells, and b, U, W denote the offset vector, input weight, and cyclic weighting of forgetting gate, respectively.

Equations (5) and (6) represent state units and hidden states that can be determined by the input as well as the forget and output gates:

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right), \quad (5)$$

$$h_i^{(t)} = \sigma\left(s_i^{(t)}\right) q_i^t. \quad (6)$$

LSTM selectively records information to the status unit through the input gate, reduces invalid information by controlling weight through the forgetting gate, and outputs with historical information through the output gate. This makes it perform better in anomaly detection tasks (Ordóñez and Roggen, 2016).

The last layer combines the fully connected layer and the ReLU classifier to tile the output of the LSTM unit into a one-dimensional vector used for vehicle risk fraud binary classification identification, shown in equation (7) as follows,

$$p_t = \sigma\left(Wh_t + b\right), \quad (7)$$

where $p \in [0, 1]$, indicating the type of fraud identification: either normal type or fraudulent type. σ is the nonlinear activation function. W and b represent the learned weights and deviation values, respectively. h_t denotes the hidden state of flat layer.

3.2 Framework

The proposed deep learning model performance depends on different network types of various parameters, such as CNN layer number, convolution kernel number, LSTM layer number, and LSTM unit number. After a number of experimental selections, the relevant parameters of the CNN and the fully connected layer become an artificially set, as shown in Table 2. In this study, we hope to identify fraudulent auto insurance claim data. As the auto insurance claim data can be divided into five categories: claimant-related information, accident-related information, motor vehicle insurance-related information, insurance claim status-related information, and claimant vehicle information, combined with experimental tuning, the model is set based on the input sequence of LSTM being 5, the sequence dimension being 5, and 2 hidden layers containing 64 neurons.

Table 2 Parameter setting of CNN and full connection layer

Type	*Filter*	*Kernel size*	*Stride*	*Param*
Convolution 1	128	3	1	3*3*1*128+128
Convolution 2	1	3	1	3*3*128*1+1
Full connection 1	-	-	-	Batch size*64*256+1
Full connection 2	-	-	-	Batch size*256*256+1

4 Experiment

4.1 Descriptive data

Since auto insurance fraud claims data often involves trade secrets and is difficult to obtain, this study selects the 'carclaims.txt' auto insurance dataset (Phua et al., 2004). According to the existing literature, this is the only open and widely used labeled claim dataset in the field of automobile insurance fraud in the real world. The dataset is shown in Table 3 and includes 31 predictive variables and 1 tag variable. This includes 15,420 historical claims data pieces, among which are 14,497 non-fraud cases and 923 fraudulent cases. Due to the unbalanced label sample data, SMOTE oversampling method is used to increase the label positive example, balancing the number of positive and negative examples of label samples (He and Garcia, 2009). In the experiment, we randomly selected 80% of the dataset as the training set, 20% of the data as the test set, and 10% of the training set as the verification set.

Due to its public availability, most researchers have successfully applied this dataset to validate the performance of their fraud identification models. Table 4 shows the indicators produced by some researchers who have used this data set to evaluate their model performance.

Since the input of CNN-LSTM model must be numbers, referring to Phua et al. (2004), it is suggested to take a series of data preprocessing steps. By converting the classified data into single heat vector coding, the numerical data is standardised in the range of [0, 1] through equation (8),

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \quad (8)$$

which helps accelerate the speed of gradient descent to find the optimal solution and being more conducive to comprehensive comparison.

Table 3 Attributes of insurance data

Num.	*Attribute*	*Description*
1	Month	Month for the accident
2	Week of month	Week of month for the accident
3	Day of week	Day of week for the accident
4	Month claimed	Month for the claim
5	Week of month claimed	Week of month for the claim
6	Day of week claimed	Day of week for the claim
7	Year	1994, 1995, or 1996
8	Make	Vehicle manufacturer (19 companies)
9	Accident area	Urban or rural
10	Gender	Female or male
11	Marital status	Married, divorced, single, or widowed
12	Age	Policy holder's age
13	Fault	Third party or policy holder
14	Policy type	Policy type (1–9)
15	Vehicle category	Utility, sport, or sedan
16	Vehicle price	Vehicle's price (6 groups)
17	Rep. number	ID of the representative who processed the claim (16 IDs)
18	Deductible	Amount to be subtracted before claim payment
19	Driver rating	Driving skill (4 groups)
20	Days: policy accident	Policy's remaining days when accident happened
21	Days: policy claim	Policy's remaining days when claim was submitted
22	Past number of claims	Number of claims in the past
23	Age of vehicle	Age of the vehicle (8 groups)
24	Age of policy holder	Age of the policy holder (9 groups)
25	Policy report filed	Yes or no
26	Witness presented	Yes or no
27	Agent type	External or internal
28	Number of supplements	Number of supplements
29	Address change claim	Number of times that address has been changed
30	Number of cars	Number of cars
31	Base policy (BP)	Collision, liability, or all perils,
32	Class	Fraudulent or not

4.2 *Results and analysis*

For the purpose of identifying auto insurance fraud, to detect as few accidental injuries as possible and to detect as accurately as possible, this study used three evaluation indicators including the accuracy rate (Accuracy), recall rate (Recall), and precision rate (Precision), which are shown in equations (9), (10), and (11) as follows,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}, \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{10}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{11}$$

in which *TP* indicates true positive, *TN* denotes true negative, *FN* represents false negative, and *FP* stands for false positive.

Table 4 Different models for the same vehicle insurance fraud dataset (unit: %)

Related work	*Model performance evaluation*		
	Accuracy	*Recall*	*Precision*
Subudhi and Panigrahi (2020)	84.34	66.67	-
Xue et al. (2010)	88.70	-	83.33
Phua et al. (2004)	60.00	-	-
Sundarkuma et al. (2015)	60.31	90.79	58.69
Nian et al. (2016)	--	91.00	-
Šubelj et al. (2011)	87.20	89.13	65.08
Xu et al. (2011)	88.7	-	-
Tao et al. (2012)	-	91.31	-

4.3 *Comparison with other machine learning algorithms*

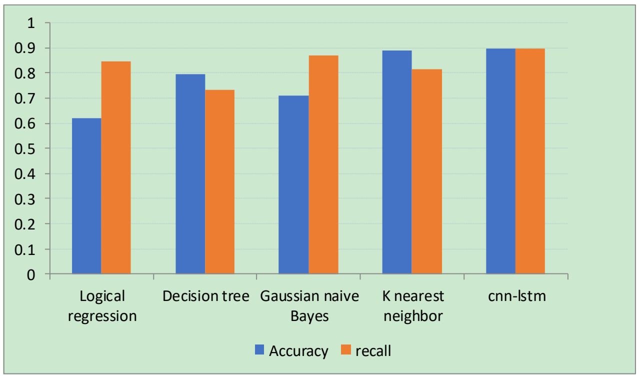
To confirm the efficacy of the proposed CNN-LSTM auto risk fraud identification model, we compare its performance with those of other machine learning algorithms by dividing the data into training sets, test sets, and verification sets. K-nearest neighbour method is a popular classification method in data mining and statistics because of its simple implementation and significant classification performance (Zhang et al., 2017). It is also one of the most widely used methods in pattern recognition (Gou et al., 2019). Decision tree is an advanced data mining technology, which is widely applied in many areas, such as classification and recognition (Li et al., 2019; Chen et al., 2017). Naive Bayes is one of the most effective and efficient classification algorithms and its classifiers still tend to perform very well under unrealistic assumptions. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful classifiers. Therefore, the acceleration of such an algorithm becomes a great asset in machine learning applications (Tzanos et al., 2019). Logical regression and decision tree are two efficient algorithms in supervised learning (Zhang et al., 2019). Therefore, four algorithms – logical regression, decision tree, k-nearest neighbour, and Gaussian naive Bayes – are selected to compare with the proposed approach. Table 5 indicates the hyper-parameter settings of the machine learning methods used in the experiments to attain high performance for each model.

Table 5 Parameters of machine learning algorithm

No.	*Algorithm*	*Description*
1	Logical regression	C parameter = 1, penalty = 'L1'.
2	Decision tree	criterion='entropy', max_depth=10, min samples split = 2, min sample leaf = 1.
3	Gaussian naive Bayes	No value.
4	K nearest neighbor	n_neighbors = 3.

The comparison shows that the suggested CNN-LSTM model accomplishes the highest accuracy and recall rate in auto insurance fraud identification. The accuracy rates of logical regression, decision tree, Gauss naive Bayes, and K-nearest neighbour are 62.52%, 80%, 71.53%, and 89.95%. The accuracy of CNN-LSTM is 90%, which is higher than those of the four algorithms: 27.48%, 10%, 18.47%, and 0.05%. The recall rates of the four algorithms are 85.11%, 74.30%, 87.61%, and 82.36% respectively, and the recall rate of CNN-LSTM is 90%. It can be seen that the recall rates of CNN-LSTM are higher than those of the four algorithms: 4.89%, 15.70%, 2.39%, and 7.64%. Figure 2 shows the accuracy and recall rate obtained through experiments.

Figure 2 Compared with the traditional machine learning algorithm (see online version for colours)



4.4 *Comparison of model combination performance*

We further conduct experiments to demonstrate that the recommended method is superior to other models based on deep learning. Experiments were performed on all combinations of the three models with CNN, LSTM, and DNN as the baselines. Table 6 contains a comparison of the performance of various deep learning model combinations.

According to the experimental results in Table 6, compared with CNN model and LSTM model, the accuracy and recall rate for the pairwise combinations among CNN, LSTM, and DNN models are significantly improved, and the precision rate is slightly decreased. Furthermore, compared with STM + DNN, CNN + DNN, and CNN + LSTM models, the accuracy and recall rate of CNN + LSTM + DNN model were improved by 2.3% and 2.3%, 5.6% and 5.6%, and 14.5% and 14.5%, respectively. The precision rate is about the same for all the combinations of CNN, LSTM, and DNN models.

Table 6 Comparison of different deep learning algorithms

Method	*Accuracy*	*Precision*	*Recall*
LSTM	0.636	0.931	0.636
CNN	0.628	0.937	0.628
LSTM+DNN	0.873	0.905	0.873
CNN+DNN	0.840	0.914	0.840
CNN+LSTM	0.751	0.915	0.751
CNN+LSTM+DNN	*0.896*	*0.907*	*0.896*

The results of the experiments in Table 6 demonstrate that the suggested CNN-LSTM model has higher accuracy, recall rate, and precision than other deep learning models, indicating its value and usefulness in the field of auto fraud detection.

5 Discussion and conclusions

The rapid development of the insurance industry has witnessed the increasing rate of insurance frauds at an alarming speed. As the main concerns of insurance frauds, Auto insurance frauds disturb the normal market order, affect the normal operation of insurance companies, and bring a series of problems and adverse effects to the society. Therefore, it is urgent to identify auto insurance fraud efficiently and effectively. As the technology to address such needs is yet to be further improved, this paper focuses on the auto insurance fraud identification method.

This study proposes a CNN-LSTM hybrid deep learning model for auto risk fraud identification. The method directly learns features from the original data, which is not dependent on feature engineering or expert knowledge in the field of auto insurance fraud. The validity of the model is confirmed by the reality that the fraud detection performance of this method is higher than that of the previous feature-based machine learning method and other independent deep learning models. Based on our experiments, the feature expression ability of the CNN on the image data is applied to obtain the abstract features of different levels in the auto insurance claim data, and the LSTM layer is used to extract the contextual information on the time axis of the auto insurance claim. In the field of fraud identification, the ability of feature expression and enhancing the connection between features is fully exploited.

There are limitations and deficiencies in this study. First, because the data of auto insurance fraud claims generally involves trade secrets and is difficult to obtain, this paper applies the only open data set. This data set contains a relatively small amount of data, so future research should obtain and apply larger and recent data sets to verify the method of deep learning. Second, since the method of use is a combination of deep learning methods, there are many parameters in the model that need to be adjusted and tried to find errors, which makes the model training a little difficult. Third, the deep learning method is similar to a 'black box'. Although it can achieve better prediction results, it is difficult to explain the reasons. Future research is needed on the parameter selection of the model and the interpretation of the model.

Acknowledgements

This research has been supported by the National Natural Science Foundation of China (71871172, Model of Risk Knowledge Acquisition and Platform Governance in FinTech based on Deep Learning; 71571139, Outlier Analytics and Model of Outlier Knowledge Management in the context of Big Data). We deeply appreciate the suggestions from fellow members of Xia's project team and Research Center of Enterprise Decision Support, Key Research Institute of Humanities and Social Sciences in Universities of Hubei Province (DSS20180204).

References

Abdallah, A., Maarof, M.A. and Zainal, A. (2016) 'Fraud detection system: a survey', *Journal of Network and Computer Applications*, Vol. 2016, No. 68, pp.90–113.

Bermúdez, L., Pérez, J.M., Ayuso, M., Gómez, E. and Vázquez, F.J. (2008) 'A Bayesian dichotomous model with asymmetric link for fraud in insurance', *Insurance: Mathematics and Economics*, Vol. 42, No. 2, pp.779–786.

Chen, W., Xie, X., Peng, J., Wang, J., Duan, Z. and Hong, H. (2017) 'GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models', *Geomatics, Natural Hazards and Risk*, Vol. 8, No. 2, pp.950–973.

Chen, X.W., and Lin, X. (2014) 'Big data deep learning: challenges and perspectives', *IEEE Access*, Vol. 2014, No. 2, pp.514–525.

Dai, Z., Galeotti, F. and Villeval, M.C. (2017) 'Cheating in the lab predicts fraud in the field: an experiment in public transportation', *Management Science*, Vol. 64, No. 3, pp.1081–1100.

Gou, J., Qiu, W., Yi, Z., Shen, X., Zhan, Y. and Ou, W. (2019) 'Locality constrained representation-based K-nearest neighbor classification', *Knowledge-Based Systems*, Vol. 167, No. 3, pp.38–52.

Hassan, A.K.I. and Abraham, A. (2016) 'Modeling insurance fraud detection using imbalanced data classification', in *Advances in Nature and Biologically Inspired Computing*, pp.117–127, Springer, Cham.

He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', *IEEE Transactions on knowledge and data engineering*, Vol. 21, No. 9, pp.1263–1284.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, Vol. 9, No. 8, pp.1735–1780.

Jin, Y., Rejesus, R.M. and Little, B.B. (2005) 'Binary choice models for rare events data: a crop insurance fraud application', *Applied Economics*, Vol. 37, No. 7, pp.841–848.

Kim, T.Y. and Cho, S.B. (2018) 'Web traffic anomaly detection using C-LSTM neural networks', *Expert Systems with Applications*, Vol. 106, No. 9, pp.66–76.

Li, M., Xu, H. and Deng, Y. (2019) 'Evidential decision tree based on belief entropy', *Entropy*, Vol. 21, No. 9, p.897.

Li, Y., Yan, C., Liu, W. and Li, M. (2018) 'A principle component analysis-based random forest with the potential nearest neighbor method for automobile insurance fraud identification', *Applied Soft Computing*, Vol. 70, No. 9, pp.1000–1009.

Ngai, E.W., Hu, Y., Wong, Y.H., Chen, Y. and Sun, X. (2011) 'The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature', *Decision Support Systems*, Vol. 50, No. 3, pp.559–569.

Nian, K., Zhang, H., Tayal, A., Coleman, T. and Li, Y. (2016) 'Auto insurance fraud detection using unsupervised spectral ranking for anomaly', *The Journal of Finance and Data Science*, Vol. 2, No. 1, pp.58–75.

Oehmcke, S., Zielinski, O. and Kramer, O. (2018) 'Input quality aware convolutional LSTM networks for virtual marine sensors', *Neurocomputing*, Vol. 275, No. 1, pp.2603–2615.

Oh, S.L., Ng, E.Y., San Tan, R. and Acharya, U.R. (2018) 'Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats', *Computers in Biology and Medicine*, Vol. 102, No. 11, pp.278–287.

Ordóñez, F. and Roggen, D. (2016) 'Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition', *Sensors*, Vol. 16, No. 1, p.115.

Phua, C., Alahakoon, D. and Lee, V. (2004) 'Minority report in fraud detection: classification of skewed data', *ACM SIGKDD Explorations Newsletter*, Vol. 6, No. 1, pp.50–59.

Pinquet, J., Ayuso, M. and Guillén, M. (2007) 'Selection bias and auditing policies for insurance claims', *Journal of Risk and Insurance*, Vol. 74, No. 2, pp.425–440.

Rayana, S. and Akoglu, L. (2016) 'Less is more: building selective anomaly ensembles', *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 10, No. 4, pp.1–33.

Šubelj, L., Furlan, Š. and Bajec, M. (2011) 'An expert system for detecting automobile insurance fraud using social network analysis', *Expert Systems with Applications*, Vol. 38, No. 1, pp.1039–1052.

Subudhi, S. and Panigrahi, S. (2020) 'Use of optimized fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection', *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, No. 5, pp.568–575.

Sundarkumar, G.G., Ravi, V. and Siddeshwar, V. (2015) 'One-class support vector machine based undersampling: application to churn prediction and insurance fraud detection', in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, IEEE, pp.1–7.

Tao, H., Liu, H. and Song, X. (2012) 'Insurance fraud identification research based on fuzzy support vector machine with dual membership', in *2012 International Conference on Information Management, Innovation Management and Industrial Engineering*, IEEE, Vol. 3, pp.457–460.

Tzanos, G., Kachris, C. and Soudris, D. (2019) 'Hardware acceleration on Gaussian naive Bayes machine learning algorithm', in *2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST)*, IEEE, May, pp.1–5.

Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M. and Baesens, B. (2017) 'Gotcha! Network-based fraud detection for social security fraud', *Management Science*, Vol. 63, No. 9, pp.3090–3110.

Viaene, S., Dedene, G. and Derrig, R.A. (2005) 'Auto claim fraud detection using Bayesian learning neural networks', *Expert Systems with Applications*, Vol. 29, No. 3, pp.653–666.

Wang, F., Shan, G.B., Chen, Y., Zheng, X., Wang, H., Mingwei, S. and Haihua, L. (2020) 'Identity authentication security management in mobile payment systems', *Journal of Global Information Management*, Vol. 28, No. 1, pp.189–203.

Wang, J., Yu, L.C., Lai, K.R. and Zhang, X. (2016) 'Dimensional sentiment analysis using a regional CNN-LSTM model', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Short Papers, Vol. 2, pp.225–230.

Wang, Y. and Xu, W. (2018) 'Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud', *Decision Support Systems*, Vol. 105, No. 1, pp.87–95.

Weisberg, H.I. and Derrig, R.A. (1998) 'Quantitative methods for detecting fraudulent automobile bodily injury claims', *Risques*, Vol. 35, No. 3, pp.75–99.

Xu, W., Wang, S., Zhang, D. and Yang, B. (2011) 'Random rough subspace based neural network ensemble for insurance fraud detection', in *2011 Fourth International Joint Conference on Computational Sciences and Optimization*, IEEE, pp.1276–1280.

Xue, Z., Shang, Y. and Feng, A. (2010) 'Semi-supervised outlier detection based on fuzzy rough C-means clustering', *Mathematics and Computers in Simulation*, Vol. 80, No. 9, pp.1911–1921.

Yan, C., Li, Y. and Sun, H. (2017) 'Research on fraud identification of automobile insurance based on ant colony optimization random forest model', *Insurance Research*, Vol. 2017, No. 6, pp.114–127.

Yu, W., Feng, G. and Zhang, W. (2017) 'Research on fraud detection system and gang identification for motor vehicle insurance', *Insurance Research*, Vol. 2017, No. 2, pp.63–73.

Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R. (2017) 'Efficient KNN classification with different numbers of nearest neighbors', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 29, No. 5, pp.1774–1785.

Zhang, X. and Zhu, J. (2020) 'Problems and countermeasures of vehicle insurance fraud', *Bank of China Insurance News*, 2020-02-25 (006).

Zhang, X., Wang, D., Qian, Y. and Yang, Y. (2019) 'Prediction accuracy analysis with logistic regression and CART decision tree', in *Fourth International Workshop on Pattern Recognition*, International Society for Optics and Photonics, July, Vol. 11198, p.1119810.