Review of Loan Fraud Detection Process in the Banking Sector Using Data Mining Techniques

Review Paper

Fahd Sabry Esmail
Helwan University,
Faculty of Commerce & Business Administration, Department of Business Information Systems
Cairo, Egypt
Fahd.Sabry21@commerce.helwan.edu.eg

Fahad Kamal Alsheref
Beni-Suef University,
Faculty of Computers and Artificial Intelligence, Department of Information Systems
Beni-Suef, Egypt
drfahad@fcis.bsu.edu.eg

Amal Elsayed Aboutabl
Helwan University,
Faculty of Computers and Artificial Intelligence, Department of Computer Science
Cairo, Egypt
amal.aboutabl@fci.helwan.edu.eg

Abstract *– At the era of digital transformation, fraud has dramatically increased, notably in the banking industry. Annually, it now costs the world's economies billions of dollars. Daily, news of financial fraud has a negative influence on the world economy. According to the harsh loss caused by fraud, effective strategies and methods for avoiding income statement fraud have to be implemented. Also, the procedure of identification should be applied. This is regarded as a result of the development of modern technology, modern invention, and the rapidity of global communications. Actually, deterrent technologies are most effective to reduce fraud and overcome cons. So, it is necessary to find ways to overcome such deterrence by depending on developed methods to identify fraud. Data mining techniques are currently the most widely used methods for the prevention and detection of financial fraud. The use of datasets for fraud detection complies with the norms of data mining, which include feature selection, representation, data gathering and management, pre-processing, comment, and summative evaluation. Methodologies for identifying fraud are essential if we want to catch criminals after fraud prevention has failed. The greatest fraud detection strategies for locating loan banking and financial fraud are compared in this article.*

Keywords*: Fraud, Loan Fraud, Loan Fraud detection, Data mining techniques.*

1. INTRODUCTION

Banks handle massive amounts of client data. It requires sustainable work to functionalize the aggregated data to detect consumer behaviors. This wastes time and effort and enables top managers to make the best decision and avert any prospective losses.

Data science is currently not only a pattern but rather a must to stay competitive in the banking industry. Moreover, data mining is increasingly emerging as a strategically significant topic for many commercial organizations, including the banking sector. It involves compiling data from many aspects and turning it into useful information. In addition, banks employ data science for target marketing, forecasting, consumer sentiment research, fraud detection, and customer support. Actually, it supports paying attention to the specifications presented to warn banks of fraud.

Furthermore, Fraud has a disastrous effect on economies all over the world, and several techniques have been applied but failed. However, machine learning is more dependable. Banks can find previously unidentified relationships in the data and look for hidden patterns in aggregations of data using data mining, which employs machine learning to make smart decisions based on the revealing insights. This enhances the capacity to identify resources, make better decisions, and perform better in loan appraisals and other categories of lending.

Obviously, many companies apply multifactor authentication as a competitive advantage. They evaluate their success in terms of profitability partially by the amount of fraud they can keep out of the hands of their competitors. These companies frequently do not have the intention to discuss or reveal their fraud prevention techniques to competitors. Fraudsters may be scared away by rapid action. This obligatory change is a vital component for companies that consider fraud control as a source of comparative advantage. The fraudsters would target their competitors to apply the scheme because their main objective is to carry out ways before their competitors.

The data are divided into two or more categories using the support vector machine (SVM) classifier, a kernel-based supervised learning method. Specifically for binary classification, SVM was developed. Within the training phase, SVM builds a model, maps the decision boundaries for each class, and finds the hyperplane that separates the multiple classes. In supervised learning, a set of input properties, such as plasma metabolite or transcriptional levels, are used to predict a quantitative sampling distribution. For example, there is the identification of loan fraud, or a qualitative one, like healthy or ill individuals. Several supervised learning techniques were handled, including multiple linear regression and random forests, as well as their usual behaviors with various sample sizes and numbers of predictor variables. We look at two popular supervised machine learning methods, linear support vector machines (SVM) and k-nearest neighbors (kNN). Both have been functionalized successfully to overcome challenging issues.

For the sake of better customer targeting and acquisition, the banking industry may profit basically from data mining tools. Customer retention is very valuable and automatic. Credit approval is used for fraud protection, and real-time fraud detection. This provides with customer analysis, segment-based solutions, historical transaction patterns for improved relationship development and retention, risk management, and marketing.

The banking industry has benefited immensely from advances in digital technology [1]. The concept of data being stored at branches has been replaced by centralized databases. Today, there are a lot more options for accessing bank accounts. Financial systems have become more client-focused and technically advanced thanks to digital purchasing, automated wire transfers, ATMs, and cash and check deposit devices [2]. The number of channels has increased along with the number of transactions and the data stored about them. Major banks have huge digital data warehouses within their computational storage systems. The quantity and quality of data have been developed [3].

Thanks to advancements in data mining methods and skills, the organization's data mountain is now proving to be its most valuable asset [4]. These data have interesting patterns and informative content. There is a great deal of potential for banks to functionalize data mining within decision-making through areas including marketing, debt management, proceeds of crime identification, liquidity management, investment banking, and the prompt detection of fraud transactions. The disability to achieve success within these fields could have negative effects on the bank, such as client loss to competitors' companies, monetary loss, reputational loss, and huge fines by the stakeholders.

In business, research, and many other fields, the need for commercial databases has expanded along with the requirement for content and retention. This increase in the amount of technologically held data can be explained by the growing acceptance of the link perspective of information preservation, as well as by the development and refinement of data access and generating contrast. In light of the need for data storage rose, this method was utilized. Recently, as previously, little consideration was given to creating software for data analysis. This changed when businesses found a resource hidden among these enormous data quantities. There is a wealth of information about their firm that has been kept and is simply waiting to be taken and used to improve a variety of elements assistance with company decisions. Functionalizing the database management systems that are used to manage these data sets, the user can presently only access information that is specifically contained in the databases. The amount of knowledge that exists is much greater than the size of a database, or the "ice shelf of knowledge," as it is called. Since this data unintentionally contains knowledge about numerous various aspects of their organization, it tends to be accessed and used for better decision-making.

Finding patterns within data represents the process of data mining. It can be beneficial in a variety of applications, including fraud detection. It combines complex data search techniques with statistical algorithms to uncover patterns and linkages. Inconsistent data, strange behavior, duplicate payments, missing invoices, abnormal transactions/vendors, and purchase and disbursement frauds, to name a few, are just a few examples of the abnormality and internal control holes that data mining can assist your firm find.

The aim of this research study is to contribute in literature concerning data mining methods used in banking to identify loan fraud. In order to categorize, extract relevant articles, and publish literature-based results, this work was completed in two parts. Stage one of classification involves locating applications, while stage two involves identifying fraud in the financial sector.

2. BACKGROUND

2.1. CONCEPT OF FRAUD

Fraud is defined as illegal deceit that is intentionally used by a person to get an unauthorized financial benefit. It may also take place with the express intent of misleading another person or organization, as in the case of making false assertions. Fraud is not a recent occurrence in contemporary life. For decades, fraudsters have been committing fraudulent acts [5]. This algorithm made somewhat more accurate predictions than the inspectors. Other reasoning systems [6] simulated the arguments of fraud experts by concentrating on two distinct tracks. The flexible anomaly classifier employed the Wang-Mendel technique to demonstrate how healthcare practitioners defrauded insurers. The search model uses an unsupervised network to discover links within data and to find clusters, after which patterns inside the clusters are found. The research gap is investigated in [7-9]. The electronic fraud detection (EFD) system [10] functionalized statistical data with expertise and knowledge to identify those whose actions deviated from the norm. Since the other clustering algorithms are frequently prohibitively costly when the datasets are very large and visualization techniques are applied for rule analysis, building mathematical synopses of the entities associated with each rule, the hot spots method combines the k-means segmentation method for cluster detection. [11-13] expanded the spots technique by generating and exploring the rules using a learning algorithm.

Several factors, including assistance from bank employees, can facilitate fraudulent activities, such as access to client databases, personal information, and information technology (IT) systems of the bank.

One definition of clumping is the discovery of groups of items that share characteristics. This method combines transactions with comparable behavior together. Segmentation can be used as developed a reputation for deciding which feature subsets to classify [14-16]. For example, in the banking industry, consumers from the tier always request a policy that guarantees more security because they are not determined on taking risks.

Similarly, people in the same middle to upper class who reside in rural settings may have tastes for some name products that are different from those who live in urban areas. Instead of mass presenting one particular "hot" product, the organization will be able to bridge other products. The company's customer service agents will have access to customer profile pages that have been enhanced through data analysis, enabling them to determine which services and goods are most meaningful to consumers.

One of the recent advancement in parallel with data processing technologies are data mining and information extraction. It incorporates the disciplines of information science, system administration, machine learning, statistics, and visualization. This is a new field. Despite this, the industry becomes more effective as a tool to research its clients and take reasonable judgments [17-18]. The process of uncovering true, fresh, possibly helpful, and ultimately intelligible data patterns is known as information retrieval from datasets. Data mining is a key step in deep learning, and the two terms are frequently used interchangeably [19].

Finding useful information from vast data repositories to address important business concerns is a technique known as data mining. It reveals hidden human analysis correlations, trends, patterns, exceptions, and oddities. Customers have a wide range of options in today's fiercely competitive market climate. In order to maintain their customer base, banks must be proactive in analyzing customer inclinations and profiles and tailoring their offerings and services accordingly [20-22]. A bank can reduce losses before it's too late by classifying customers into problematic and excellent consumers [23]. A bank can identify credit card fraud by examining average demand before it has an impact on its earnings [24]. Data analysis could be useful for achieving these highly desired qualities.

Fraudulent activities vary by severity, sector, complexity, manner, and difficulty of discovery or prevention of fraud differently. The summary is a well-rounded, non-exhaustive collection of several fraud categorizations.

2.1.1. Credit Card Fraud

Credit Card Fraud is referred as unauthorized use of a credit card account. A methodology with a malfeasance property and a clustering time regular by a classifier without an embezzlement attribute were both recommended by the credit card theft model [25-27]. Automobile injury claims were divided into different categories based on the degree of deception suspicion using a soul feature map [28-31]. The correctness of the feature map was then tested using a back propagation method and recurrent neural network networks. This occurs if neither the cardholder nor the card issuers are aware that a third party is using the card. As a result, fraudsters are able to buy things for free or acquire access to money in an account [32-34].

2.1.2. Insurance Fraud

Insurance fraud is known as an effort to take advantage of or abuse insurance coverage. Insurance is designed to cover losses and guard against dangers. Fraud happens when an insured utilizes their insurance policy to gain an unauthorized profit [35-37].

2.1.3. Money Laundering

Money laundering is a technique criminals use to conceal the source and final location of money obtained illegally to make it appear genuine [8].

2.1.4. Telecommunication Fraud

As relevant to telecommunications, fraud is defined as the usage of any carrier service without the purpose of paying. Other motivations, such as political or personal motivations, could be available.

2.1.5. Financial Statement Fraud

Financial fraud, commonly referred to as accounting fraud, is the deliberate misrepresentation of financial information in order to deceive the reader. Specifically, they mislead lenders and investors about a business's strategic stability [9].

2.1.6. Monetary deception Investment fraud

Monetary deception Investment fraud, commonly referred to as financial markets fraud, describes dishonest actions when securities are offered and sold [38-39]. High-yield investment fraud is a frequent kind of securities fraud. Affinity fraud, pyramid scams, and Ponzi schemes are a few well-known examples.

2.2. LOAN FRAUD OVERVIEW

Loan fraud, also known as lending fraud, refers to any deceptive action taken to gain a financial advantage during the loan process. Loan fraud can take many forms, including mortgage fraud, payday fraud, and loan scams. All of them will result in someone losing money, while the counterparty gains money and disappears. Fig.1 illustrates several types of secured and unsecured loans.

Secured loans have a pledge of something of value as security; if the borrower defaults on the loan, the bank or financial institution may sell the asset to recoup the loan balance. A good example of this kind of loan is a mortgage or home loan when the house or the property is used as collateral or security. If the borrower defaults, the bank may foreclose on the loan and change the loan balance.

Unsecured loans lack an underlying asset or another kind of security. The only thing the bank or other financial institution receives is the borrower's guarantee. Personal loans are an excellent example of an unsecured loan when the lending institution provides no security.

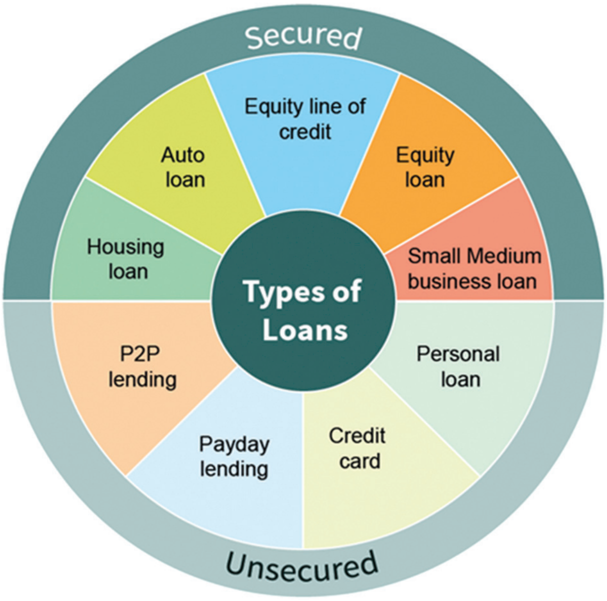

Fig. 1. Different types of secured and unsecured loans

2.3. LOAN APPLICATION FRAUD DEFINITION

Credit fraud, also known as loan application fraud, is a type of financial fraud in which the criminal obtains an illegal loan that they have no intention of repaying. They can take out the loan in their own name, which would be considered first-party fraud. However, it is more common for the line of credit to be opened in the name of a third party using forged or stolen identification documents. As shown in Table 1, there are two types of loan application fraud: first-party and third-party fraud.

Table 1. Types of Loan Application Fraud

First-Party Fraud	Third-party fraud
First-party fraud occurs when a criminal applies for a credit line or a personal loan using their legal name and documents. The criminal will then withdraw all of the funds from the account and vanish without leaving a trace.	Third-party loan fraud, on the other hand, involves receiving loans in another person's name. Criminals can accomplish this by using either stolen or forged identity documents. Read our article on document fraud to learn more about how they do it.

Personal loans, mortgages, commercial loans, and other sorts of loans are handled by banks and given to their customers. The bank will decide the customer's loan eligibility when the consumer submits an application [40-42]. Banks' main source of income comes from loans, but if borrowers don't make their payments on time, there will be bad loans [43-45]. To remain in business and earn the trust of their clients, banks must adhere to strict standards. In order to lower the risk that may hurt the bank, the most crucial criterion is to always investigate the behavior of the customer seeking a loan [13]. Numerous clients request bank loans, but because banks occasionally have limited resources and can only lend to a small number of customers, customers themselves are frequently forced to look for loans from third parties [14].

Naturally, fraudsters face a high-risk level because they must voluntarily hand over their personal information to the lender. While an evil actor could simply cross a state border and start a new life a century ago, this is nearly impossible in today's digital world. As a result, this method is becoming less popular by the year.

3. DATA MINING APPLICATIONS IN BANKING

As a method of identifying beneficial patterns and correlations, data mining has a specific place in financial modeling. Virtually all data mining techniques, like other computational methods, can be used in financial modeling. In the banking industry, data mining is beneficial.

Large volumes of data are explored through data analysis to enhance the market segment for businesses. You can develop a tailored loyalty promotion, particularly for that consumer segment by looking at the links between factors like user age, gender, etc. Additionally, it may be used to predict which people are most likely to pay to a company, what their search histories indicate about their preferences, or what subscribers should include on email lists to boost response times.

Banks can find patterns in a group and uncover hidden links in data using data mining. The bank fully profiles each customer of the bank. The customer data includes specific items regarding the client's financial circumstances and spending habits before and after the credit was approved.

Directed learning is another name for classifiers. A dependent attribute or aim that was previously understood guides the learning process. A set of independent qualities or predictors are used in directed data mining to explain the target's behavior. Predictive models are typically the result of supervised learning. In contrast, the aim of unsupervised classification is pattern recognition.

A supervised model must be trained, which entails having the computer examine numerous instances where the target value is already known. The model "learns" the reasoning behind the prediction during the training process. For instance, a model that aims to predict which customers are most likely to respond to an offer must be trained by studying the traits of several individuals who are known to have previously responded or not to a campaign.

Data mining techniques have been applied in the banking industry for a number of purposes, such as

Predicting bank collapse [15–16], identification of potential bank customer churns [17], fraudulent transaction detection [18], customer segmentation [19-20], bank telemarketing predictions [21-24], sentiment analysis for bank customers [25], and bank loan prediction [26-28]. Some categorization studies in the banking sector are compared in Table 2. This table displays the aims of the previous researches, the years they were carried out, the ensemble learning techniques and the applied algorithms, the nationality of the bank, and the outcomes.

Table 2. Categorization studies in the banking sector

Reference	Year	Algorithms						Ensemble learning		Description	Country of the bank	Result
		DT	NN	SVM	KNN	NB	LR	Bagging (RF)	Boosting (AB, XGB)			
Rabihah Md. Sum et al[30]	2022			✓			✓			Personal loan applicants	Malaysia	AUC 0.64
Zahra Faraji[31]	2022	✓			✓		✓		✓	Credit card fraud detection	-	ACC 99%
Dhanashri A. et al[32]	2022	✓					✓	✓		Loan Approbation prediction	Afghanistan	ACC 88.53%
Doumpos M et al. [2]	2020			✓			✓		✓	Bank collapse expectation	USA	AUC 0.97
Khikmah L et al. [21]	2019	✓	✓	✓	✓	✓	✓	✓		Long-term deposit forecast	Portugal	ACC 97.07%
Huang J et al. [18]	2019		✓							Discovery of bank account fraud	-	ACC 97.39%
Ravi V et al. [25]	2019	✓	✓	✓	✓	✓	✓	✓	✓	Analysis of customer satisfaction for banks	India	AUC 0.826
Farooqi et al. [22]	2019	✓	✓	✓	✓	✓				Prediction of the results of telemarketing by banks	Portugal	ACC 91.2%
Climent F et al. [15]	2019						✓	✓	✓	Bank collapse expectation	USA	ACC 94.74%
Jing et al. [16]	2018		✓	✓			✓			Bank collapse expectation	USA	AUC 0.916
Lahmiri S. [23]	2017		✓							Prediction of the results of telemarketing by banks	Portugal	ACC 71%
Marinakos et al. [4]	2017	✓	✓	✓	✓	✓	✓			Classification of the bank's customers for marketing directly	Portugal	AUC 0.90
Ghaneei H et al. [17]	2016	✓								Prediction of customer attrition in banks	-	AUC 0.929
Yue et al. [29]	2016	✓		✓	✓			✓	✓	Forecasting loan defaults	China	AUC 0.965

All of the studies mentioned in the table above us only one method for categorization techniques in loan applications, rather than a combination of methods. Additionally, studies integrating the support vector machine (SVM) and the K-nearest neighbor algorithm (KNN) have also been conducted. This research aims to identify the algorithms used to detect fraud in the banking sector, specifically loan fraud.

Data mining may assist the banking sector in acquiring new customers and keeping hold of current ones. Customer recruitment and retention are issues that should concern any industry, but finance should be given special attention. Today's consumers have many different opinions regarding where to transact [17]. Executives in the banking industry must therefore be aware that if they do not give their consumers their entire attention, they can readily locate another bank that will. By delivering advantages tailored to each customer's needs and using data collection to identify target clients seeking services, a bank may be able to keep its existing clients and goods and learn about a customer's past purchasing habits. To prevent losing its profitable customers to other banks, Chase Madison Banks in New York started using data collection to review client accounts and alter its parameters for creating new accounts [25]. Medical centre Fleet Bank uses data mining to identify the most suitable new buyers for its equity investment offers. To determine which customers could be more likely to purchase a corporate bond, the bank analyses client demographics and financial accounts across several product lines. This data is then utilized to attract those people. With client profiles compiled through data mining, the contact centre for Bank of America is prepared to offer new services and offers that are most pertinent to each caller. Another difficulty is that the finances are client retention. Using slashing Web-based technologies, making predictions, and consumer advertising helps lenders attract new customers and retain existing ones.

According to the findings of this study, SVM and KNN may be used to forecast the likelihood that a customer will be approved for a loan.

4. FRAUD DETECTION IN BANKING SECTOR

In the Banking sector, data mining may be used to detect fraud. Since fraud detection is a priority for many businesses, data mining has increased the amount of fraud that is being identified and reported. Two separate methods have been developed by financial organizations to identify fraud trends.

In the first method, a bank employs data mining software and a third-party data warehouse to discover fraud trends. The bank can then check for indications of internal issues by comparing those patterns to its own database. The second method relies only on internal bank data to identify fraud patterns. The majority of banks take a hybrid approach [1]. The banking industry is putting in more effort to detect fraud. Fraud management requires a lot of expertise. Because it reveals which transactions were not allowed by the user, it is essential in the identification of fraud.

Advertising is one of the often-used data mining applications in the banking sector. In order to evaluate customer information and create statistically accurate portraits of customers' preferences for items and services, banks' marketing teams can utilize data mining [42]. By just offering the goods and services that customers genuinely want, banks can save a significant amount of funds on marketing and discounts that would be useless [32]. As a result, bank marketers must focus on their customers by learning more about them. Obviously, to increase sales and improve service quality, Citigroup uses marketing technology. Due to the unification of four years' supply of client history paperwork, the business was able to market to and provide consumers with tailored services.

A key use of data mining in the financial system is fraud detection. Too many businesses are concerned about being able to spot fraudulent conduct, and data analysis is helping to find and report more suspicious transactions. Two unique techniques have been developed by financial organizations to spot fraud patterns. The first technique involves a bank obtaining a third coalition's data center, which may include metadata from several organizations, and using information retrieval techniques to identify fraud patterns. The bank can then check for any signs of internal problems by comparing those characteristics to its own database. In the second method, only data that the bank already has is used to identify fraud trends. The vast majority of banks use a "hybrid" approach. One approach that has been successful in detecting fraud is Falcon's "fraud assessment." It examines the activity for 60% of the cards that consumers in the country have, and it is used by nine of the top ten banks that offer credit cards. Mellon Bank can better safeguard itself and its customers' assets from prospective fraudulent transactions by using data mining for fraud prevention.

This section provides a review of previous research on fraud detection in the loan banking sector.

Goyal et al. [1] provided an example of several extensively used DM and MLT for detecting credit card fraud. Investigations into credit card fraud have been done in several ways. It began by outlining the significance of the subject and the present limitations in customary practices. The danger associated with counterfeit transactions varies; hence it is important to develop efficient and precise methods for identifying high-risk transactions. Standard data mining techniques are insufficient to identify these transactions. To find the best solution, advanced algorithms should be used.

The examination of each feature's information gain ratio serves as the foundation for feature selection. The division and conquer technique is imitated in the construction of the lower directions in data mining, which also follows the same process of information gain evaluation. Three essential components are included while creating a technique for solving a problem.

The technique of extracting knowledge from vast quantities of unstructured data is known as data mining. The learning must be fresh, obscure, applicable, and defensible in the field in which it was discovered.

Singh et al. [33] proposed a structure based on three-layer verification techniques. In order to detect and reduce fraudulent credit card applications and transactions, they used threshold values, a genetic algorithm, community detection, and spike detection to accomplish this. The outcomes demonstrate the superiority of the suggested methodology. This essay also covers an extensive list of security recommendations that have helped credit card customers avoid fraudulent actions. The framework can make it easier for a rookie researcher to spot credit card (financial) fraud.

Kavipriya et al. [34] developed a system that uses effective clustering and classification techniques like apriori and support vector machines to analyze, spot, and identify fraudulent transactions. The results show that the proposed method outperforms the existing hidden Markov model in terms of fraud coverage while also having a low false alarm rate. The credit card fraud detection system was discussed in this thesis. The suggested technique has involved extensive testing on several different kinds of transactions. The findings were encouraging; nearly all fraudulent transactions were successfully identified, and when the new technique was compared to the current method, the outcome showed that it outperformed the latter.

Suresh et al. [35] presented a survey that represents a systematic examination of data mining techniques and how they are used in the processing of credit cards. The primary focus of the study was on data mining techniques, especially as they are used in credit card processing, which helps to spotlight considerably bigger components. As a result, this survey should be highly helpful for both academics doing a thorough evaluation of the literature in their subject and credit card companies choosing an effective solution for their problem. This article served as an overview of methods for identifying fraudulent credit card use and credit card fraud. Missions for categorization and prediction are particularly important in the credit card procedure.

A reliable and serious approach for identifying accounting information was put out by Yao et al. [36]. Businesses with a fraudulent financial statement (FFS) and non-FFS cases between 2002 and 2013 are their research subjects. Support vector machine (SVM), decision tree (DT), artificial neural network (ANN), and bayesian belief network (BBN) were utilized to detect FFS. Conventional statistical methods, such as regression models, have a greater mistake rate than data mining methods.

Zhou et al. [37] surveyed the evidence on financial products and fraud detection methods based on ensemble, transfer, supervised, unsupervised, and semi-supervised techniques. The bulk of fraud detection systems has been found to employ at least one supervised learning technique. There are still challenges to be resolved in this area, though. Beginning with feature engineering, parameter selection, and hyper parameter tweaking, data mining-based credit scoring and fraud detection encounter the same challenges as other categorization tasks. Second, it is practically hard to describe complicated financial situations, particularly those in China, and researchers do not have access to sufficient public data to train and evaluate their models.

Talavera et al. [38] developed a modern strategy for using computational intelligence approaches to identify fraudulent credit transactions. To train a classification and clustering algorithm, they divided a dataset of historical customer data from a financial institution. They train a radial basis function network to assess if a client engages in credit fraud. Then, to create customer profiles, they construct a fuzzy c-means clustering. This algorithm can give a degree of membership in the form of points outside of clusters and group data inside clusters.

Dushyant Singh et al. [39] focused on analyzing credit card transaction datasets to identify patterns that fraudulent credit card transactions follow to help implement and design a credit card fraud detection algorithm. The data is analyzed by creating histograms of the variables in the dataset and a correlation matrix of the variables. The data analysis also assists us in determining the best machine-learning techniques to use when implementing the algorithm. The algorithm is implemented using the local outlier factor machine learning technique, which shows that this technique is highly accurate but there is a low accuracy in detecting credit card fraud. The algorithm can detect fraudulent transactions that match the pattern.

A comprehensive program for detecting and preventing fraud must include quality as the key for managers and employees to fraud awareness. Employee tips are the primary source of professional fraud detection. Actually, my study demonstrates that companies with anti-fraud training programs for leaders, executives, and personnel have lower losses and shorter scams than companies without such initiatives. At the very least, staff members should be informed about what constitutes fraud, how it negatively affects everyone in the business, and how to report suspicious activity.

Kirkos E. et al. [40] investigated the efficiency of several classification methods of fraudulent financial statement detection (FFS) using data mining (DM).

Researchers also determined the main FFS potential risks. The authors used neural networks, decision trees, and bayesian belief networks for classification (BBN). Different ratios and variables derived from financial statements are used as input for their experiments, such as total assets, working capital, sales to total assets, net income, quick assets, liabilities, fixed assets to total assets, and earnings before interest and taxes. They also concentrated on management fraud, which is induced by managers in order to meet targets while concealing losses or debt. Financial distress, they claim, is also a motivator for management fraud.

Various forms of fraudulent activity were categorized by West et al. [41] as follows: 1) Economic fraud 2) Business fraud the theft of insurance. Bribery, financial crimes, and fraudulence were their additional three classifications for banking fraud. Corporate fraud also includes deception involving financial statements, commodities, and securities. Fig.2 displays various variations in fraud detection methods. Additionally, they categorize financial crimes into two groups: healthcare fraud and motor insurance fraud. They said that standard audit committee fraud prevention is no longer feasible in the era of big data. In the given figure, the blue color represents health care fraud and the green color represents motor insurance fraud.

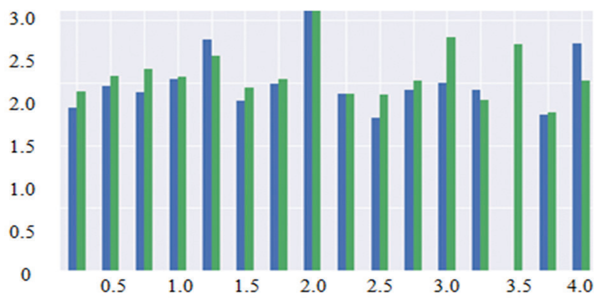

Fig. 2. Fluctuations in fraud detection tasks

Finally, we found out numerous challenges in the detection of financial fraud from table 3, as following:

- Financial fraud is an ever-changing field. Being one step ahead of the offenders is essential.
- Financial fraud detection methods vary depending on the situation.
  - There is no one-size-fits-all approach to data mining.
  - Sometimes hybrid methods are more effective.
  - Parameter tuning improves results. It took a lot of trial and error to find the best set of parameters.
- Privacy concerns led to corporate reluctance to share information, which resulted in various experimental limitations such as under-sampling.
- To avoid loss, financial fraud requires near-real-time detection.
- Misclassification has a financial and/or business cost. As a result, more emphasis should be placed on performance (accuracy and time) versus misclassification cost.
- Having a common guideline that can handle fraud cases across several domains could be advantageous.
- The financial dataset is highly skewed [a few fraudulent transactions among millions of transactions].

Besides the previously mentioned challenges, traditional approaches have limitations.

Every lending decision made by a bank involves some level of risk. This risk may be quantified to facilitate risk management and lower the possibility of financial loss for the bank. Credit management can make better judgments if they are aware of the repayment capacity of their customers.

Additionally, data mining may be used to identify whether clients would skip or default on loan installments. This increased information can help the bank make the necessary corrections to prevent losses. Consideration should be given to behavioral patterns, balance sheet numbers, turnover trends, limit use, and check return patterns when making such forecasts. Default patterns from the past can also be utilized to anticipate future defaults when comparable patterns are found.

To increase the accuracy of credit score predictions and decrease default probability, data mining techniques are applied. A borrower's creditworthiness is shown by their credit score. To predict a client's future behavior in a range of circumstances, behavioral scores are created using probability models of customer behavior. By examining the borrower's previous debt repayment practices and the accessible credit history, data mining can determine this score, which will be used in our subsequent articles.

Table 3. Summary of fraud detection in banking sector

Author	Year	Proposed Work	Gap/ Future Work	Classification / Approach Used	Other Model / Techniques
Zahra Faraji[31]	2022	This study aims to present the commonly used supervised algorithms for fraud detection. In addition, this work aims to apply specific strategies, evaluate how well they work on actual data, and create an ensemble model as a viable solution to this problem.	The data is only one of the drawbacks of this study. The results of this study cannot be generalized to all banks or financial institutions because the data were limited to a single financial institution. Future studies could investigate machine learning methods with larger data sets. Another disadvantage is that unsupervised techniques were not used in this study.	decision tree, logistic regression, KNN ,random forest, and XGBoost	Not Used
M Sathy et al [42]	2019	They suggested approach (the credit card expenditure model) is better at lowering the percentage of false alarms since it looks at the correlation between fraudulent transactions and those that are only suspected of fraud.	Future work on this study is By integrating more rules in the rule engine that models expand, the system's accuracy may be increased (Hidden Markov and K-mean clustering).	Not Used	Clustering, Hidden Markov Model, k-mean
Janaki K. et al [43]	2019	They suggested a mechanism to identify and stop fraudulent transactions and acts to lessen the economic industry's unit of drop.	The restrictions are that Web-based research models are utilized to reach out by taking these tactics into account. On the other hand, you may investigate the alternative models online. The location of extortion cases may be rapidly determined by the internet.	Naïve Bayes, Decision Tree, Random Forest	Not Used
Aswathy et al [44]	2018	They frequently used techniques such as genetic algorithms, rule induction, SVM, ANN, decision tree, and logistic regression. The neural network is the most commonly used algorithm for detecting fraud. Furthermore, these algorithms can be used singly or in combination to create models.	Not Mentioned	SVM, Logistic Regression, Neural Network, Decision Tree.	Rule Induction, Genetic Algorithm
S.Vimala et al [45]	2017	They presented a research paper on using data mining to identify fraud in credit cards. They conclude that the optimum method for detecting fraud is to use a hidden Markov model, and decision tree.	The clustering algorithm and Markov chain model need to be enhanced further to prevent future fraud.	Neural Network, Decision Tree	Genetic Algorithm, Hidden Markov Model, K-mean

5. FINDINGS

The study's most recent findings investigate the fraud detection tools currently in use. We examined indicators for model performance that were based on the statistics. We demonstrated the application of data mining approaches to model evaluation. Considering evaluation, performance metrics including accuracy, recall, and sharpness are obtained. Analysis using data mining techniques offers a very visual summary of a model's performance. It is significant with regard to class skew, making it a trustworthy performance metric in numerous significant application domains for fraud detection. A comprehensive program for preventing and detecting fraud must include targeted training for managers and employees on fraud awareness. Employee tips are the primary source of occupational fraud detection, but my study also demonstrates that companies with anti-fraud training programs for managers, ceos, and personnel have lower losses and shorter scams than companies without such initiatives. Staff members ought to be informed at a minimum about what constitutes fraud, how it hurts the corporation as a whole, and how to report suspicious activity.

6. CONCLUSIONS

Data mining is a technique that supports the banking and retail industries for making better decisions by sifting through the vast amount of already available evidence to find particular information. Data processing condenses a variety of information into a manageable form so that it may be mined. The organization as a whole uses research methodology to gather data to support for decision-making. The financial industry can benefit significantly from the use of data mining techniques for better client acquisition, automatic lending for fraud detection, real-time fraud detection, segment product design, analysis of existing money transfer data for better client service and client retention, risk management, and advertising.

To conclude, loan fraud prevention became more and more challenging as research develops and the variety of commitment to giving rises. The conventional auditing approach for detecting loan fraud is no longer applicable because it is manual, labor-intensive, expensive, and inaccurate. Fraud is a serious issue in the financial industry. Fraudsters always devise new schemes. As they attempt to avoid detection, the plans become more intricate, making it difficult to detect and stop fraud. This paper aims to provide a comprehensive analysis of fraud detection and data mining applications in the banking sector. A framework will soon be proposed as an improvement over the limitations offered by the approaches examined for the study.

7. REFERENCES:

[1] R. Goyal, A. Kumar, "Review on Credit Card Fraud Detection using Data Mining Classification Techniques & Machine Learning Algorithms", International Journal of Research and Analytical Reviews, Vol. 7, No. 1, 2020, pp. 972-975.

[2] M. Georgios, M. Doumpos, C. Zopounidis, E. Galariotis. "An Ordinal Classification Framework for Bank Failure Prediction: Methodology and Empirical Evidence for US Banks", European Journal of Operational Research, Vol. 282, No. 2, 2020, pp. 786-801.

[3] A. Gupta, V. Pant, S. Kumar, P. K. Bansal, "Bank Loan Prediction System using Machine Learning", Proceedings of the 9th International Conference on System Modeling and Advancement in Research Trends, Moradabad, India, 4-5 December 2020, pp. 423- 426.

[4] G. Marinakos, S.Daskalaki, "Imbalanced customer classification for bank direct marketing", Journal of Marketing Analytics, Vol. 5, No. 1, 2017, pp. 14-30.

[5] B. Baesens, V. Van Vlasselaer, W. Verbeke, "Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection", John Wiley & Sons, 2015.

[6] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, "Data mining for credit card fraud: A comparative study", Decision support systems, Vol. 50, No. 3, 2011, pp. 602-613.

[7] A. Abdallah, M. A. Maarof, A. Zainal, "Fraud detection system: A survey", Journal of Network and Computer Applications, Vol. 68, 2016, pp. 90-113.

[8] M. Kharote, V. P. Kshirsagar, "Data mining model for money laundering detection in financial domain", International Journal of Computer Applications, Vol. 85, No. 16, 2014.

[9] O. Gottlieb, C. Salisbury, H. Shek, V. Vaidyanathan, "Detecting corporate fraud: An application of machine learning", A publication of the American Institute of Computing, 2006, pp. 100-215.

[10] K. Golmohammadi, O. R. Zaiane, "Data mining applications for fraud detection in securities market", Proceedings of the IEEE European Intelligence and Security Informatics Conference, Odense, Denmark, 22-24 August 2012, pp. 107-114.

[11] M. A. Sheikh, A. K. Goel, T. Kumar, "An approach for prediction of loan approval using machine learning algorithm", Proceedings of the IEEE International Conference on Electronics and Sustainable Communication Systems, 2-4 July 2020, pp. 490-494.

[12] J. Sanjaya, E. Renata, V. E. Budiman, F. Anderson, M. Ayub, "Prediksi Kelalaian Pinjaman Bank Menggunakan Random Forest dan Adaptive Boosting", Jurnal Teknik Informatika Dan Sistem Informasi, Vol. 6, No. 1, 2020, pp. 50-60.

[13] K. Gupta, B. Chakrabarti, A. A. Ansari, S. S. Rautaray, M. Pandey, "Loanification- Loan Approval Classification using Machine Learning Algorithms", Proceedings of the International Conference on Innovative Computing & Communication, 24 April 2021, pp. 1-4.

[14] A. Kulothungan, "Loan Forecast by Using Machine Learning", Turkish Journal of Computer and Mathematics Education, Vol. 12, No. 7, 2021, pp. 894-900.

[15] F. Climent, P. Carmona, A. Momparler, "Predicting failure in the US banking sector: An extreme gradient boosting approach", International Review of Economics & Finance, Vol. 61, 2019, pp. 304-323.

[16] Z. Jing, Y. Fang, "Predicting US bank failures: A comparison of logit and data mining models", Journal of Forecasting, Vol. 37, No. 2, 2018, pp. 235-256.

[17] H. Ghaneei, A. Keramati, S. M. Mirmohammadi, "Developing a prediction model for customer churn from electronic banking services using data mining", Financial Innovation, Vol. 2, No. 1, 2016, pp. 1-13.

[18] J. Huang, W. Wang, Y. Wei, Y. Sun, "A two-route CNN model for bank account classification with heterogeneous data", PlosOne, Vol. 14, No. 8, 2019, p.e0220631.

[19] I. Smeureanu, G. Ruxanda, L. M. Badea, "Customer segmentation in private banking sector using machine learning techniques", Journal of Business Economics and Management, Vol. 14, No. 5, 2013, pp. 923-939.

[20] F. N. Ogwueleka, S. Misra, R. Colomo, L. Fernandez, "Neural network and classification approach in identifying customer behavior in the banking sector: A case study of an international bank", Human factors and ergonomics in manufacturing & service industries, Vol. 25, No. 1, 2015, pp. 28-42.

[21] L. Khikmah, A. Ilham, A. Indra, "Long-term deposits prediction: a comparative framework of classification model for predict the success of bank telemarketing", Journal of Physics: Conference Series, Vol. 1175, No. 1, 2019, p. 12035.

[22] R. Farooqi, N. Iqbal, "Performance evaluation for competency of bank telemarketing prediction using data mining techniques", International Journal of Recent Technology and Engineering, Vol. 8, No. 2, 2019, pp. 5666-5674.

[23] S. Lahmiri, "A two-step system for direct bank telemarketing outcome classification", Intelligent Systems in Accounting, Finance and Management, Vol. 24, No. 1, 2017, pp. 49-55.

[24] S. Moro, P. Cortez, P. Rita, "A data-driven approach to predict the success of bank telemarketing", Decision Support Systems, Vol. 62, 2014, pp. 22-31.

[25] V. Ravi, G. Krishna, B. Reddy, M. Zaheeruddin, "Sentiment classification of Indian Banks' Customer Complaints", Proceedings of the IEEE in 10th Annual International Conference, Kochi, India, 17-20 October 2019, pp. 429-434.

[26] H. Ramachandra, G. Balaraju, R. Divyashree, H. Patil, "Design and Simulation of Loan Approval Prediction Model using AWS Platform", Proceedings of the IEEE International Conference on Emerging Smart Computing and Informatics, 5 March 2021, pp. 53-56.

[27] M. Alaradi, S. Hilal, "Tree-Based Methods for Loan Approval", Proceedings of the International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy, 2020, pp. 1-6.

[28] M. A. Sheikh, A. K. Goel, T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm", Proceedings of the IEEE International Conference on Electronics and Sustainable Communication Systems, Coimbatore, India, 2-4 July 2020, pp. 490-494.

[29] Z. Yue, J. Wan, D. Yang, Y. Zhang, "Predicting non performing loan of business Bank with data mining techniques", International Journal of Database Theory and Application, Vol. 9, No. 12, 2016, pp. 23-34.

[30] A. D. Wani, S. Adarsha, "In Banking Prediction of Loan Approbation Using Machine Learning", International Journal of Research Publication and Reviews, Vol. 3, No. 5, 2022, p. 7421.

[31] Z. Faraji, "A Review of Machine Learning Applications for Credit Card Fraud Detection with A Case study", SEISENSE Journal of Management, Vol. 5, No. 1, 2022, pp. 49-59.

[32] R. Sum, M. Rabihah, "A New Efficient Credit Scoring Model for Personal Loan using Data Mining Technique Toward for Sustainability Management", Journal of Sustainability Science and Management, Vol. 17, No. 5, 2022, pp. 60-76.

[33] Singh, Ajeet, Anurag, "A Novel Framework for Credit Card Fraud Prevention and Detection (CCFPD) Based on Three Layer Verification Strategy", Proceedings of the ICETIT: Emerging Trends in Information Technology, 2020, pp. 935-948.

[34] T. Kavipriya, N. Geetha, "An identification and detection of fraudulence in credit card fraud transaction system using data mining techniques", International Research Journal of Engineering and Technology, Vol. 5, No. 1, 2018.

[35] G. Suresh, R. J. Raj, "A Study on Credit Card Fraud Detection using Data Mining Techniques", International Journal of Data Mining Techniques and Applications, Vol. 7, No. 1, 2018, pp. 21-24.

[36] J. Yao, J. Zhang, L. Wang, "A financial statement fraud detection model based on hybrid data mining methods", Proceedings of the IEEE International Conference on Artificial Intelligence and Big Data, Chengdu, China, 26-28 May 2018, pp. 57-61.

[37] X. Zhou, P. Xu, "A state of the art survey of data mining based fraud detection and credit scoring", Proceedings of MATEC Web of Conferences, EDP Sciences, Vol.189, 2018, p. 3002.

[38] Talavera, Alvaro, "Data Mining Algorithms for Risk Detection in Bank Loans", Proceedings of Springer Annual International Symposium on Information Management and Big Data, 2018, pp. 151-159.

[39] D. Singh, S. Vardhan, N. Agrawal, "Credit Card Fraud Detection Analysis", International Research Journal of Engineering and Technology, Vol. 5, No. 11, 2018, pp. 1600-1603.

[40] E. Kirkos, C. Spathis, Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements", Expert systems with applications, Vol. 32, No. 4, 2007, pp. 995-1003.

[41] J. West, M. Bhattacharya, "Intelligent financial fraud detection practices: an investigation", Proceedings of the Springer International Conference on Security and Privacy in Communication Networks, 2014, pp. 186-203.

[42] M. Sathyapriya, V. Thiagarasu, "A Cluster Based Approach for Credit Card Fraud Detection System using HMM with the Implementation of Big Data Technology", International Journal of Applied Engineering Research, Vol. 14, No. 2, 2019, pp. 393-396.

[43] K. Janaki, V. Harshitha, S. Keerthana, Y. Harshitha, "A Hybrid Method for Credit Card Fraud Detection Using Machine Learning Algorithm", International Journal of Recent Technology and Engineering, Vol. 7, No. 6S4, 2019, pp. 235-239.

[44] M. S. Aswathy, L. Sameul, "Survey on Credit Card Fraud Detection", International Research Journal of Engineering and Technology, Vol. 05, No.11, Nov 2018, pp.1291-1294.

[45] S. Vimala, K. C. Sharmili, "Survey Paper for Credit Card Fraud Detection Using Data Mining Techniques", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, No. 11, 2017, pp.357-364.