Journal of Management Science and Engineering 7 (2022) 589–607

Contents lists available at ScienceDirect

Journal of Management Science and Engineering

journal homepage: www.keaipublishing.com/en/journals/journal-of-management-science-and-engineering/

Refined analysis and a hierarchical multi-task learning approach for loan fraud detection

Liao Chen [a], Ning Jia [a], Hongke Zhao [a,*], Yanzhe Kang [b], Jiang Deng [c], Shoufeng Ma [a]

[a] College of Management and Economics, Tianjin University, Tianjin, 300072, China
[b] Postdoctoral Research Station, Guosen Securities Co. Ltd., Shenzhen, 518001, China
[c] Beijing Fantaike Technology Co. Ltd., Beijing, 100012, China

ARTICLE INFO

Article history:
Received 3 May 2021
Received in revised form 5 December 2021
Accepted 1 June 2022
Available online 3 July 2022

Keywords:
Loan application
Fraud detection
Information falsification
Multi-task learning

ABSTRACT

Fraud problems in loan application assessment cause significant losses for finance companies worldwide, and much research has focused on machine learning methods to improve the efficacy of fraud detection in some financial domains. However, diverse information falsification in individual fraud remains one of the most challenging problems in loan applications. To this end, we conducted an empirical study to explore the relationships between various fraud types and analyzed the factors influencing information fabrication. Weak relationships exist among different falsification types, and some essential factors play the same roles in different fraud types. In contrast, others have various or opposing effects on these types of frauds. Based on this finding, we propose a novel hierarchical multi-task learning approach to refine fraud-detection systems. Specifically, we first developed a hierarchical fraud category method to break down this problem into several subtasks according to the information types falsified by customers, reducing fraud identification's difficulty. Second, a heterogeneous network with a meta-path-based random walk and heterogeneous skip-gram model can solve the representation learning problem owing to the sophisticated relationships among the applicants' information. Furthermore, the final subtasks can be predicted using a multi-task learning approach with two prediction layers. The first layer provides the probabilities of general fraud categories as auxiliary information for the second layer, which is for specific subtask prediction. Finally, we conducted extensive experiments based on a real-world dataset to demonstrate the effectiveness of the proposed approach.

1. Introduction

In recent years, consumer finance has become an indispensable part of society and its penetration rate has risen rapidly. The consumer credit amount reached 13.91 trillion by the end of 2019 (*China Financial Stability Report 2020*, 2020), a 135% increase compared to 2015. Fraud is one of the most significant challenges for consumer finance companies. The average default rate of consumer credit loans was 2.63% in 2019, slightly higher than that for credit cards (Consumer Finance Committee of China

* Corresponding author.
E-mail address: hongke@tju.edu.cn (H. Zhao).

https://doi.org/10.1016/j.jmse.2022.06.001

Banking Association, 2020). Wang et al. (2006) defined financial fraud as "a deliberate act that is contrary to law, rule, or policy with the intent to obtain an unauthorized financial benefit." Ngai et al. (2011) categorized financial fraud into four types: bank fraud, insurance fraud, securities and commodities fraud, and other related financial fraud. Financial institutions involving money and services in these areas face serious fraud events that may cause great losses. There has been extensive research on fraud detection; however, the target products of current studies are undiversified. Many studies have proposed data mining methods for credit card fraud in banks (Bhattacharyya et al., 2011; Duman and Ozcelik, 2011; Olszewski, 2014; Quah and Sriganesh, 2008), healthcare insurance (Francis et al., 2011; Mailloux et al., 2010; Ortega et al., 2006), automobile insurance (Brockett et al., 2006; Wen et al., 2005), and online auctions (Almendra, 2013; Shah et al., 2002). Ngai et al. (2011) classified the data mining techniques involved into six categories: classification, clustering, prediction, outlier detection, regression, and visualization.

To the best of our knowledge, there have been few studies on detecting information falsification in loan fraud, which is one of the most critical problems that results in bad lending. There are several challenges to detecting information falsification in loan fraud. First, the diversified fake information falsified in loan applications leads to a failure in using a single standard detection technique. For example, customers must submit marriage, occupation, and social relationship information for assessment, but it is possible to fabricate all of this information. Second, owing to the explosive growth of data and information, it is challenging to find useful information from a large amount of data to characterize fraudulent customers and their behaviors, which increases the difficulty of data analysis and feature engineering. Finally, the imbalance problem resulting from the relatively small number of fraud samples leads to low accuracy and poor performance of the detection system. Specifically, standard classifiers are suitable for balanced datasets, and when facing imbalanced cases, these techniques may only cover the majority of the samples, and the minority examples are distorted (López et al., 2013).

Therefore, we conducted an empirical study to analyze the relationships among different falsification types and the impact of various types of information on different fraud behaviors. Specifically, we found weak connections among various fabrication types using label analysis. Furthermore, we employed binary logistic regression to explore the impact of the essential elements on different fabrication behaviors. The results show that some factors play the same role in all fraud types, while others have different or even opposing effects on various tasks, which means that these problems can neither be solved by one unified method nor independently.

Based on the empirical study's conclusion, a hierarchical multi-task learning (HMTL) approach is proposed to refine loan application fraud detection. Specifically, we propose a hierarchical fraud classification method to classify fraud customers and divide the fraud detection task into several subtasks, after which a more specific method is considered for each task. In the first level, all fraud is categorized as either individual information fraud or social relationship fraud. In the second level, fraud is classified into more specific categories. Furthermore, a heterogeneous graph network generates representation learning of various customer information. Here, the main reason for utilizing a heterogeneous graph network is to learn the representations of feature relations and explore hidden information in the data. Finally, the HMTL predicts multiple fraud-detection tasks using two prediction layers. The first is for general-type prediction, that is, whether the customer has committed individual information fraud or social relationship fraud. In this layer, an oversampling method is utilized to solve the imbalance problem. Specifically, different fraud types are combined to obtain general fraud-type samples, which increases the sample size of the falsification application. This method increases the minority class samples (He and Garcia, 2009) and eliminates the negative impact of skewed distributions (Chawla et al., 2002) in the learning process. Then, the prediction result is used as auxiliary information in the second prediction layer to make the final task prediction easier. In the second prediction layer, specific fraud-detection tasks are predicted through several shared hidden and task-specific layers. Extensive experiments on an auto loan dataset provided by a Chinese automobile finance company were conducted to evaluate the performance of HMTL, and the experimental results clearly show the effectiveness of HMTL compared with several state-of-the-art methods.

In summary, the main contributions of this study can be summarized as follows. First, we conducted empirical research to study the relationships among fabrication types and adopted a binary logistic regression model to explore factors that affect varied falsification. The results suggest weak relevance among fraud categories, and many factors perform distinctly in various identification tasks. Second, based on the empirical results, we propose detecting loan application fraud by decomposing complex-component fraud into a two-layer classification system. A novel HMTL approach was designed, including a heterogeneous graph network for feature embedding and two prediction layers. Finally, the proposed approach is evaluated using a real-life dataset. The results show that HMTL outperforms other state-of-the-art methods in various aspects.

The remainder of this paper Is organized as follows: Section 2 reviews the relevant literature on loan fraud and detection techniques, Section 3 describes the data and develops an empirical analysis, Section 4 presents the proposed detection method, Section 5 provides the experimental results and discussion. Finally, Section 6 concludes the paper.

2. Literature review

In this section, we summarize related research to explore the factors influencing loan fraud and discuss fraud-detection techniques.

2.1. Influencing factors for loan fraud

Loan fraud leads to significant losses in consumer finance, and there are three main elements in a financial fraud detection system: data, rule-based strategies, and artificial intelligence models (*Intelligent risk control: principles, algorithms and*

practice, 2020). We explore the important factors influencing loan fraud from previous studies to find advantageous data for detection. These data usually consist of basic information, behavioral information, and other external information for risk assessment and management (*Intelligent risk control: principles, algorithms and practice, 2020*). Hartmann-Wendels et al. (2009) proposed that fraud risk is highly related to demographic and socioeconomic data, including gender, marriage, age, and occupation. Dorfleitner and Jahnes (2014) identified determinants for fraudulent applications, including gender, marital status, and age, and they considered product information, such as loan amount, in a fraud risk management framework. Many studies have adopted a vast amount of data to develop risk control methods. Wheeler and Aitken (2000) employed identity information, including names and addresses, in credit card fraud detection. Hamid and Ahmed(2016) selected important information for a loan risk detection model, including basic demographic information (e.g., gender and age), loan information (e.g., loan purpose), and job information (e.g., occupation type).

2.2. Fraud detection techniques

Traditional rule-based models have been widely used in many fraud-detection systems as the main strategy, and they are still the basic element. Leonard (1995) designed a rule-based system with expert experience, and Sánchez et al. (2009) proposed association rules for fraud detection (Shah et al., 2002).

In recent years, with the development of data-driven technology, more advanced data mining methods have been proposed for fraud detection systems. Bolton and Hand (2002) classified statistical fraud detection methods into supervised and unsupervised learning approaches. Specifically, supervised methods refer to situations in which fraudulent and non-fraudulent samples are utilized to establish models; thus, a new observation can be assigned to one of the two groups. Unsupervised methods do not require labeled data and seek to distinguish between the different behavior patterns of the samples. Many studies have focused on fraud detection using supervised approaches including logistic regression (Awoyemi et al., 2017; Sahin and Duman, 2011), decision trees (Almendra, 2013; Baesens et al., 2003; Bhowmik, 2011; Mailloux et al., 2010), support vector machines (Bhattacharyya et al., 2011; Francis et al., 2011; Ravisankar et al., 2011), artificial neural networks (Brockett et al., 2006; Kirkos et al., 2007; Ortega et al., 2006; Quah and Sriganesh, 2008), and genetic algorithms (Duman and Ozcelik, 2011). Additionally, many studies have used unsupervised methods, such as a self-organizing map algorithm (Olszewski, 2014; Zaslavsky and Strizhak, 2006) and outlier detection methods (Malini and Pushpa, 2017). These methods are mainly employed in credit card, insurance, and other financial domains to identify fraudulent behavior; however, when facing complicated characteristics and interconnections among various information falsification behaviors in loan applications, performance is poor, especially in problems with massive amounts of data. Considering the credit card fraud detection problem as an example: Although credit card transactions and loan applications are both financial transaction behaviors, there are many differences between them in the frequency of transactions and the attributes considered to assess transactions (Dal Pozzolo et al., 2017). Specifically, credit card transaction frequency is much higher than that of loan applications, which helps accumulate data on customer behavior. Moreover, the timeline to approve loan application transactions is much longer than that for credit cards, which results from more complex material and data to assess. Thus, we must also consider more complete attributes in our detection technique.

In this study, we introduce multi-task learning (MTL) approach, which has achieved great success in many industrial applications. In machine learning tasks, the core idea of MTL is to provide inductive bias with auxiliary tasks so that the model will prefer a solution that chooses hypotheses to explain more problems, which contributes to a more generalized solution (Ruder, 2017; Zhao et al., 2019). This framework is widely used in natural language processing (Collobert and Weston, 2008), speech recognition (Deng et al., 2013), computer vision (Yu et al., 2020), etc. and has also received considerable attention in new scenarios such as online gaming (Zhao et al., 2022) and entrepreneur fund-raising (Zhao et al., 2022). In the finance domain, some research has utilized a multi-task architecture to predict the volatility of financial assets (Yang et al., 2020) or stock prices (Sawhney et al., 2020) to harness multimodal data such as text and audio data. Shao et al. (2022) applied MTL to identify potential portfolio clients by extracting valuable features from diverse user characteristics. Kang et al. (2022) adopted a conditional Wasserstein generative adversarial network with a gradient penalty (CWGAN-GP)-based multi-task learning model to solve the imbalance problem in credit scoring. However, to the best of our knowledge, there have been few studies on MTL in fraud detection. We employ this framework in our study because there are interactions and differences among the various fake information types in detecting fraud. One of the most crucial steps in MTL is to tune the task weightings, and the most common approach is to use uniform or manually tuned loss weights (Eigen and Fergus, 2014; Kokkinos, 2017; Sermanet et al., 2013). However, these weight parameters are costly to train, and a considerable amount of time is often required for trials. Thus, we employed homoscedastic uncertainty (Kendall et al., 2018) to balance the task weightings. In addition, using a related task as an auxiliary task in MTL is a classical optimization method. For example, Osman and Sierra (2016) predicted the features of roads as auxiliary tasks and forecasted the steering direction in a self-driving task. Cheng et al. (2015) predicted the existence of a name in a sentence to identify name errors. In this study, we use general fraud-type identification as the auxiliary task for specific fabrication detection.

Additionally, we designed a heterogeneous network for representation learning in feature engineering. Representation learning is a tool for extracting useful information from data for downstream tasks and provides crucial inputs for task prediction (Bengio et al., 2013). Many studies have focused on matrix factorization methods (Hoff et al., 2002; Lei et al., 2011; Neville and Jensen, 2005) that generate latent dimension variables; however, these methods require extensive computational costs to decompose a large matrix. Other studies have focused on neural-network-based representation learning models.

Perozzi et al. (2014) proposed DeepWalk, which utilizes random walkers to record walking paths over networks and regards the chain of nodes as a sentence and then employs the word2vec framework (Mikolov et al., 2013) to learn the representations of words. Grover and Leskovec (2016) introduced a biased random walk procedure that explores diverse neighborhoods using a mixture of breadth-first and width-first search procedures. To represent the heterogeneous structural network in our study, we used metapath2vec (Dong et al., 2017) for various types of nodes in the network.

3. Data and analysis

This section describes a preliminary empirical study of loan application fraud detection. We introduce the data and variables used in this study, followed by the analysis methods, and then we present the empirical results, including the falsification type correlation and performance of different factors in various loan fraud cases.

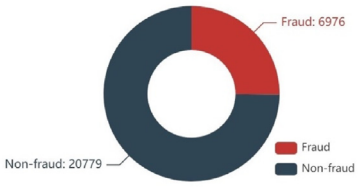
3.1. Data and variables

3.1.1. Data source and labeling

There is no public dataset available for studying fake information detection in loan application assessments. Obtaining real data from financial companies is extremely difficult, and it is costly for companies to complete data labeling, that is, to label customers with fake information (Phua, 2010). We obtained a dataset from a Chinese automobile finance company as the data source, which encompasses 31 provinces and more than 500 cities across the country. The dataset consists of 27,755 loan applications from March 19, 2015, to November 27, 2020, and 6976 of the applications involve information falsification, as shown in Fig. 1(a). It should be noted that we selected all the fraud samples and part of the non-fraud samples from all the application data to improve computing efficacy. The company completed data labeling mainly through three methods—phone review, home visits, and third-party data verification—to check whether fake information existed during loan application assessment.

This study focused on four common types of fake information, and we discuss their characteristics and some traditional identification methods below. Note that more than one type of falsification can exist in an application, and Fig. 1(b) shows the distribution of applications with different fabrication types.

- Fake occupation information: Applicants with fake occupation information often falsify their occupation, working units, monthly income, etc. Loan reviewers commonly check the authenticity of job information by contacting applicants to query the working information. They may lack relevant professional knowledge or the capacity to verify their information. For instance, if the applicant claims that he/she serves as a salesman in an electrical appliance store,



(a) Number of fraud and non-fraud applications

(b) Number of applications with different types of fraud

Fig. 1. Number of fraud and non-fraud applications, and number of applications with different types of fraud.

the reviewer may ask the applicant what brands of refrigerators they have and then make a judgment based on the applicant's reply, reaction, and even tone.

- Fake ability information: Applicants provide fake ability certifications or information to pretend to be high-quality customers, such as income certificates or driving ability. Taking the income certificate as an example, reviewers may check whether the transaction and saving amounts that reflect the applicant's consumption behavior are in accordance with his/her profile because people with lower income may have lower transaction and saving amounts. In this study, we used auto loan application data and used driving ability to estimate customers' related abilities.
- Fake marriage information: Applicants may show false marriage information, including their spouse's basic information and working information. They want to show a stable marriage relationship, because a good family relationship represents a lower probability of default and fraud. Specifically, they may offer incorrect information regarding a spouse's occupation and income, or sometimes they may even fabricate a spouse; however, a false spouse may not be familiar with the applicant's or other family members' information. Loan reviewers often contact the spouse and query simple but helpful questions, such as children's names and the marriage date for evidence.
- Fake contact information: Applicants may falsify contact information, including the contact relationship and information of the contact person. For instance, they may add their friend as a contact person but claim another type of relationship (such as a cousin) to show a closer link, which increases the approval probability. Reviewers can call them and judge this based on the responder's reaction and reply.

Most fraud applications have one or two types of fake information, and we present the number of each falsification type and application with two false information types in Fig. 2.

3.1.2. Variables

The definitions of the main dependent variables are listed in Table 1, which are widely used for fraud detection. In particular, basic demographic information, including sex, age, education, and household type, can be used to construct a customer profile illustrating a person's basic characteristics. Furthermore, loan amount, interest, term, and down payment ratio are related to the loan product information. Additionally, the working year and monthly income show customers' occupation information, and the driving year is related to driving ability. Moreover, marital status and the relationship with the contact person can reflect a customer's social relationships, and the consistency between different addresses focuses on physical distance relationships.

3.2. Analysis methods

We first conducted a correlation analysis of various fake types using Cramer's V-value and then utilized logistic regression to analyze the essential elements influencing falsification behaviors. In this problem, we define $x_1, \ldots, x_n$ as the explanatory variables mentioned above and denote a binary dependent variable Y representing whether the customer has fabrication behavior. Here, P is the conditional probability that $Y = 1$ (i.e., that fake information exists). Thus, we can express P as:

$$P(Y|x_1, \ldots, x_n) = \frac{e^{b_0+b_1x_1+\ldots+b_nx_n}}{1 + e^{b_0+b_1x_1+\ldots+b_nx_n}} \tag{1}$$

where $b_0, b_1, \ldots, b_n$ represent the estimated coefficients.

Finally, the factors' performances were analyzed for the four fraud types, and a predictive model was developed based on the empirical conclusion.

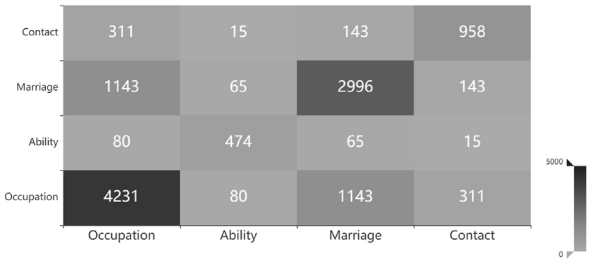


Fig. 2. Number of each falsification type and application with two false information types. The heat map is "square," with the same items shown in the rows and columns. The cells along the main diagonal show the number of applications of each falsified information type. The others show the number of applications for two types of fake information. This map is symmetrical, with the same value being shown above the main diagonal as a mirror image of those below the main diagonal.

Table 1
Description and statistical summary for empirical variables.

Variable	Description	Type	Obs	Mean	SD	Max	Min
Age	Age(log).	Numerical	27755	3.52	0.26	4.19	2.89
Amount	Amount of loan application(log).	Numerical	27755	11.24	0.69	14.91	3.37
interest	Interest ratio of loan.	Numerical	27755	0.12	0.06	9.00	0.00
term	Loan term(log).	Numerical	27755	3.55	0.25	4.09	1.79
down_pay_ratio	Down payment ratio.	Numerical	27755	0.29	0.11	1.00	0.15
work_year	Working year(log).	Numerical	27755	1.25	0.81	4.25	−1.00
income	Monthly income(log).	Numerical	27755	9.19	0.89	15.61	−1.00
dirving_year	Driving year(log).	Numerical	27755	0.31	1.49	4.73	−1.00
sex	Sex: female = 0, male = 1.	Categorical	27755	0.76	0.43	1.00	0.00
edu	Education level: 0–4, a larger value indicates a higher education level.	Categorical	27755	0.92	0.86	4.00	0.00
household	Household type: non-local = 0, local = 1.	Categorical	27755	0.14	0.35	1.00	0.00
marriage	Marriage status: married = 1, others = 0.	Categorical	27755	0.59	0.49	1.00	0.00
contact	Closeness of relationship between applicant and contact: 0–5, a larger value indicates a closer relationship.	Categorical	27755	0.91	1.23	5.00	0.00
household_addr_con	Consistency of household province with other addresses, including the residential address of applicant, the company address of the applicant, the company address of the spouse, and the address of the contact: 0–4, a larger value indicates a higher level of consistency.	Address	27755	2.27	1.11	4.00	0.00
resid_addr_con	Consistency of applicant residence city with other addresses, including the company address of the applicant, the company address of the spouse, and the address of the contact: 0–3, a larger value indicates a higher level of consistency.	Address	27755	1.65	0.56	3.00	0.00

3.3. Empirical analysis

Customers can falsify various information in loan applications, and some common types include occupation, ability certificate, marriage, and contact information. We first explored the relationships between these fake types based on the correlation matrix in Table 2. We used Cramer's V-value to measure correlation. We found that all the correlation coefficients were smaller than 0.3, which shows a weak relationship or irrelevance between the different types. This increases the difficulty of convergence of the algorithm if all types are integrated into one model. However, although there is no direct and robust relationship between them, customers with fabricated information may still have common characteristics. Therefore, we developed a more profound analysis approach using binary logistic regression.

Table 3 lists the results of this empirical study. Some vital elements played the same role in all of the tasks or several tasks, whereas some had different effects on different fabrication behaviors. Specifically, customers from local households had a higher probability of falsifying information for all fraud types. However, several factors only had the same impact on some falsification types and no significant influence on others, including sex, edu, amount, interest, term, down_pay_ratio, work_year, income, contact, and household_addr_con. For example, female customers were more likely to provide false job, marriage, and contact information in loan applications, and applicants with lower education levels more often submitted fake ability certifications. Another example is that fraudulent applicants with lower monthly incomes (representing poorer economic states) were more likely to fabricate their jobs, abilities, and marriage information. In addition, age, marriage, and resid_addr_con had opposite effects on different tasks. For instance, younger fraudulent customers prefer submitting fake jobs and contact information, whereas older people prefer falsifying marriage and ability information.

In conclusion, some variables have similar effects on different tasks, but some have different influences. Therefore, we could neither detect varying fraud types with a unified technique nor treat them entirely as independent tasks. Thus, we propose to decompose this task into several subtasks and solve the problem with the MTL framework. Furthermore, variables related to applicant's residential address information worked significantly in three tasks (job, marriage, and fake contact). To further explore the complex relationships among address information, we developed a heterogeneous network and adopted representation learning to transfer address nodes into vectors.

In summary, we found a weak correlation or irrelevance between different fake information types, and different variables may have varying performances in the subtasks; thus, we propose a hierarchical fraud-classification MTL approach to decompose and solve this challenging task (Zhao et al., 2021). In addition, we found that address information plays a critical role; thus, we constructed a heterogeneous network to study the interaction of these variables.

Table 2
Label correlation matrix.

	Fake Occupation	Fake Ability	Fake Marriage	Fake Contact
Fake Occupation	**1**	**0.0056**	**0.2215**	**0.0903**
Fake Ability	**0.0056**	**1**	**0.0119**	**0.0013**
Fake Marriage	**0.2215**	**0.0119**	**1**	**0.0248**
Fake Contact	**0.0903**	**0.0013**	**0.0248**	**1**

Table 3
Empirical analysis using binary logistic regression.

	Fake Occupation		Fake Ability		Fake Marriage		Fake Contact	
	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio	Coefficient	Odds ratio
sex	−0.221***	0.802***	−0.119	0.888	−0.395***	0.674***	−0.501***	0.606***
edu	−0.031	0.970	−0.108*	0.897*	−0.031	0.970	0.051	1.052
age	−0.369***	0.691***	0.862***	2.367***	0.792***	2.209***	−0.919***	0.399***
household	0.099**	1.105**	0.429***	1.536***	0.161***	1.174***	0.386***	1.471***
amount	0.576***	1.779***	0.281***	1.325***	0.151***	1.163***	0.001	1.001
interest	0.434	1.544	0.460*	1.584*	0.245	1.278	0.517**	1.677**
term	−0.010	0.990	0.078	1.082	0.188**	1.206**	−0.187	0.830
down_pay_ratio	−0.343*	0.710*	0.649	1.914	−0.571***	0.565***	−0.429	0.651
work_year	−0.273***	0.761***	−0.049	0.953	−0.045*	0.956*	0.028	1.029
income	−0.440***	0.644***	−0.137*	0.872*	−0.103***	0.903***	0.014	1.014
driving_year	0.013	1.013	−0.011	0.989	0.009	1.009	0.028	1.029
marriage	−0.156***	0.856***	−1.016***	0.362***	0.330***	1.392***	−1.875***	0.153***
contact	−0.011	0.989	−0.060	0.942	−0.027	0.973	0.471***	1.601***
household_addr_con	−0.136***	0.872***	−0.001	0.999	−0.006	0.995	0.040	1.041
resid_addr_con	−0.073**	0.929**	−0.135	0.873	0.219***	1.244***	−0.285***	0.752***
Constant	−1.783	0.168	−8.550	0.000	−6.396	0.002	0.730	2.076

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$.

4. Detection methods

This section introduces HMTL for fake loan application information detection. We first provide an overview of HMTL and then present the details of the approach, including the hierarchical fraud classification method, input layer, embedding layer, and prediction layer. In addition, certain techniques have been introduced to enhance the performance of the corresponding parts.

4.1. Overview

The HMTL framework is illustrated in Fig. 3. As mentioned above, there are various types of fake information in fraud applications, and the relationships among them are complicated, which increases the prediction difficulty; therefore, the main idea is to perform the desired tasks using hierarchical fraud classification as well as shared and specific feature representation. First, the hierarchical fraud classification method decomposes fabrication identification into several subtasks according to the type of fabrication information. The proposed HMTL has three main components: input, embedding, and prediction layers. The input layer is the foundation layer of the framework, and there are three main types of data: numerical, categorical, and address features. The second layer is the embedding layer, in which we developed specific feature engineering based on different data types. Specifically, we constructed a heterogeneous network to learn the vector

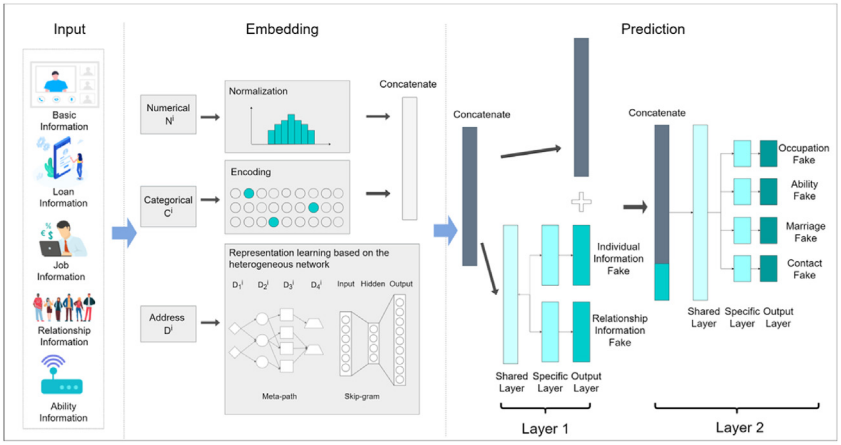


Fig. 3. An overview of the HMTL framework. It contains the input layer, embedding layer, and a two-level prediction layer. The input layer includes basic, loan, job, relationship, and ability information, which can be categorized into numerical, categorical, and address features. The embedding layer transfers the numerical and categorical features via normalization and the embedding method, respectively. Then, representation learning is conducted based on a heterogeneous network for addressing features. Lastly, the first prediction layer provides the general type-probability as auxiliary information to the second layer, which then predicts the final sub-tasks.

representation of each address feature. The final layer is the prediction layer, which contains two levels. In the first level of the prediction layer, the general category of fake information is predicted. Then, the results of the first level are utilized as auxiliary information to predict the final category of the counterfeit information.

4.2. Hierarchical fraud classification

To manage all types of fraudulent customers more efficiently, we propose a hierarchical fraud classification method, as shown in Fig. 4, to classify different types of fraudulent customers according to their fake information, as the foundation of our detection approach. There are two primary levels in the classification system: the first level represents the general category of fake information, and the second level represents the corresponding category. Specifically, in the first level, all customers with false information are classified as individual fake or relationship-fake. It should be noted that a customer could belong to both categories if he/she falsifies both types of information. In the second level, all false information is categorized into a more specific level. In particular, individual fake includes false that is only related to the applicant himself/herself, such as fake job and fake ability. Fake relationship contains falsifications related to other people, such as fake marriage and fake contacts.

4.3. Input layer

The input layer is the basic layer of our framework. This layer categorizes all data into three types. Specifically, feature groups that were proven useful for task prediction were used, including basic, loan, job, relationship (marriage and contact), and ability information. We categorized all these features into three types: numerical, categorical, and address features. The types of variables used in the empirical study are listed in Table 1. Here, the numerical and categorical features are denoted by N^i and C^i, respectively, where i represents the number of features. Furthermore, all address information was decomposed into three parts, including province, city, and district, and there were three character-types in our dataset: applicant, applicant's spouse, and contact. For the applicant, we obtained the household address, residential address, and company address information. As for the spouse and contact person of the applicant, we only had their company-address information. Moreover, we obtained address information related to the loan business. Thus, we denote the district component of the address information as $D^i = (D^i_{ah}, D^i_{ar}, Dd^i_{ac}, D^i_{sc}, D^i_{cc}, D^i_l)$, where each element represents the district information of the applicant's household, residence, applicant company, spouse company, contact company, and loan-related addresses, respectively. Because district information can be understood as a more detailed representation of cities and provinces, we focused only on district information in our study.

4.4. Embedding layer

The embedding layer is a crucial step for learning the representation of features and developing feature engineering for model training and prediction. Specifically, the numerical features N^i were standardized using the z-score method (Bolstad et al., 2003). HMTL uses one-hot encoding to transfer the categorical features C^i (Singh et al., 2020). A heterogeneous network was developed to learn the deep relationships between address features d^i (Dong et al., 2017; Wu et al., 2022). We constructed a heterogeneous graph to contain all related nodes. Formally, a heterogeneous network is defined as a graph $G = (V, E, T_V, T_E)$ consisting of a node set V and edge set E, where an edge type represents the link between two types of nodes. Each node v is associated with a node-type mapping function $\varphi(v) : V \rightarrow T_V$, and each edge e is associated with an edge-type mapping function $\varphi(e) : E \rightarrow T_E$, where T_V and T_E denote the sets of all node types and edge types, respectively, and $|T_V| + |T_E| > 2$ (Wang et al., 2016). For example, in a heterogeneous network, two applicants' residential addresses can be homogeneous neighbors and an applicant's residential address and company address can be heterogeneous neighbors. The dense vector representation of nodes can then be learned as follows: $X \in \mathbb{R}^{(|V| \times d)}, d \ll |V|$, which captures the relationships among the nodes.

To learn the embedding vector of nodes, metapath2vec combines meta-path-based random walks and a heterogeneous skip-gram model (Guthrie et al., 2006) in heterogeneous networks. First, we define a meta-path scheme P as $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \rightarrow$

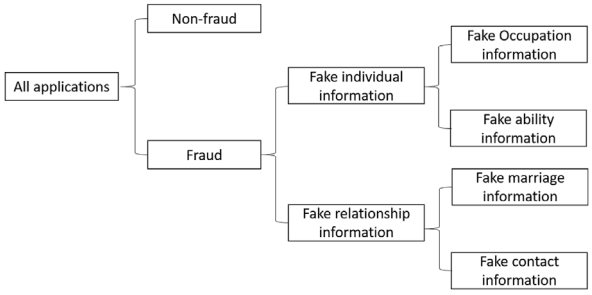


Fig. 4. Hierarchical fraud classification.

$V_t \xrightarrow{R_t} V_{t+1} \xrightarrow{R_{i-1}} V_l$, where R_i denotes the relationship between node types V_i and V_{i+1}. The walker flow is determined by the defined walking path; that is, the probability of transition at step i is defined as follows:

$$P\left(v^{i+1}\middle|v_t^i\right) = \begin{cases} \dfrac{1}{\left|N_{t+1}\left(V_t^i\right)\right|}, & \left(v^{i+1},\ V_t^i\right)\in E,\ \varphi\left(v^{i+1}\right) = t+1 \\ 0, & \left(v^{i+1},\ V_t^i\right)\in E,\ \varphi\left(v^{i+1}\right)\neq t+1 \\ 0, & \left(v^{i+1},\ V_t^i\right)\notin E \end{cases} \tag{2}$$

where $v_t^i \in V_t$, and $N_{t+1}(V_t^i)$ represents the V_{t+1} type of neighborhood of node V_t^i. Using this strategy, the structural relationships between different types of nodes can be transformed into a skip-gram model. Subsequently, the skip-gram model learns the representation of the nodes. Given node v, the probability of the context $N_t(v)$, $t \in T_V$ can be maximized as follows:

$$argmax_\vartheta \sum_{v\in V}\sum_{t\in T_V}\sum_{C_t\in N_t(v)} \log P(C_t|v;\vartheta) \tag{3}$$

where $N_t(v)$ denotes the neighborhood of v's t^{th} type of node, and $P(C_t|v;\vartheta)$ is defined as $P(C_t|v;\vartheta) = \frac{e^{X_{C_t}\cdot X_v}}{\sum_{u\in V} e^{X_{C_t}\cdot X_v}}$, where X_v is the v^{th} row of the node embedding vector.

In summary, with all encoded and transferred features, HMTL predicts two-level tasks in the next layer.

4.5. Prediction layer

As mentioned above, the novel HMTL has two sublayers for prediction tasks: predicting the general category and the specific category. The tasks in the first level predict the general category, and the output contains two features representing the probabilities of individual information and relationship information being fake, which means that an applicant can be classified under non-fraud, only false individual information, only false relationship information, and both false individual and relationship information. Although the result may not be completely accurate because there are several fake types to be predicted in one task, it contributes to improving the final results. Specifically, our final tasks include occupation, ability, marriage, and contact information fraud detection; however, the imbalanced distribution caused by the different sample sizes of two classes is a crucial reason for poor prediction performance. Thus, in general fraud-type detection tasks, HMTL increases the fraudulent data sample size by combining different types of fraud. This oversampling method increases the minority class samples (He and Garcia, 2009) and helps solve the imbalance problem resulting from the small fraudulent dataset size. Subsequently, the prediction result serves as auxiliary information for the second prediction layer. In the second level, the specific fake category of loan applications is predicted through the MTL mechanism along with the auxiliary information (Ruder, 2017) from level one. Similar to the first layer, the output of the second layer shows whether an applicant's information can independently contain false occupation, ability, marriage, or contact information. Note that there are two possible results (fake or not fake) for each fake type and four types in total; therefore, the output of layer 2 has 16 classes.

Formally, all the features N^i, C^i, D_{ah}^i, D_{ar}^i, D_{ac}^i, D_{sc}^i, D_{cc}^i, D_l^i are concatenated, and a network containing shared layers and specific layers is employed to predict the probability of the i^{th} task in level one, which is denoted as Y_1^i. That is,

$$H_{1sh} = X\,W_{1sh} + b_{1sh} \tag{4}$$

$$H_{1sp}^i = H_{1sh}W_{1sp}^i + b_{1sp}^i \tag{5}$$

and

$$Y_{t1}^i = Softmax\left(H_{1sp}^i W_1^i + b_1^i\right) \tag{6}$$

where W_{1sh}, b_{1sh}, W_{1sp}^i, b_{1sp}^i, and W_1^i, b_1^i are the parameters to learn, and H_{1sh} and H_{1sp}^i represent the hidden states of the shared and specific layers, respectively. Here, HMTL adopts a widely used approach for multi-task learning frameworks in neural networks called hard parameter sharing (Caruana, 1993), which is achieved by sharing the hidden layers between tasks while maintaining several task-specific layers. Specifically, HMTL employs a shared layer to realize parameter sharing, which helps reduce the risk of overfitting and improves the generalization ability. Subsequently, a specific layer is utilized to fit the specific parameters in different tasks to avoid underfitting (Ruder, 2017). Therefore, from the above results, the probability of the general categories is obtained, and the fake individual and relationship information can be a strong hint for the tasks in the next prediction layer. The input features are concatenated with the results from the first level, and then, a similar-structured multi-task network is employed using the first level to obtain the final results. That is,

$$H_{2sh} = \left(X \oplus Y_{t1}^{1} \oplus Y_{t1}^{2}\right) W_{2sh} + b_{2sh} \tag{7}$$

$$H_{2sp}^{i} = H_{2sh} W_{2sp}^{i} + b_{2sp}^{i} \tag{8}$$

and

$$Y_{t2}^{i} = Softmax\left(H_{2sp}^{i} W_{2}^{i} + b_{2}^{i}\right) \tag{9}$$

where W_{2sh}, b_{2sh}, W_{2sp}^{i}, b_{2sp}^{i}, and W_{2}^{i}, b_{2}^{i} are the parameters to learn, and H_{2sh} and H_{2sp}^{i} represent the hidden state of the shared and specific layers, respectively.

Another crucial element of HMTL is the combination of multiple loss functions. Traditionally, most methods utilize a weighted sum of losses, in which the weights are simply uniform or chosen manually. In our method, homoscedastic uncertainty is used as a basis for loss weighting, which is dependent on the task itself rather than on input data or other factors (Kendall et al., 2018). In the classification problem, where $f^{w}(x)$ is the output of the task and w is the weight, the likelihood with scalar σ can be defined as follows:

$$p\left(y \mid f^{w}(x),\ \sigma\right) = Softmax\left(\frac{1}{\sigma^{2}} f^{w}(x)\right) \tag{10}$$

where σ can be interpreted as the temperature and can be learned or fixed in a Boltzmann distribution. The log likelihood can be written as:

$$\log p\left(y = c \mid f^{w}(x),\ \sigma\right) = \frac{1}{\sigma^{2}} f_{c}^{w}(x) - \log \sum_{c'} \exp\left(\frac{1}{\sigma^{2}} f_{c'}^{w}(x)\right) \tag{11}$$

where $f_{c}^{w}(x)$ is the c^{th} element of output $f^{w}(x)$. Then, considering an example where two classification tasks exist, the combined loss $L(W,\ \sigma_1,\ \sigma_2)$ can be expressed as follows:

$$L(W,\ \sigma_1,\ \sigma_2) = \frac{1}{\sigma_1^{2}} L_1(W) + \log \frac{\sum_{c'} \exp\left(\frac{1}{\sigma_1^{2}} f_{c'}^{w}(x)\right)}{\left(\sum_{c'} \exp\left(f_{c'}^{w}(x)\right)\right)^{\frac{1}{\sigma_1^{2}}}} + \frac{1}{\sigma_2^{2}} L_2(W) + \log \frac{\sum_{c'} \exp\left(\frac{1}{\sigma_2^{2}} f_{c'}^{w}(x)\right)}{\left(\sum_{c'} \exp\left(f_{c'}^{w}(x)\right)\right)^{\frac{1}{\sigma_2^{2}}}} \tag{12}$$

and

$$L(W,\ \sigma_1,\ \sigma_2) \approx \frac{1}{\sigma_1^{2}} L_1(W) + \frac{1}{\sigma_2^{2}} L_2(W) + \log\sigma_1 + \log\sigma_2 \tag{13}$$

where $\frac{1}{\sigma_1}\left(\sum_{c'} \exp(f_{c'}^{w}(x))\right) \approx \left(\sum_{c'} \exp(f_{c'}^{w}(x))\right)^{\frac{1}{\sigma_1^{2}}}$ is simplified when σ_1 approaches 1. From the final equation, a larger σ can decrease the contribution of $L(W)$ but receive a larger penalization with $\log\sigma$. In our study, we trained the scale of σ for each subtask to obtain the weights of the losses.

5. Experiments

This section illustrates the effectiveness of HMTL through experiments. First, we compared the proposed approach with baseline methods (Zhu et al., 2020) and HMTL variants on a real-world dataset. Then, we conducted an in-depth analysis of HMTL, including the impact of different loss weighting methods on model performance and hyper-parameter sensitivity in representation learning. Finally, we provide a case study to illustrate the rationality and effectiveness of HMTL.

5.1. Datasets

We used the same dataset from a Chinese automobile finance company to conduct our experiments. In our experiments, we focused on four specific tasks: finding applications with fake occupation information, fake driving ability information (driving ability is a crucial point for automobile loan application assessment), fake marriage information, and fake contact information. The first two tasks belong to the fake individual information category and the other two belong to the fake relationship information category. We randomly split the entire dataset into three parts at a ratio of 3:1:1 to obtain training,

validation, and test datasets. The basic information for the three datasets is shown in Fig. 5. We trained the model using the training dataset, tuned the parameters using the validation dataset, and obtained the final results using the test dataset.

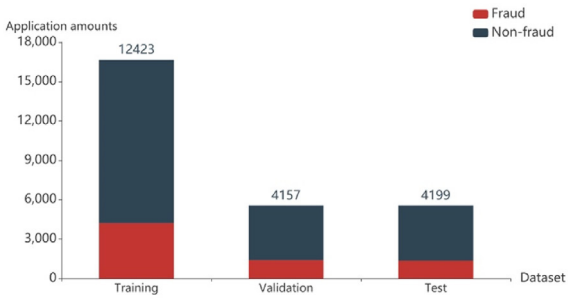
Based on the variables in the empirical study, we employed more features in feature construction, as shown in Table 4.

5.2. Experimental settings

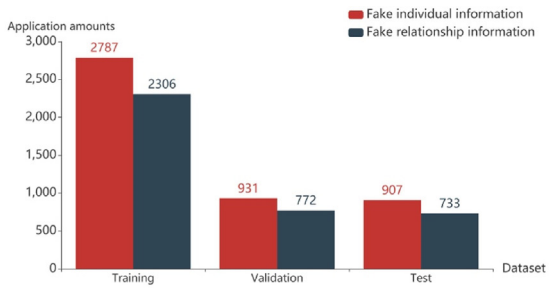
5.2.1. Baseline approaches

We conducted extensive experiments to compare the performance of HMTL with that of state-of-the-art models. Specifically, some of the approaches for feature engineering that we chose are as follows:

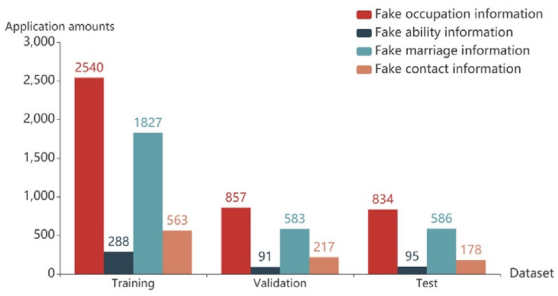
- LabelEncoding (Guedrez et al., 2016): An encoding method for transforming non-numerical features into numerical features.
- ConsistCheck: A statistical method for feature construction based on the consistency of different types of addresses.



(a) Number of fraudulent and non-fraudulent applications



(b) Number of applications for each general fraud type



(c) Number of applications for each specific fraud type

Fig. 5. Application distribution of different types in the training, validation, and test datasets. We randomly split the entire dataset into three parts at a ratio of 3:1:1, which are named the training, validation, and test datasets. We trained the model using the training dataset, tuned the parameters using the validation dataset, and obtained the final results using the test dataset.

Table 4
Examples of features used for modeling.

Information	Feature example
Basic information	Sex, age, nationality, etc.
Loan information	Loan interest, loan ratio, product price, etc.
Job information	Monthly income difference from average, working years, district of company address, etc.
Ability information	The age of obtaining driver's license, driving experience, etc.
Marriage information	Marriage status, district of spouse's company address, etc.
Contact information	Relationship with contact, district of contact's company address, etc.

We then employed the following popular algorithms in fraud-detection problems with the above methods for feature engineering to train the model. Because of the limited availability of benchmark data for testing methods and complicated fraud types in information falsification detection problems (Dal Pozzolo et al., 2017), there have been few studies on related detection techniques. Thus, we adopted several state-of-the-art methods that are widely used for similar financial problems, including online credit payment default detection (Zhong et al., 2020), loan default detection (Hu et al., 2020), etc.

- Logistic Regression (LR) (Peng et al., 2002): An extension method of the linear regression model for classification problems, which is popular in credit scoring card modeling.
- Neural network (NN) (Hecht-Nielsen, 1992): A network of "neurons" organized in layers that can model nonlinear and complicated relationships between the features and labels. In our experiment, MLP with one hidden layer is used.
- Random Forest (RF) (Liaw et al., 2002): A bagging-based algorithm that constructs many decision trees for training and then improves the predictive result by the averaging method.
- Gradient Boosting Decision Tree (GBDT) (Rao et al., 2019): A machine learning technique that produces many weak prediction decision trees as the prediction model by boosting methods with a differentiable loss function.
- XGBoost (Chen et al., 2015): A popular implementation of GBDT with improvements in regularization, handling sparse data, etc.
- LightGBM (Ke et al., 2017): A fast, distributed, high-performance gradient boosting framework based on the decision tree algorithm.
- DeepForest (Hu et al., 2020; Zhong et al., 2020; Zhou and Feng, 2017): A deep model based on the decision tree ensemble approach with tree characteristics, i.e., layer-by-layer processing, in-model feature transformation, and sufficient model complexity.
- HAN (Hu et al., 2020; Wang et al., 2019; Zhong et al., 2020): A semi-supervised heterogeneous graph neural network based on hierarchical attention, including node-level and semantic-level attentions.

Additionally, we adopted some variants of the HMTL approach to explore the contributions of different components in our model, including the following.

- HMTL-HP: This removes the hierarchical prediction layers of HMTL, which predicts the four subtasks without a general category probability.
- HMTL-Hete: This only develops traditional feature engineering with LabelEncoding and ConsistCheck rather than constructing a heterogeneous graph.
- HMTL-Single: This fits and predicts data independently without the MTL framework.

5.2.2. Parameter settings

For HMTL and its variants, we examined the hyperparameter sensitivity and selected the best parameters. To construct the representation of address information, we set the window size, vector dimension, batch size, and size of negative samples to 3, 64, 128, and 5, respectively. The learning rate was initialized to 0.01. Regarding the MTL approach, we trained every model for 20 epochs with a batch size of 500 and vector dimension of 128. The learning rate of the Adam optimizer was initially set to 10^{-3}.

5.2.3. Evaluation metrics

To evaluate the performance of HMTL on the automobile loan dataset, we chose three metrics most commonly used in classification problems: area under the curve (AUC), accuracy, and F1-score. A higher value of AUC, accuracy, and F1-score shows a better performance of the method. To the best of our knowledge, owing to a lack of studies on auto loan fraud detection, there are few reports on detection performance. Jerzy et al. (Błaszczyński et al., 2021) adopted a random forest and support vector machine algorithm for auto loan detection, and the performance measured by the F1-score was under 0.30. In our study, we compared the proposed methods with state-of-the-art baselines.

Table 5
Predictions of state-of-the-art baselines.

Baseline	F1-score	Accuracy	AUC
LR	0.3894	0.3181	0.5303
NN	0.3938	0.2818	0.539
RF	0.4186	0.4115	0.6636
GBDT	0.4587	0.5476	0.6963
XGB	0.4555	0.5374	0.7035
LGB	0.4578	0.5305	0.6997
DeepForest	0.4609	0.5306	0.7176
HAN	0.4777	0.5111	0.7233

Table 6
Overall performance of the HMTLs.

Method	Overall performance	
	F1-score	Accuracy
HMTL	**0.4865**	**0.7019**
HMTL-HP	0.4658	0.6905
HMTL-Hete	0.4624	0.6921
HMTL-Single	0.4590	0.6658

5.3. Experimental performance

To evaluate the effectiveness of our novel framework, Tables 5–7 list the experimental results of the baseline models and HMTL with its variants.

Overall Performance. The results show that our HMTL method and its variants outperform the baseline methods. It should be noted that we integrated the results of the four subtasks to evaluate the overall performance of HMTL and its variants. Specifically, we chose the thresholds that yielded the best F1-score values on the validation dataset and then applied these thresholds to the test data to obtain the corresponding values of accuracy and F1-score. We obtained the overall performance results by combining the sub-task results (thus, we only had F1-score and accuracy values for the overall performance of HMTL but no AUC values). For the baseline methods, we obtained the overall prediction performance for all types of fake information detection from the model directly, without categorizing them into specific tasks. The following observations were made based on the experimental results:

1) The proposed HMTL and its variants outperformed all the baseline models in terms of overall performance. Specifically, HMTL achieved remarkable improvement over the baselines with respect to the F1-score by 3.95–24.94% and accuracy by 28.17–49.06%. The results demonstrate the effectiveness of HMTL for falsification detection.
2) Among all the baseline models, the tree-based models performed better than the linear models. Their AUC values were 25.14–35.32% higher than those of logistic regression, and the F1-score and accuracy values were also 7.50–18.36% and 29.36–72.15% higher, respectively. This indicates that tree-based models can capture nonlinear effects and interactions to improve the model performance. In addition, we found that DeepForest had the best performance among all the tree-based models. It achieved the best AUC by deeper modeling of the attributes, which was improved by 2.00–3.06% compared with other tree-based models. Moreover, HAN improved the AUC further than DeepForest, which demonstrates that the GNN-based method can better model attributes and structural information.

Table 7
Performance of the HMTL subtasks.

Method	Task1 Fake Occupation			Task2 Fake Driving Ability		
	AUC	F1-score	Accuracy	AUC	F1-score	Accuracy
HMTL	**0.7051**	0.3615	0.7028	**0.6927**	0.0653	0.8143
HMTL-HP	0.6855	0.3391	0.6811	0.6692	0.0577	0.8116
HMTL-Hete	0.6850	0.3293	0.6793	0.6554	0.0604	0.6864
HMTL-Single	0.6747	0.3257	0.6941	0.6362	0.0599	0.7116
Method	Task3 Fake Marriage			Task4 Fake Contact		
	AUC	F1-score	Accuracy	AUC	F1-score	Accuracy
HMTL	**0.7210**	0.3041	0.8483	**0.8503**	0.2730	0.9146
HMTL-HP	0.6901	0.2999	0.8377	0.8356	0.2412	0.9377
HMTL-Hete	0.7061	0.2982	0.8380	0.8472	0.1806	0.9542
HMTL-Single	0.6838	0.2849	0.8200	0.8342	0.2067	0.9447

Table 8
Overall Performance of different loss weighting methods.

Method	F1-score	Accuracy
Uncertainty	0.4865	0.7019
AVG	0.4787	0.7179
SampleSize	0.4914	0.7028

Ablation Results. Further we evaluated the contribution of each component to the proposed HMTL. First, we found that HMTL-HP performed worse than HMTL in terms of overall performance as well as subtask performance when removing auxiliary information representing general categories, which shows that this two-level prediction layer facilitates improved predictive results. Specifically, the first prediction layer provides general fraud-type attributes by solving the imbalance problem (by increasing the sample size of the minority class), and the results serve as auxiliary information in the second layer. This mechanism learns the features as hints for the final tasks and works for tasks that might not be easy to learn using only the original information. Second, when we adopt the LabelEncoding and ConsistCheck methods to replace the construction of heterogeneous graphs with learning vector representations, the smaller values of the evaluation metrics illustrate the crucial role of using the heterogeneous graph to learn deep interaction information between features. Furthermore, when utilizing single tasks to predict these subtasks, we noticed that the performance was poorer than that of HMTL and other variants, which demonstrates the effectiveness of the MTL framework. Specifically, the multi-task learning framework improves the generalization ability and avoids overfitting, and the superior improvement of the MTL framework versus other components illustrates the efficacy of this design.

In addition, comparing the subtask results shows that the difficulties of various tasks were different, where fake contact information detection had the best outcome and fake driving ability information detection was the most challenging. One possible reason for this is that there are fewer features related to driving ability than in other tasks, which increases the difficulty of detection. Moreover, we found that one component may perform differently for different tasks. For example, the multi-task learning framework most improved the performance in the fake driving ability task, and its AUC was increased by 8.88%; however, it improved the AUC value of the fake contact task by only 1.93%. The main reason for this is that fake contact detection is easier than the driving ability task, which means that contact information is more suitable compared with the others; thus, the contribution of learning the hints and auxiliary information is not so obvious.

5.4. In-depth analysis

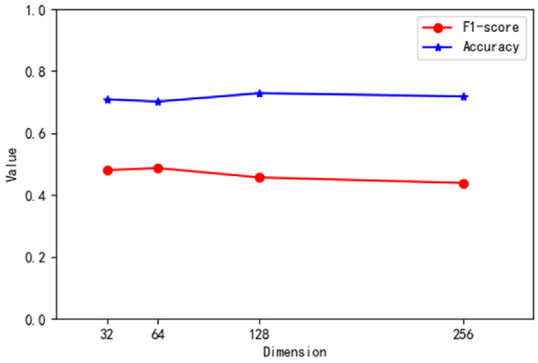
5.4.1. Loss weighting in MTL

For approaches based on MTL, the weights of the task losses play an essential role in the model performance. In the proposed HMTL, we use uncertainty for weight losses, but sometimes weights can be set manually for different purposes. For instance, when we attribute more importance to one task, we assign more weight to this task, and less weight is assigned to the other tasks that can be regarded as minor. Thus, we tested HMTL under different loss weights to observe the influence on model performance, and the results are shown in Tables 8 and 9.

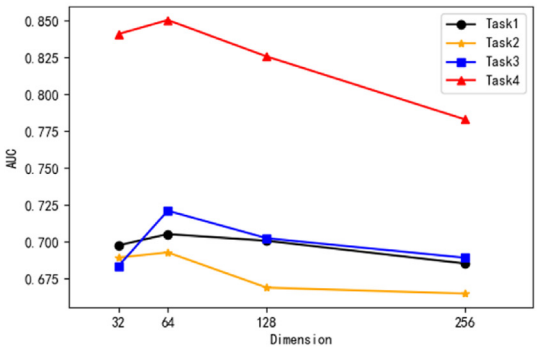
We tested three situations: HMTL (named Uncertainty), average weights on all the tasks (AVG), and a ratio proportional to the number of fake samples in the tasks (named SampleSize), whereby the ratios were 0.29 : 0.21 : 0.28 : 0.22, 0.25 : 0.25 : 0.25 : 0.25, and 0.5 : 0.05 : 0.35 : 0.1, respectively. In general, the overall performance was stable when different loss combinations were used; hence, we were able to manually set various loss weights according to the specific situation. Consider the identification of gang fraud as an example. When detecting gang fraud behaviors during loan application assessments, reviewers need to evaluate the applicant's materials to determine the complicated relationship among possible fraud customers. Thus, the authenticity of marriage information plays an essential role in this process, and we chose a proper weight ratio to obtain the best fake marriage task performance.

Table 9
Subtask performance of different loss weighting methods.

Method	Task1 Fake Occupation				Task2 Fake Driving Ability			
	Weight	AUC	F1-score	Accuracy	Weight	AUC	F1-score	Accuracy
Uncertainty	0.2900	0.7051	0.3615	0.7028	0.2100	**0.6927**	0.0653	0.8143
AVG	0.2500	0.7162	0.3679	0.7307	0.2500	0.6826	0.0602	0.8368
SampleSize	0.5000	**0.7284**	0.3783	0.7418	0.0500	0.6479	0.0578	0.6826
Method	Task3 Fake Marriage				Task4 Fake Contact			
	Weight	AUC	F1-score	Accuracy	Weight	AUC	F1-score	Accuracy
Uncertainty	0.2800	0.7210	0.3041	0.8483	0.2200	**0.8503**	0.2730	0.9146
AVG	0.2500	**0.7228**	0.2918	0.8453	0.2500	0.8473	0.2359	0.9440
SampleSize	0.3500	0.7147	0.2878	0.8440	0.1000	0.8129	0.2422	0.8997



(a) Overall performance: F1-score and Accuracy



(b) Subtask performance: AUC

Fig. 6. Performance of different dimensions in constructing the heterogeneous network. The overall performance was relatively stable, and the AUC was higher when choosing dimension = 64 for the subtasks.

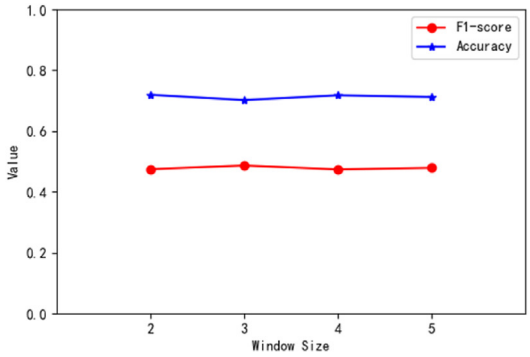
Moreover, an increased weight of another task's loss was not associated with increased performance in this task, which illustrates that other tasks contributed to this task; thus, reducing the loss of other tasks may lead to lower performance. For example, when assigning much more weight to the fake marriage task in SampleSize than the AVG method (the weight is 0.35 and 0.25, respectively), the AUC decreased from 0.7228 to 0.7147, which indicates that more weight does not result in higher performance.

5.4.2. Hyper-parameter sensitivity for a heterogeneous network

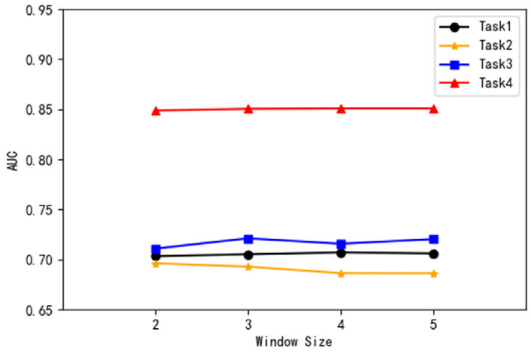
To learn the vector representation of various address information, we constructed a heterogeneous network and then utilized metapath2vec, which combines meta-path-based random walks and the heterogeneous skip-gram model. There are several critical parameters in this part, including the vector dimension d of the learned vector and window size w in the skip-gram model. Here, we explored varying d and w, and the results are shown in Figs. 6 and 7. For vector dimension d, we propose that the overall performance is relatively stable compared with the subtasks when increasing d; however, regarding the subtasks, higher AUC values could be achieved at $d = 64$. Similarly, the overall trend of performance with varied window sizes was stable, and when focusing on the AUC, we chose a window size $w = 3$, where a balance between the computational cost and efficacy can be obtained.

5.5. Case study

We provide a case study to demonstrate the effectiveness and rationality of HMTL. In Fig. 8, we provide an example of a loan applicant with her application information and the prediction result from our proposed method. From the prediction result, we found that the probability of fake occupation was relatively high. After a phone review and home visit, the credit reviewer labeled the applicant with a fake occupation. The consistency between the model prediction and confirmed truth illustrates the effectiveness of our method. Additionally, from the application information, we found that the applicant was young (26 years old) and had a low education level (junior college). She had only one year of work experience and lived in a



(a) Overall performance: F1-score and Accuracy



(b) Subtask performance: AUC

Fig. 7. Performance under different window sizes when constructing the heterogeneous network. The overall performance was relatively stable, and the AUC was higher when choosing window size = 3 for the subtasks.

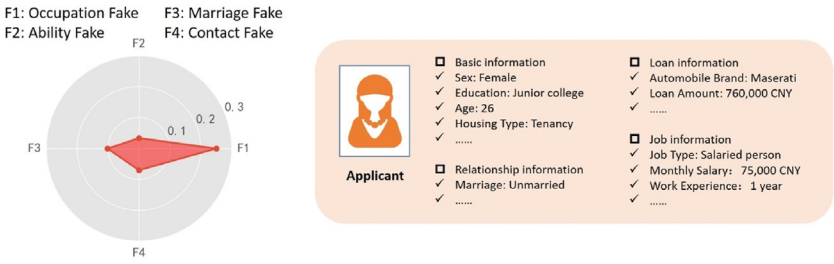


Fig. 8. An example of a loan applicant. The radar figure on the left shows the prediction results from our HMTL method, i.e., the fraud probabilities of the four specific falsification types. The description on the right shows some information about the applicant.

rented place. However, she stated that her monthly salary was as high as 75,000 CNY and hoped to apply for a loan of 760,000 CNY for a Maserati, which was not in accordance with her basic situation. Thus, it is reasonable to label her as having a fake occupation.

6. Conclusion

In this study, we examined the relationships among fraud types and explored the performance of various factors influencing fraud detection through an empirical analysis. The results suggest weak correlations among fabrication types and that crucial elements may perform similarly or differently in various tasks.

Based on the empirical results, we introduced a refined HMTL approach with a two-level fraud classification system for the fake information detection problem. Specifically, we defined a fake information category system to classify all customers and decompose this detection task into several subtasks based on falsified information types. Moreover, we constructed a heterogeneous network to learn crucial feature representations using a meta-path-based random walk method and skip-gram model, which can contribute to the deep mining of relationship information between features. Furthermore, the novel MTL framework predicts general categories of fake information and then adopts the results as auxiliary information to predict the final subtasks by utilizing uncertainty to weight task losses. Finally, the proposed method was evaluated on a real-world dataset through extensive experiments, and the results illustrate the effectiveness of HMTL.

In summary, loan application assessment faces more challenges with the development of finance and technology, which results in the demand for fraud-detection approaches with higher efficacy. Based on the empirical results, the proposed HMTL approach improves individual fraud-detection performance.

Furthermore, we are still facing challenges such as data isolation and data privacy, which exist in most industries. Specifically, the confidentiality of individual information in loan applications may cause the unavailability of data for machine learning. Thus, federated learning can be introduced in the future to realize information falsification detection with data isolation problems (Konečný et al., 2016). We will consider horizontal federated learning, vertical federated learning, and federated transfer learning, which are utilized in different scenarios depending on whether datasets share the same feature space and have the same samples (Yang et al., 2019). Specifically, we will consider adopting horizontal federated learning to increase the number of training samples to improve detection performance and add more application information for feature engineering through vertical federated learning. In addition, federated transfer learning can be used to identify different types of fraud in various business scenarios.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgment

This study was partially funded by the support of the NSFC Project of International Cooperation and Exchanges under Grant No. 72010107004, National Natural Science Foundation of China (72101176) and Beijing Fantaike Technology Co. Ltd.

References

Almendra, V. (2013). Finding the needle: a risk-based ranking of product listings at online auction sites for non-delivery fraud prediction. *Expert Syst. Appl., 40*(12), 4805–4811.

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: a comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1–9).

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *J. Oper. Res. Soc., 54*(6), 627–635.

Bengio, Yoshua, Courville, Aaron, Vincent, & Pascal. (2013). Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis \> Machine Intelligence, 35*(8), 1798–1828.

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: a comparative study. *Decision Support Systems, 50*(3), 602–613.

Bhowmik, R. (2011). Detecting auto insurance fraud by data mining techniques. *Journal of Emerging Trends in Computing and Information Sciences, 2*(4), 156–162.

Błaszczyński, J., de Almeida Filho, A. T., Matuszyk, A., Szeląg, M., & Słowiński, R. (2021). Auto loan fraud detection using dominance-based rough set approach versus machine learning methods. *Expert Systems with Applications, 163*, 113740.

Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2), 185–193.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: a review. *Statistical science, 17*, 235–255.

Brockett, P. L., Golden, L. L., Jang, J., & Yang, C. (2006). A comparison of neural network, statistical methods, and variable choice for life insurers' financial distress prediction. *Journal of Risk and Insurance, 73*(3), 397–419.

Caruana, R. A. (1993). Multitask learning: a knowledge-based source of inductive bias. *Machine Learning Proceedings, 10*(1), 41–48.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & others. (2015). Xgboost: extreme gradient boosting. *R Package Version 0.4-2, 1*(4).

Cheng, H., Fang, H., & Ostendorf, M. (2015). Open-domain name error detection using a multi-task RNN. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

China financial stability report 2020. (2020). China Financial Publishing House.

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167).

Consumer Finance Committee of China Banking Association. (2020). *Report on the Development of China Consumer Finance Companies 2020*.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3784–3797.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8599–8603).

Dong, Y., Chawla, N. V., & Swami, A. (2017). metapath2vec: scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 135–144).

Dorfleitner, G., & Jahnes, H. (2014). What factors drive personal loan fraud? Evidence from Germany. *Review of Managerial Science, 8*(1), 89–119.

Duman, E., & Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications, 38*(10), 13057–13063.

Eigen, D., & Fergus, R. (2014). *Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture*. IEEE.
Francis, C., Pepper, N., & Strong, H. (2011). Using support vector machines to detect medical fraud and abuse. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 8291–8294).
Grover, A., & Leskovec, J. (2016). *node2vec: Scalable Feature Learning for Networks*. ACM.
Guedrez, R., Dugeon, O., Lahoud, S., & Texier, G. (2016). Label encoding algorithm for MPLS segment routing. In *2016 IEEE 15th International Symposium on Network Computing and Applications (NCA)* (pp. 113–117).
Guthrie, D., Allison, B., Liu, W., Guthrie, L., & Wilks, Y. (2006). A closer look at skip-gram modelling. *LREC, 6*, 1222–1225.
Hamid, A. J., & Ahmed, T. M. (2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal (MLAIJ), 3*(1).
Hartmann-Wendels, T., Maehlmann, T., & Versen, T. (2009). Determinants of banks' risk exposure to new account fraud - evidence from Germany. *Journal of Banking \> Finance, 33*(2), 347–357.
He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.
Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural Networks for Perception* (pp. 65–93). Elsevier.
Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Publications of the American Statistical Association, 97*(December), 1090–1098.
Hu, B., Zhang, Z., Zhou, J., Fang, J., Jia, Q., Fang, Y., Yu, Q., & Qi, Y. (2020). Loan default analysis with multiplex graph learning. In *Proceedings of the 29th ACM International Conference on Information \> Knowledge Management* (pp. 2525–2532).
Intelligent risk control: principles, algorithms and practice. (2020). China Machine Press.
Kang, Y., Chen, L., Jia, N., Wei, W., Deng, J., & Qian, H. (2022). A CWGAN-GP-based multi-task learning model for consumer credit scoring. *Expert Systems with Applications*, Article 117650.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154.
Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7482–7491).
Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications, 32*(4), 995–1003.
Kokkinos, I. (2017). UberNet: training a 'Universal' convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). *Federated Learning: Strategies for Improving Communication Efficiency*. *ArXiv Preprint ArXiv:1610.05492*.
Lei, Tang, Huan, & Liu. (2011). Leveraging social media networks for classification. *Data Mining and Knowledge Discovery, 23*(3).
Leonard, K. J. (1995). The development of a rule based expert system model for fraud alert in consumer credit. *European Journal of Operational Research, 80*(2), 350–356.
Liaw, A., Wiener, M., & others. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.
López, V., Fernández, A., Garc\'\ia, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113–141.
Mailloux, A. T., Cummings, S. W., & Mugdh, M. (2010). A decision support tool for identifying abuse of controlled substances by ForwardHealth Medicaid members. *Journal of Hospital Marketing \> Public Relations, 20*(1), 34–55.
Malini, N., & Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. In *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics* (pp. 255–258). AEEICB).
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Neville, J., & Jensen, D. (2005). Leveraging relational autocorrelation with latent group models. *IEEE*, 49–55.
Ngai, E., Hu, Y., Wong, Y. H., Chen, Y., & 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, 50*(3), 559–569.
Olszewski, D. (2014). Fraud detection using self-organizing map visualizing the user profiles. *Knowledge-Based Systems, 70*, 324–334.
Ortega, P. A., Figueroa, C. J., & Ruz, G. A. (2006). A medical claim fraud/abuse detection system based on data mining: a case study in Chile. *DMIN, 6*, 26–29.
Osman, N., & Sierra, C. (2016). *Autonomous Agents and Multi-Agent Systems*. Kluwer Academic Publishers.
Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14.
Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). *DeepWalk: Online Learning of Social Representations*. ACM.
Phua, C. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.
Quah, J. T. S., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert Systems with Applications, 35*(4), 1721–1732.
Rao, H., Shi, X., Rodrigue, A. K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X., & Gu, L. (2019). Feature selection based on artificial bee colony and gradient boosting decision tree. *Applied Soft Computing, 74*, 634–642.
Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems, 50*(2), 491–500.
Ruder, S. (2017). *An Overview of Multi-Task Learning in Deep Neural Networks*. *ArXiv Preprint ArXiv:1706.05098*.
Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. In *2011 International Symposium on Innovations in Intelligent Systems and Applications* (pp. 315–319).
Sánchez, D., Vila, M. A., Cerda, L., & Serrano, J.-M. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications, 36*(2), 3630–3640.
Sawhney, R., Mathur, P., Mangal, A., Khanna, P., Shah, R. R., & Zimmermann, R. (2020). Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 456–465).
Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & Lecun, Y. (2013). *OverFeat: Integrated Recognition, Localization and Detection Using Convolutional Networks*. Eprint Arxiv.
Shah, H. S., Joshi, N. R., Sureka, A., & Wurman, P. R. (2002). Mining eBay: bidding strategies and shill detection. In *International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles* (pp. 17–34).
Shao, Q., Yu, R., Zhao, H., Liu, C., Zhang, M., Song, H., & Liu, Q. (2022). Toward intelligent financial advisors for identifying potential clients: a multitask perspective. *Big Data Mining and Analytics, 5*(1), 64–78. https://doi.org/10.26599/BDMA.2021.9020021
Singh, V. K., Maurya, N. S., Mani, A., & Yadav, R. S. (2020). Machine learning method using position-specific mutation based classification outperforms one hot coding for disease severity prediction in haemophilia 'A. *Genomics, 112*(6), 5122–5128.
Wang, D., Cui, P., & Zhu, W. (2016). Structural deep network embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1225–1234).
Wang, J. H., Liao, Y. L., Tsai, T. M., & Hung, G. (2006). Technology-based financial frauds in taiwan: issues and approaches. In *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*.
Wang, X., Ji, H., Shi, C., Wang, B., Ye, Y., Cui, P., & Yu, P. S. (2019). Heterogeneous graph attention network. In *The World Wide Web Conference, 2022–2032*.
Wen, C.-H., Wang, M.-J., & Lan, L. W. (2005). Discrete choice modeling for bundled automobile insurance policies. *Journal of the Eastern Asia Society for Transportation Studies, 6*, 1914–1928.

Wheeler, R., & Aitken, S. (2000). Multiple algorithms for fraud detection. In *Applications and Innovations in Intelligent Systems VII* (pp. 219–231). Springer.
Wu, L., Li, Z., Zhao, H., Liu, Q., & Chen, E. (2022). Estimating fund-raising performance for start-up projects from a market graph perspective. *Pattern Recognition, 121*, Article 108204.
Yang, L., Ng, T. L. J., Smyth, B., & Dong, R. (2020). Html: hierarchical transformer-based multi-task learning for volatility prediction. In *Proceedings of the Web Conference 2020* (pp. 441–451).
Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST), 10*(2), 1–19.
Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). BDD100K: a diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
Zaslavsky, V., & Strizhak, A. (2006). Credit card fraud detection using self-organizing maps. *Information and Security, 18*, 48.
Zhao, C., Zhao, H., Wu, R., Deng, Q., Ding, Y., Tao, J., & Fan, C. (2022). *Multi-dimensional Prediction of Guild Health in Online Games: A Stability-Aware Multi-Task Learning Approach*.
Zhao, H., Cheng, Y., Zhang, X., Zhu, H., Liu, Q., Xiong, H., & Zhang, W. (2022). What is market talking about market-oriented prospect analysis for entrepreneur fundraising. *IEEE Transactions on Knowledge and Data Engineering, 1*. https://doi.org/10.1109/TKDE.2022.3174336
Zhao, H., Jin, B., Liu, Q., Ge, Y., Chen, E., Zhang, X., & Xu, T. (2019). Voice of charity: prospecting the donation recurrence \& donor retention in crowdfunding. *IEEE Transactions on Knowledge and Data Engineering, 32*(8), 1652–1665.
Zhao, H., Liu, X., Zhang, X., Wei, Y., & Liu, C. (2021). The effects of person-organization fit on lending behaviors: empirical evidence from Kiva. *Journal of Management Science and Engineering, 7*, 133–145.
Zhong, Q., Liu, Y., Ao, X., Hu, B., Feng, J., Tang, J., & He, Q. (2020). Financial Defaulter Detection on Online Credit Payment via Multi-View Attributed Heterogeneous Information Network. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020* (pp. 785–795). https://doi.org/10.1145/3366423.3380159
Zhou, Z.-H., & Feng, J. (2017). *Deep Forest. ArXiv Preprint ArXiv:1702.08835*.
Zhu, N., Zhu, C., & Emrouznejad, A. (2020). A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies. *Journal of Management Science and Engineering, 6*(4), 435–448.