Addressing data imbalance in insurance fraud prediction using sampling techniques and robust losses

Nhu-Tai Do[1†], Loc Dinh Tan[2†], Di Khanh Le[2], Quoc-Huy Nguyen[1*]

[1]Saigon University, Vietnam.
[2]University of Economics Ho Chi Minh City, UEH, Vietnam.

*Corresponding author(s). E-mail(s): nqhuy@sgu.edu.vn;
Contributing authors: dntai@sgu.edu.vn;
locdinh.31221020226@st.ueh.edu.vn;
khanhle.31211022918@st.ueh.edu.vn;
†These authors contributed equally to this work.

Abstract

Fraud in the auto insurance industry is a significant challenge for insurers globally. Policyholders engage in fraudulent activities like falsifying documents and creating fake evidence, leading to substantial financial losses for insurance companies. Traditionally, insurers have relied on financial examinations and machine learning models to detect fraud, which require extensive data processing. This study uses deep learning models, particularly convolutional neural networks, to tackle fraud detection. Using a dataset of 1,000 car collision claims from seven US states in 2015, we aim to demonstrate the effectiveness of deep learning, even with small datasets. Despite the limitations of small datasets, our deep learning models achieved performance comparable to traditional methods through rigorous efforts. This research supports sustainable development by promoting innovation and technological advancement (SDG 9) and contributes to fraud prevention, organizational integrity, and transparency in the insurance industry (SDG 16). Future work will focus on leveraging other advanced deep-learning techniques to enhance fraud detection further.

Keywords: fraud detection, sustainable development, deep learning models, imbalance problem

1 Introduction

In automobile insurance, various techniques detect fraudulent activities based on claims, which can occur skillfully or unskillfully [1]. The goal is to identify multiple frauds associated with behavioral changes, which can be challenging for traditional machine learning approaches [2]. Conventional methods for detecting insurance fraud using financial assessment techniques are less effective when dealing with imbalanced data and fail to consider the significance of outliers in the data [3]. Additionally, traditional machine learning models face significant difficulties during data processing and feature selection, especially with highly complex datasets such as those involved in fraud detection. These methods often yield suboptimal results for minority groups.

To optimally address the fraud detection problem, we propose employing deep learning models instead. Deep learning models can overcome the limitations of the aforementioned traditional methods. However, these models must undergo training and evaluation on imbalanced and low-observation datasets to be genuinely effective in fraud detection. In deep learning, convolutional neural networks (CNNs) have achieved remarkable success and are widely applied in image classification tasks, achieving significant milestones. Therefore, this study focuses on demonstrating the feasibility of CNN models compared to traditional methods for tabular data.

Data generation and augmentation techniques have been explored to address the data imbalance issue. Oversampling has proven effective in increasing accuracy for minority groups, resulting in higher precision and fewer false negatives [4]. In addition to directly addressing imbalance through data augmentation techniques, we aim to enhance the performance and efficiency of deep learning models through experiments with loss functions commonly used to address imbalance issues. Well-known loss functions such as focal loss and dice loss are reputed to resolve imbalance issues in most computer vision cases effectively.

Experiments use the SMOTE data generation technique and multi-loss functions to enhance the effectiveness of deep learning models in classifying fraudulent contracts. The VGG16 model, combined with these techniques, outperforms traditional models like Decision Tree and Random Forest, demonstrating their versatility across various scenarios.

2 Materials and Methods

2.1 Fraud Detection Dataset

Based on the dataset of automobile insurance compensation claims [5], our study evaluated our approach against significant challenges. With 40 variables and 1000 observational samples, the dataset was geared towards distinguishing between valid and fraudulent insurance contracts. However, our analysis unearthed several hurdles that underscored the complexity of the task. Firstly, the imbalance between the data labels emerged as a prominent issue. We noted a substantial disproportion, with a marked abundance of valid insurance contracts (753 cases) compared to their fraudulent counterparts (247 cases). This imbalance posed a formidable obstacle in effectively training our model, as it hindered its ability to discern fraudulent cases amidst the

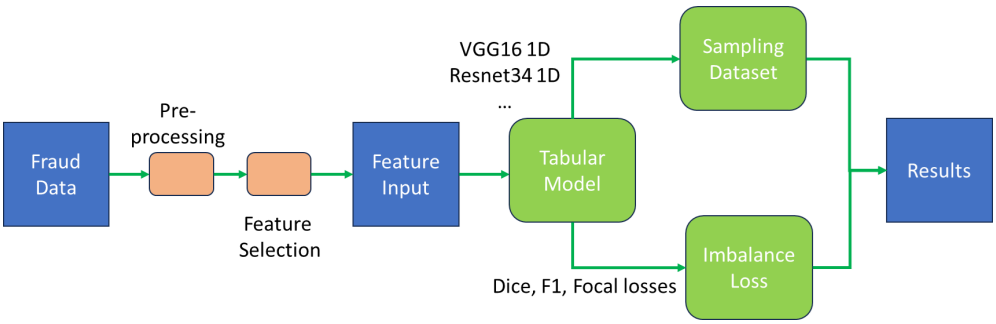

Fig. 1: Overall proposed method

overwhelming majority of valid contracts. Moreover, our examination revealed a lack of meaningful correlation between certain variables within the dataset. This absence of correlation complicated our efforts to identify underlying relationships and patterns, impeding the model's predictive capabilities.

The study faced challenges due to the variability in compensation claims distribution and the small size of the dataset. The data was subjected to rigorous analysis and preprocessing to ensure its quality and suitability for machine learning models. This approach helped navigate the complexities of the dataset and develop robust methodologies for effectively classifying insurance contracts.

2.2 Problem Overview

Research consistently indicates that most solutions for insurance fraud detection stem from traditional models. Decision trees and random forests are widely used and highly effective models because they can classify based on data characteristics, thereby mitigating the impact of underlying data issues [6]. These models have been favored for their high accuracy in most fraud detection scenarios and relatively low implementation complexity. However, they may encounter challenges with complex data processing and less effectiveness with datasets containing numerous observation samples.

The novel aspect of our research lies in extending the scope to include deep learning models for insurance fraud classification tasks. Our exploration will progress from simple architectures, such as MLPs, towards more complex ones, focusing on deep learning CNN models. This expansion aims to leverage the potential of deep learning in handling intricate data patterns and enhancing fraud detection accuracy.

The complexity inherent in deep learning model architectures, particularly CNNs, may diminish efficacy when applied to tabular data, especially in scenarios where the input data is sparse. Across most cases of imbalance, interventions aimed at rebalancing sample proportions between classes, notably in binary classification tasks, have demonstrated positive effects, enhancing the predictive capabilities of the models. Moreover, in the realm of image classification, the effectiveness of loss functions in combating issues stemming from imbalanced data has been well-established. Therefore, we will integrate data processing strategies and imbalance handling techniques

alongside refining the architectures of the research entities in this study to suit tabular data shown in Fig. 1.

2.3 Proposed Model

In this section, we will list the experimental subjects of the study, including both traditional machine learning models and deep learning models. We will then focus on a detailed analysis of the architecture of each group of deep learning models, concentrating on explaining their suitability for tabular data, along with an in-depth discussion of sampling techniques and loss functions.

As previously mentioned, the CNN model is renowned in deep learning and has achieved numerous milestones in computer vision [7]. Consequently, the input data for the model is typically a three-dimensional array (samples, features, channels), often consisting of images, videos, and similar formats. The CNN models employed in this study will be fine-tuned and reconstructed based on several well-known architectures, such as VGG16 and ResNet34, which will be detailed in the following sections. We anticipate that the CNN models will yield positive results with the dataset used in this study.

An in-depth analysis reveals that selecting the appropriate architecture alone is insufficient to address this problem thoroughly; combining it with various techniques and methods, specifically sample adjustment techniques and related procedures, is necessary. Building on previous efforts and recent investigations, we will employ a combination of class weight calculations and the Synthetic Minority Over-sampling Technique (SMOTE) to address the issue.

Furthermore, we propose that an appropriate loss function can enhance the predictive capabilities of deep learning models. Previous studies in image classification have demonstrated the effectiveness of loss functions in mitigating issues caused by imbalanced data. Our experimental strategies will be built upon applying several commonly used segmentation loss functions, such as dice loss, focal loss, and f1 loss, enabling us to rank their effectiveness in improving the performance of our study.

By implementing these strategies, we aim to enhance the model's ability to handle imbalanced tabular data and improve its overall performance in insurance fraud detection.

2.4 Implementation Details

Our research explores deep learning models, ranging from simple Multi-Layer Perceptrons (MLPs) to more complex models like Transformers and Convolutional Neural Networks (CNNs), which are applied to our study.

MLP Model: For tabular data, complex architectures are often suboptimal. We constructed an MLP network with an input layer (features, 1) and superficial Dense layers using ReLU activation functions. The output is a classification layer with a SoftMax classifier. This model was built and adjusted to evaluate its classification ability in addressing the issues encountered in the data.

Transformer Model: TabNet, designed explicitly for tabular data, combines attention mechanisms and decision trees, leveraging the strengths of both traditional

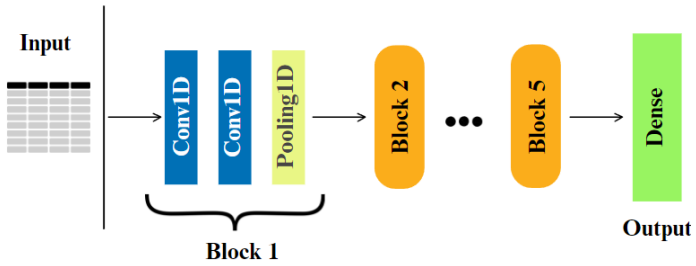

Fig. 2: Model architecture designed based on VGG16 for tabular data.

and modern methods [8]. Due to TabNet's superior capabilities, we aim to assess how data imbalance impacts this specialized model.

CNN Model: CNNs are renowned in deep learning, excelling in computer vision with 3D input arrays (samples, features, channels). To adapt CNNs for tabular data, we redesigned classical architectures with 1D convolution functions and pooling layers, adjusting the intrinsic parameters [9]. Since the study focuses on evaluating CNN models, we redesigned and fine-tuned several renowned deep learning models based on modern techniques such as VGG16 in Fig. 2, ResNet34, ResNet50, Inception V3, and Inception V2 + ResNet50.

In addition to data manipulation through sampling adjustment techniques, we recognized that applying certain renowned loss functions could mitigate the effects of data imbalance. Therefore, we experimented with different loss functions to evaluate the model's improvement, aiming to derive the most objective and accurate conclusions regarding these loss functions.

F1 Loss: Designed to optimize the F1 score, balancing precision and recall harmoniously. This loss function is handy in highly imbalanced data scenarios, where one of these metrics might be overlooked when using traditional loss functions like cross-entropy loss. F1 Loss balances precision and recall, making it suitable for situations where detecting both positive and negative instances is crucial, such as in fraud contract classification.

$$\text{F1 Loss} = 1 - \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{1}$$

Focal Loss: Designed to address data imbalance by reducing the weight of easy samples and increasing the weight of more complex samples. This focus helps the model concentrate on samples that are more difficult to classify correctly. The model relies on the class weight ratio, making it highly suitable for the dataset discussed in this study. Focal Loss mitigates bias towards dominant classes and enhances performance on less frequent classes [10].

$$\text{Focal Loss} = -(1 - p_t)^{\gamma} \log(p_t) \tag{2}$$

where p_t is the predicted probability for the target class and γ is a focusing parameter.

Dice Loss: Originating from the Dice coefficient, this loss function is commonly used in medical image segmentation but is also effective for imbalanced tabular data

[11]. It measures the similarity between sets, optimizing the Dice coefficient directly. This innovative approach in our research involves applying a loss function traditionally used in computer vision to evaluate its effectiveness on tabular data.

$$\text{Dice Loss} = 1 - \frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{3}$$

where X and Y are the predicted and ground truth sets.

Multi Loss: A combination of various loss functions to leverage the unique advantages of each. Multi Loss provides a comprehensive approach, allowing the model to learn from multiple aspects of the data and minimizing the weaknesses of individual loss functions [12].

$$\text{Multi Loss} = \alpha \times \text{Loss}_1 + \beta \times \text{Loss}_2 + \gamma \times \text{Loss}_3 \tag{4}$$

where α, β, γ are weighting parameters.

3 Experiment and Results

3.1 Experiments Setup

Training Process: We divided the dataset into two parts, comprising test data and training data, with a ratio of 66.7/33.3. Subsequently, we applied several strategies, as previously mentioned, to address the imbalance issue, such as weight balancing combined with the use of the SMOTE method. Next, we segmented the strategy to construct and enhance the model into two distinct processes: the traditional machine learning models and the deep learning models we had previously proposed.

RFE, Lasso, and Adam were used in feature selection and training for deep-learning models. The AUC was then used to evaluate the performance of the models. The process involved fine-tuning parameters to select the best-performing traditional model and refine the parameters for optimal results.

The deep learning models were restructured and optimized for specific problems. The Tabnet model was built using Keras' TabNet library, while the VGG16 model was adapted from VGG16's structure. Strategies were set for each step, including data normalization, output activation function, and loss functions. Callback functions were employed for efficient learning. Results were visualized and monitored using Wandb.

3.2 Results

3.2.1 Results on traditional models experiments

We conducted cross-validation using K-fold and then analyzed the model performance through the metrics mentioned: accuracy, F1 score, and the AUC curve.

The results shown in Fig. 3 on accuracy and f1 score in testing are visualized using Boxplot charts, which help us understand the metrics and the variability of each model. The baseline models were selected based on performance criteria and had high stability. Consequently, models such as Decision Tree (DTC), and Random Forest (RF) are appropriate choices for the group of baseline models. We aim to compare whether using deep learning models for fraud prediction is feasible.

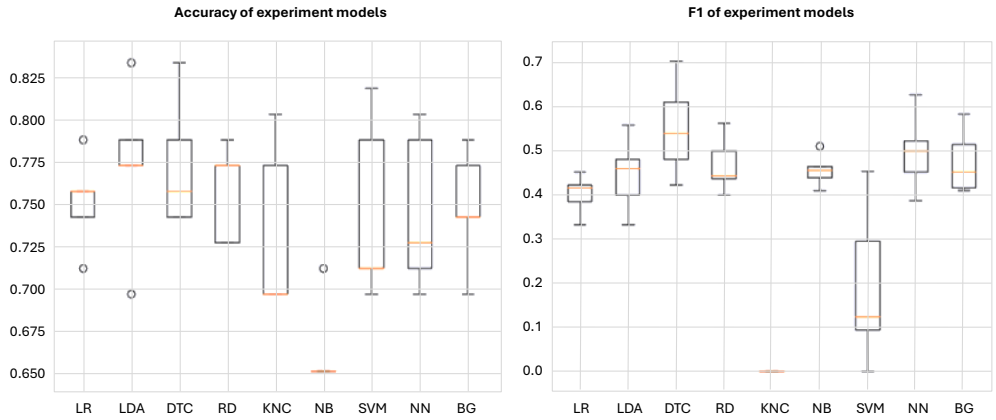

Fig. 3: Accuracy and F1-score of traditional models. .

The SVM model, initially used for insurance fraud detection, has shown low performance due to class imbalance. To improve performance, the model's hyperparameters were fine-tuned using the Grid Search technique.

3.2.2 Results on deep learning models experiments

As mentioned earlier, the issue of data imbalance has significantly impacted the performance of our problem-solving efforts. Initially, our deep learning MLP models seemed to achieve an impressive performance rate of 73.7%. However, other metrics indicated that the model's predictions were biased towards the majority class. This can be attributed to the relatively simple structure of the model, which is heavily influenced by the data imbalance issue. More importantly, a deeper evaluation of other relevant metrics revealed a meager f1 score and an AUC of only 0.5, indicating that the model had no discriminative power. This provides a deeper insight into the impact of the dataset on the model's learning capability.

The study evaluated deep learning models like TabNet and VGG16 in Table 1, showing higher performance. However, data imbalance affected their AUC scores below 0.7. To address this, the SMOTE technique was applied during data processing, significantly improving the models.

Table 1: Result of deep learning models experiments

	Base training[1]		Base training + class weight + SMOTE[2]	
Model	Acc	AUC	Acc	AUC
MLPs model	73.7%	0.5	67.28	0.65
Tabnet	64.19%	0.60	79.32%	0.75
VGG16	70.16%	0.65	73.77%	0.69

[1]No data imbalance, [2] SMOTE + class weights

The TabNet model's classification capability is excellent due to its specialized architecture and attention mask mechanisms. Despite the significant impact of data on the model's classification ability, it still demonstrated the capability to make predictions when applied to tabular data, validating the hypothesis of using CNN models for fraud detection.

Table 2: Results of the CNN models applying imbalance handling strategies.

	Base training[1]		Base training[1]+ SMOTE	
Model	Acc	AUC	Acc	AUC
VGG16	72.22%	0.68	73.77%	0.69
ResNet34	69.75%	0.65	71.91%	0.68
ResNet50	69.44%	0.65	70.67%	0.65
Inception V2[2]	69.14%	0.64	74.07%	0.65
Inception V3	74.07%	0.68	74.38%	0.64

[1]This experiment applies class weights to train model. [2] InceptionV2 + ResNet50

It can be observed that among the CNN models discussed in Table 2, the VGG16 model demonstrated the best classification performance, with the AUC reaching the highest level and showing improved classification ability after applying various data strategies. This can be attributed to the relatively more straightforward architecture of VGG16 compared to other CNN models while still maintaining the intrinsic characteristics of convolutional networks.

Therefore, the strategies will be focused on and revolve around the VGG16 model to present the experimental results more smoothly. The results in Table 3 presented below illustrate the changes in the evaluation metrics of the VGG16 model with different experimental strategies and loss functions.

Table 3: Accuracy and AUC of VGG16 models through situations

Model	Loss	SMOTE	Acc	AUC
VGG16			72.22%	0.68
VGG16		x	73.77%	0.69
VGG16	focal loss	x	70.06%	0.69
VGG16	f1 loss	x	77.16%	0.71
VGG16	dice loss	x	79.32%	0.75
VGG16	multi loss[1]	x	74.69%	0.73
VGG16	multi loss[2]	x	75%	0.61

[1]F1, dice losses. [2]F1, Focal and Dice losses.

The VGG16 model shows significant improvements when using various loss functions, including f1 loss, focal loss, and dice loss. Focal loss adjusts class weights, balanced f1 loss harmonizes precision and recall, and dice loss boosts metrics. These functions improve classification capability, precision, and recall, making them effective in medical image segmentation and fraud detection.

This led us to investigate whether combining multiple loss functions could further improve the VGG16 model. However, experimental results indicated that incorporating multiple loss functions into the model reduced VGG16's learning capability, suggesting that the model had reached a saturation point.

In the process of extending to other deep learning model groups, we obtained two significant findings. First, VGG16 demonstrated the highest classification capability and performance among the CNN models, likely due to its well-suited architecture. Second, applying loss functions such as Dice loss and Focal loss not only failed to enhance performance but also reduced the classification capability of the model. This indicates that these loss functions are incompatible with non-CNN models, particularly those with architectures and attention mechanisms like TabNet. The results in Table 4 of the models are organized in ascending order of architectural complexity.

Table 4: Evaluation metrics of all experiment models

Model	Loss	Acc	AUC
Decision Tree		81.3%	0.83
Random Forest		80.8%	0.77
SVM		75%	0.71
Tabnet	categorical loss	80.12%	0.76
MLPs	dice loss	72.53%	0.71
VGG16	dice loss	79.32%	0.75
ResNet34	dice loss	75.93%	0.73
ResNet50	dice loss	75.93%	0.7
Inception V3	dice loss	70.37%	0.7
Inception V2 + ResNet50	dice loss	70.67%	0.71

The study demonstrates that applying CNN models to fraud detection is feasible and promising, with strong classification capabilities and stability comparable to traditional model groups. Integrating these models into tabular data can yield promising results, demonstrating the potential of advanced techniques.

4 Conclusions

Despite numerous challenges, the research team's relentless efforts have yielded commendable results. Deep learning models have shown the ability to boost productivity by overcoming traditional machine learning limitations and maintaining stability and accuracy. The detailed findings in the report highlight how applying well-known loss functions from computer vision enhances CNN model performance, affirming deep learning's viability in predicting fraudulent activities. Future endeavors will explore and implement advanced deep learning techniques to improve problem-solving efficiency and broaden dataset scopes for enhanced research objectivity.

References

[1] Kini, A., Chelluru, R., Naik, K., Naik, D., Aswale, S., Shetgaonkar, P.: Automobile

insurance fraud detection: An overview. In: 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), pp. 7–12 (2022). IEEE

[2] Caruana, M.A., Grech, L.: Automobile insurance fraud detection. Communications in statistics: case studies, data analysis and applications **7**(4), 520–535 (2021)

[3] Ali, A., Shamsuddin, S.M., Ralescu, A.L.: Classification with class imbalance problem. Int. J. Advance Soft Compu. Appl **5**(3), 176–204 (2013)

[4] Zarzà, I., Curtò, J., Calafate, C.T.: Optimizing neural networks for imbalanced data. Electronics **12**(12), 2674 (2023)

[5] Abdelhadi, S., Elbahnasy, K., Abdelsalam, M.: A proposed model to predict auto insurance claims using machine learning techniques. Journal of Theoretical and Applied Information Technology **98**(22), 3428–3437 (2020)

[6] Viaene, S., Derrig, R.A., Baesens, B., Dedene, G.: A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. Journal of Risk and Insurance **69**(3), 373–421 (2002)

[7] Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., Ghayvat, H.: Cnn variants for computer vision: History, architecture, application, challenges and future scope. Electronics **10**(20), 2470 (2021)

[8] Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.: Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems (2022)

[9] Kashyap, S.K., Mahalle, P.N., Shinde, G.R.: Human activity recognition using 1-dimensional cnn and comparison with lstm. In: Sustainable Technology and Advanced Computing in Electrical Engineering: Proceedings of ICSTACE 2021, pp. 1017–1030. Springer, ??? (2022)

[10] Yun, P., Tai, L., Wang, Y., Liu, C., Liu, M.: Focal loss in 3d object detection. IEEE Robotics and Automation Letters **4**(2), 1263–1270 (2019)

[11] Hashemi, S.R., Salehi, S.S.M., Erdogmus, D., Prabhu, S.P., Warfield, S.K., Gholipour, A.: Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection. IEEE Access **7**, 1721–1735 (2018)

[12] Du, Z., Zhang, H., Wei, Z., Zhu, Y., Xu, J., Huang, X., Yin, B.: Merge loss calculation method for highly imbalanced data multiclass classification. IEEE Transactions on Neural Networks and Learning Systems (2023)