International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I12P111

Volume 72 Issue 12, 89-99, December 2024
© 2024 Seventh Sense Research Group®

Review Article

Evaluation of Financial Data Processing Life Cycle for Risk Prediction: A Survey

M.V. Narayana[1], T. Sumallika[2], M. Vijaya Sudha[3], P.V.M. Raju[4]

[1]*Department of CSE, Guru Nanak Institutions Technical Campus, Telangana, India.*
[2]*Department of IT, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, India.*
[3]*Department of IT, Sir C.R. Reddy College of Engineering, Andhra Pradesh, India.*
[4]*Department of Business and Management Studies, Seshadri Rao Gudlavalleru Engineering College, Andhra Pradesh, India.*

[1]*Corresponding Author : mvnarayanacse@gmail.com*

Received: 31 October 2024 Revised: 26 November 2024 Accepted: 11 December 2024 Published: 30 December 2024

Abstract *- Financial risk analysis is integral to financial planning and investment at organizational and personal levels. Due to the higher fluctuation of the financial trend, many inverters consider the risk prediction strategy during the investment portfolio generation. The risk prediction for financial assets is highly challenging due to the dependency of financial trends on various technical and non-technical factors. Hence, the use of computer-aided processes is becoming popular for risk prediction. Recently, with the enhancements in machine learning algorithms, the risk prediction processes have improved the accuracy of the prediction. These algorithms have two phases: training the model and deploying the models to predict. Nonetheless, the available machine learning algorithms for risk prediction have many limitations. The limitations primarily concern the correctness of the data to be deployed for building the predictive model for prediction as these data are collected from various sources, sometimes with human interventions, and are prone to insufficient and incorrectness. Hence, the frameworks or the processes for financial predictions must perform an additional step, such as data pre-processing, and then further perform the actual task, risk predictions. In the recent past, a good number of research works have aimed to predict financial risks with higher accuracy by designing a complete life cycle of the data for financial predictions, starting from data pre-processing to the conclusion of risk analysis. Nevertheless, these works are criticized for not performing the prediction task with the best possible accuracy and compromising on the time complexity, as time complexity can be a critical measure of performance in financial risk analysis. Hence, this work aims to analyze the various strategies and works for data pre-processing and predictions on financial data. This work finally contributes to the research domain by analyzing the strategies mathematically, algorithmically and result wise to identify the unsolved challenges in this domain.*

Keywords *- Computational modelling, Data analytics, Data collection, Machine Learning, Risk prediction.*

1. Introduction

Financial risk prediction is one of the key components of banking and personal investment strategies. Organizations collect data from the complete process of their banking risk prediction and further process the data using dedicated frameworks for risk analysis and predictions. These frameworks or processes are widely applied and can be adapted for various purposes. Hence, banking organizations always seem to adopt the most sophisticated framework. Most investment organizations consider the outcomes of these frameworks to identify risks, such as bankruptcy or insolvency, for the investing organization or product before investing. The work by N. Yerashenia et al. [1] has listed numerous examples.

The task of analysis and prediction of risk analysis for financial domains is highly critical because the trend of the financial processes depends on a wide variety of parameters. These parameters can be internal, such as organization policies or outcomes from a previous policy or completely external, such as the influence of an external policy as government regulations.

Thus, identifying the right characteristics to analyse the data for prediction is highly complex and cannot be done manually. Thus, higher-performing computing algorithms from the domain of machine learning are used.

The work by H. Son et al. [2] has justified using machine learning techniques for such purposes. With a moderate number of corrections and customizations, these machine learning methods can be highly effective for risk analysis and predictions with higher accuracy and timely outcomes. Nonetheless, financial organizations use these algorithms for risk analysis or predictions and other information expected to be concluded from these algorithms. Thus, the demand for a framework to be deployed on such aspects constantly grows. The work by H. A. Alaka et al. [3] clearly indicated the need for such frameworks and elaborated on the desirable outcomes. It is evident from the work by Y. Roh et al. [4] that the accuracy or the correctness of the predictions (Figure 1) by these frameworks strongly depends on the quality or cleanliness of the data to be deployed to train such models. Hence, the demand for proper and justified data cleaning or pre-processing before deploying to the framework is increasing.

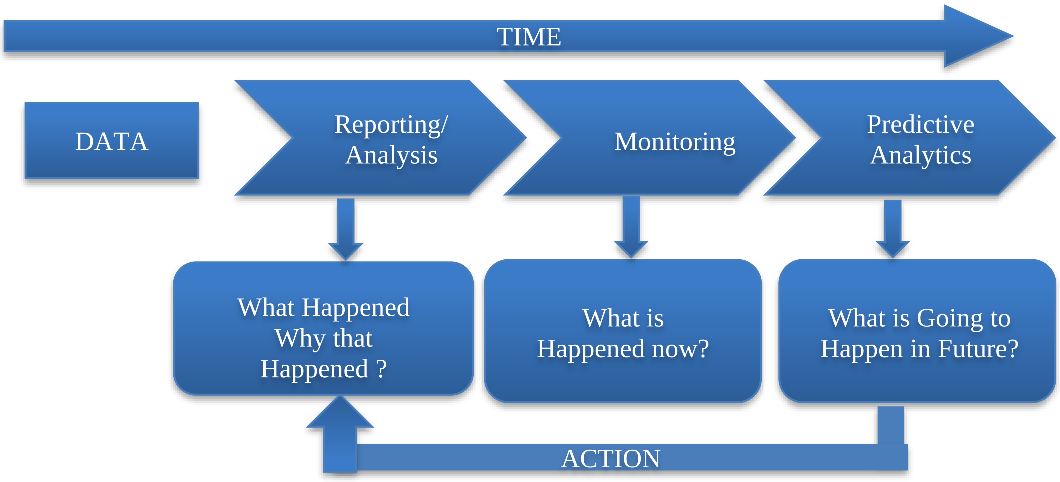


Fig. 1 Predictive analysis

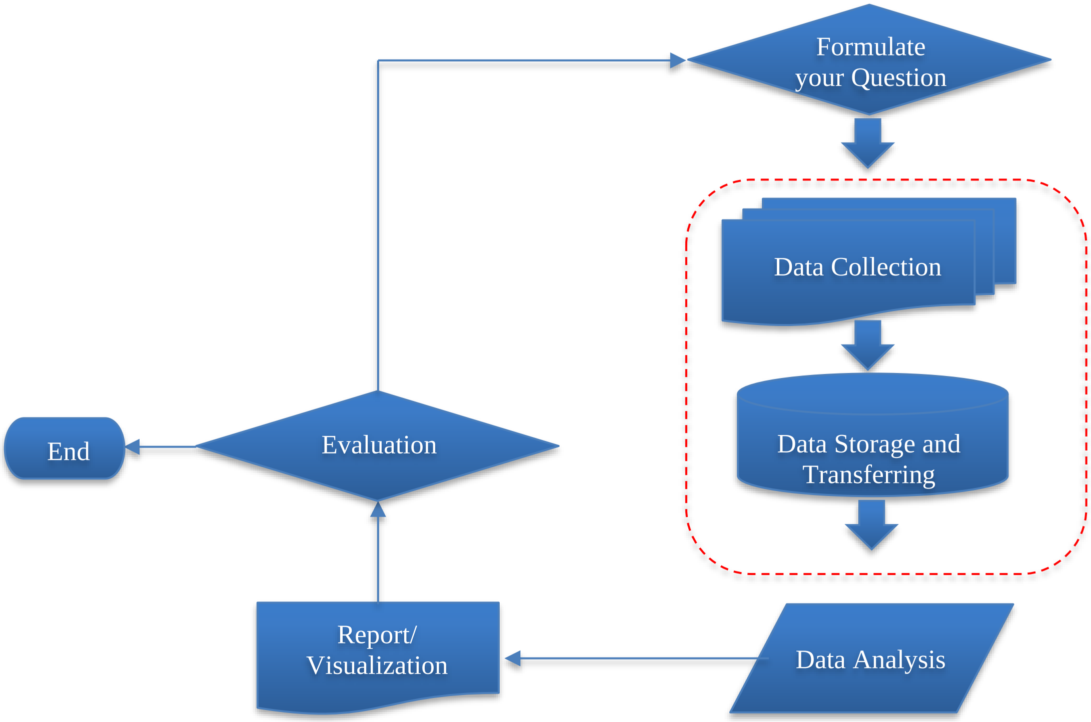


Fig. 2 Data processing framework for financial risk prediction [5]

Henceforth, in this work, the complete analysis of the data processing lifecycle is carried out based on four major life cycle phases such as missing value identification and imputations, outlier detection and replacements, noisy data identification and replacements and finally, the available models for analysis and predictions of the risks. First, this work analyses the complete data processing life cycle in Sections II and III, using mathematical models to analyse the basic methods for performing the data processing life cycle. Further, based on the mathematical model understanding of the life cycle, critically analyses the existing works by various researchers in Section - IV and immediately identifies the challenges found in all these methods in Section – V. Also, a few research outcomes are analysed after implementing the existing algorithms proposed by these researchers in Section – VI; the final survey conclusion is furnished in Section VII.

2. Data Processing Life Cycle and Methods

With the context and foundational need for this research domain to identify the most suitable method for financial risk analysis in the previous section of this work, this section visualizes the generic framework for data processing for financial risk analysis.

Many parallel research outcomes have identified the ideal framework or process description to manage the data most effectively and increase the accuracy of the analysis and predictions. Tao Huang et al. [5] have furnished the framework (Figure 2) and explained various phases of this research.

As demonstrated in the above diagram, this process uses machine learning methods. The initial phase, which is highly evident in this claim, focuses on identifying the questions or data analytics hypothesis identification or formulation. The hypothesis driven analysis is only adopted in the case of machine learning driven methods such as correlations.

Once the hypothesis is furnished, which will be validated in the later stages of the process, the next immediate stage is to collect the information and further transform or transfer the collected data to the storage segments for final processing. During the data collection process, one of the most important tasks takes place, which is data cleaning or data pre-processing.

The need for data pre-processing is significant and notably mentioned in all the parallel research attempts as the accuracy of the final outcome, in this case, the risk analysis and prediction, relies partly on the correctness of the data.

Many parallel research outcomes have also mentioned that the process for correcting the data, the pre-processing, can be adopted just before storing the data. This process is also very popular, as is a sub-process called extract–transform – Load or the ETL process. Here, the transform phase takes care of the impurities of the data as,

- Missing Values: The missing values, or the MV referred to by many authors, are missing data points in the collected data. This missing value impurity can happen for various reasons, such as incorrect strategies for data collection, human error, or customized processes, where the data point for a specific case cannot be collected. Nonetheless, the impact of the missing values on the final analysis outcome can be very high and lead to higher inaccuracy of the machine learning models as these missing values can completely hide some aspects of the data, or process, from which the data is collected, or characteristics of the organization.
- Outliers: The outliers are some data points beyond the expected values of any data characteristics. These values are not impossible. Instead, these values are random incidents entirely irrelevant to machine learning algorithms modelling. Also, including these outliers during the model design can demonstrate higher deviation from most of the other data points, which are regular data.
- Noisy Data: Noisy data are rare for textual data analysis projects or research as the noisy data points can primarily be seen in the datasets collected using some devices or sensors. Hence, this research does not focus on these types of impurities.

After the pre-processing phases are completed, the primary data analysis task can be adopted. Many researchers have demonstrated various methods, some hybrid methods that produce the most effective outcomes. In the further sections of this survey, all methods and phases are discussed in detail.

3. Fundamental Process - The Mathematical Models

After understanding the fundamental process of data analysis for risk analysis and prediction in the previous section of the work, a deeper understanding of each phase is realized using mathematical models in this section.

Many parallel research outcomes have significantly demonstrated the benefits of realizing problems and identifying solutions using mathematical models. TA Lin et al. [6] demonstrated the benefits of process modelling and risk identification for internet-based transactions using mathematical models. The designed factors for risk identification using the parametric approach can only be realized using mathematical models and can be highly beneficial for identifying the problems to be addressed. Few of the parallel research have also stated that the risk modelling fundamentals can be highly complex during analyzing the micro-level transactions as the components in the mathematical models can be multiple, and identifying the relations between these parameters can be challenging, as demonstrated in the work by MA Li et al. [7]. Nonetheless, this is one of the distinct cases, and this work does not particularly focus on microfinance.

Another benefit of the financial models is the applicability of the process in various other domains, such as manufacturing or production, where the same risk analysis model can be deployed. The work by LI Yuanyuan et al. [8] is one of the notable proofs of this claim. Also, the mathematical models can be of real benefit for identifying the connections between various other entities in the organization, as showcased in the work by YANG Songling et al. [9]. In the deeper aspects of the financial mathematical models (Figure 3), it is often observed that the components used in the predictive analysis carry various importance or weightages. With a regular method of analysis or modelling, considering the weightages can be highly challenging. However, this can be realized easily in mathematical modelling methods, as suggested by LI Hua et al. [10]. This can influence the correctness of the analysis, as showcased by BAI Xue et al. [11] and YE Li et al. [12].

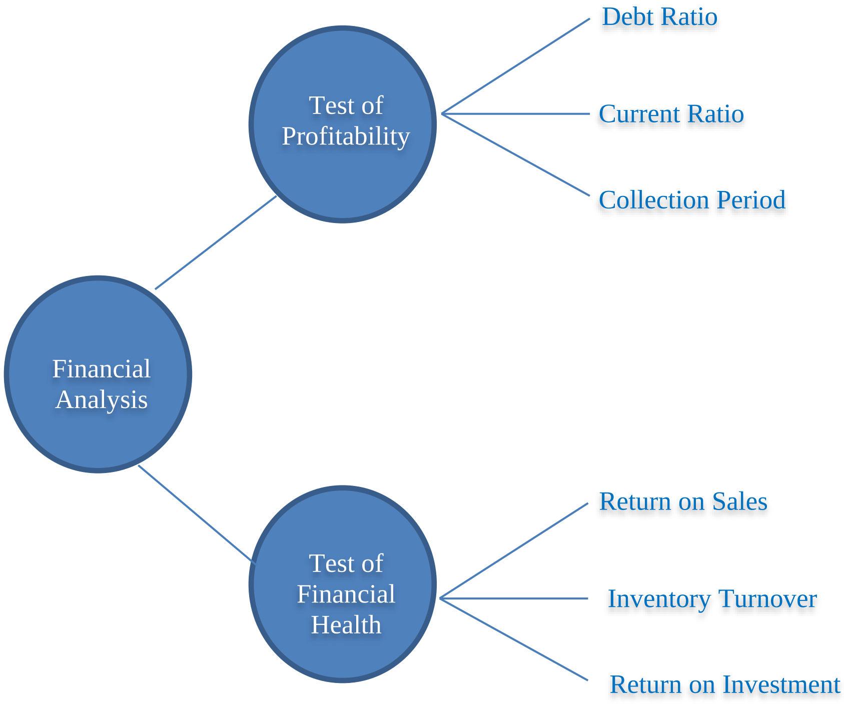


Fig. 3 Financial analysis

On the other hand, many researchers have also highlighted other benefits of the mathematical model-based analysis, such as the use of probabilistic analysis as demonstrated in the work by ZHU Xiaoqian et al. [13], and in the work by LIU Zhiyang et al. [14].

Hence, at this point of the discussion, it is evident that the mathematical model for analysing the existing systems is highly beneficial for identifying the problems. Thus, this section of the work analyses the foundational mathematical models for the risk prediction models.

3.1. Missing Value Detection and Imputations

As seen in many parallel research outcomes, the first step during the data processing or modelling method is treating the missing values in the dataset. Hence, the mathematical model for missing value analysis is furnished and discussed here.

Assuming that the complete dataset, FD, under analysis is a collection of multiple attributes, A[], and any individual attribute can be identified as A_i[]. Thus, for n number of total attributes, the relation can be formulated as,

$$FD = A[] \tag{1}$$

And,

$$FD = < A_1[], A_2[], A_3[], \ldots.., A_n[] > \tag{2}$$

Further, assuming that each attribute has its domain, D_X[], each domain consists of individual data items such as DI_X. Thus, for m number of elements in each domain, this can be formulated as,

$$A_i[] = < D_i[] > \tag{3}$$

And,

$$D_i[] = < DI_1, DI_2, \ldots.DI_m > \tag{4}$$

Here, the identification of the missing values can be fairly easy as any data item, DI_X, may contain a null value and can be detected as,

$$DI_X \Rightarrow \Phi \tag{5}$$

Thus, this process must be replicated for the complete dataset as,

$$\Phi[][] = \prod_{DI_X \Rightarrow \Phi} A[] \tag{6}$$

Here, $\Phi[][]$ denotes the collection of attribute indexes from the entire dataset with missing values.

Further, the identified missing value must be imputed with a value generated from the domain of that specified attribute. The generic method suggests that the mean value of that attribute's specified domain acts as an imputation value. This means MVE, a value from the attribute domain, can be formulated as,

$$MVE(x) = \sum_{x=1}^{m} DI_X \Big/ \lambda(A[x]) \tag{7}$$

Here, $MVE(x)$ is the mean value for the domain X or A_X.

Further, the missing value must be imputed as,

$$DI_X \Leftarrow MVE(x) \tag{8}$$

The generic problems are discussed in Section V of this work.

3.2. Outlier Analysis and Imputation

After the missing value treatment, the next important task is identifying the outliers and performing suitable treatments to clean the data.

Assuming that, from Equations 4 and 7, any data item that is above the expected range of the specified attribute shall be considered as an outlier, O_x, and can be formulated as,

$$O_x \Leftarrow \forall DI_X > MVE(x) \tag{9}$$

Or,

$$O_x \Leftarrow \forall DI_X > [\sum_{x=1}^{m} DI_X \Big/ \lambda(A[x])] \tag{10}$$

Further, the outlier imputation process in the generic method is similar to the missing value imputation process, as demonstrated using Equation 8. The generic problems are discussed in Section V of this work.

3.3. Risk Prediction

Finally, after the relevant data cleaning process, the most relevant stage in the life cycle will be the risk analysis and prediction. Continuing from Equation 1, assuming that out of the total n number of attributes in the dataset, only one element denotes the risk level, and the rest of the (n-1) attributes or parameters are independent attributes. Thus, the final clustering process for the risk analysis will be driven by one attribute, assuming the nth attribute. This can be formulated as,

$$FD =< A_1[], A_2[], A_3[], \ldots.., A_{n-1}[], A_n[] > \tag{11}$$

Further, assuming that XX is the function to identify unique elements in the $A_n[]$ set, then this can be formulated as,

$$G[] = f\theta(A_n[]) \tag{12}$$

These unique values, G[], denote each dataset group level and, in this case, the risk levels. Once the unique values are identified, the complete dataset can be separated into G[] number of groups as,

$$FD'[] = \prod_{A_n[]=G[]} FD \tag{13}$$

Where, $FD'[]$ denotes the separated dataset based on the risk factors. Further, risk levels can be predicted using the same strategy for new data items.

The generic problems are discussed in Section V of this work. Further, the next section of this work discusses the recent improvements in these fundamental methods.

4. Parallel Research Outcomes

Considering the recent research outcomes in this domain, it is evident that the generic methods for data processing are highly customized for various purposes by various research attempts. These attempts are discussed in this section of the work.

4.1. Missing Value Detection and Imputations

The recent improvements over the last few years for deploying machine learning driven methods for identification and improvements on the missing value detection strategies, this work considers analysing the recent improvements. The recent report published [15] on the selection criteria for the population generally considered as the dataset for analysis under training and testing demonstrates that under the population, the cleanliness of the data is one of the primary concerns. Data cleaning, especially missing value reduction, is highly desired as machine learning methods focus on the relations between multiple attributes and the importance of the attributes for model training. Hence, the attribute's importance and relations are sometimes ignored during the missing value analysis, which is a critical mistake. The work by S. I. Koval et al. [16] has demonstrated that considering the attributes for selecting these attributes for missing value analysis can be a highly time complexity-reducing factor. The work showcases a greater reduction in time complexity and reports a reduction in the model complexity.

Apart from the machine learning methods for detection and imputations of the missing values (Figure 4), it can also be realized using the visual methods as showcased in the work by F. Antoniazzi et al. [17]. This work shows a greater possibility of detecting the missing values using visual methods by deploying various tools. Nevertheless, these frameworks are criticised by various other parallel research works for higher dependencies on the system architecture, and they often cannot consider the priority of the tasks and finally land up in a process requiring higher time complexity.

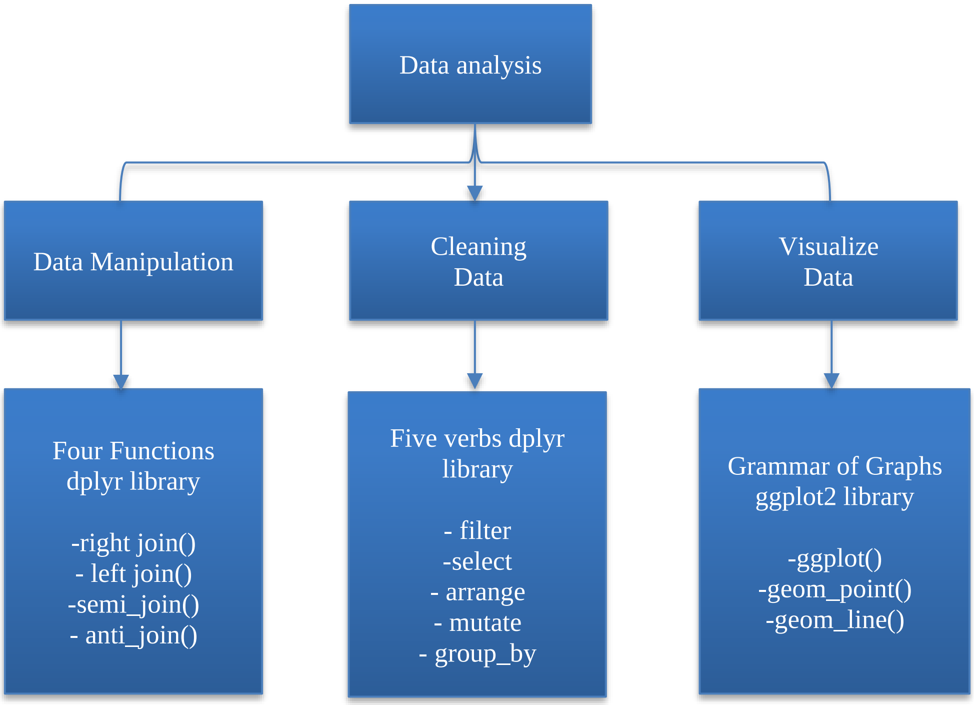


Fig. 4 Missing value analysis

The applicability of machine learning or tool-based methods for identifying and imputations the missing values is inevitable because these missing values can influence the outcomes of the processes to a huge extent in terms of the accuracy of the models. The work by L. Medsker et al. [18] have demonstrated multiple significant situations where the anomalous data reduced the process or model accuracy to a huge extent.

4.2. Outlier Analysis and Imputation

The parallel research attempts have showcased a higher inclination towards cleaning the outliers from the datasets. The outliers, equally critical as missing values, can occur in the dataset for various reasons, such as human or data collection device errors. Due to the higher automation in the data analytics domain, most data are collected automatically without human intervention, reducing the data anomalies. Nonetheless, the data collection process is usually a secondary process, where the actual business process is given priority, and the parallel or secondary process is usually without validations and cannot ensure the correctness of the data. Hence, the outlier treatment of the data before deploying to the models is highly crucial.

The outlier detection (Figure 5) processes are always customised for business situations or highly depend on the data capture process or devices. The work by P. Gil et al. [19] has demonstrated one such example of outlier detection process customization to a greater extent. Nevertheless, the functional process for outlier detection is a statistical process, which is highly adopted. Various versions of this statistical method are discussed in the work by V. Barnett et al. [20]. During a higher dataset analysis, outlier detection can be a highly time-consuming process due to the size of the data. Hence, multiple parallel research attempts have demonstrated methods that consider probability distributions to optimize the detection process. The work by V. L. Brailovsky et al. [21] is significant in this direction.

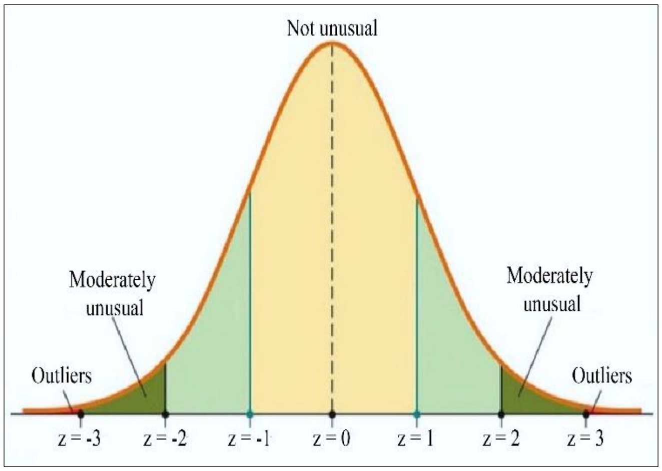


Fig. 5 Outlier analysis

Machine learning-driven methods are also highly adopted by researchers for the detection and imputation of outliers in the datasets. These machine learning methods can identify the trends of the outliers and critically analyse the domain's impact to select the values that are truly out of the acceptable range. The work by H. Kaneko et al. [22] showcased a similar approach for outlier detection using the regression method.

Similar to the advanced machine learning methods for outlier treatment, traditional machine learning methods such as clustering are still very popular. The work by E. M. Knorr et al. [23] has demonstrated a significant and evident improvement over the existing methods using the distance matrix to detect the outliers. Continuing the discussion in the same direction, another work by M. M. Breunig et al. [24] showcased the clusters' density analysis to identify specific detection processes for outliers.

4.3. Risk Prediction

The final objective of this work is to identify the most notable works demonstrated for financial risk predictions (Figure 6) with detailed analysis. The risk prediction or the analysis can be effective after the data pre-processing, as discussed in the previous two sub-sections of this work.

Risk analysis is a widely accepted strategy, and multiple parallel research works have focused on improving these strategies. The work by S. Subudhi et al. [25] have demonstrated the clustering process to be the most effective for risk predictions. Various other parallel research attempts criticise this work for many reasons. Firstly, the clustering method, guided or supervised machine learning, must decide on the clusters or, in this case, the risk labels. The labelling processes for risk analysis can be highly challenging as the risk factors can be subjective according to the domain of the financial institute. Secondly, these labels cannot be pre-decided or decided with many assumptions, which makes the complete model subjected to customisation and cannot be adopted as a generic model.

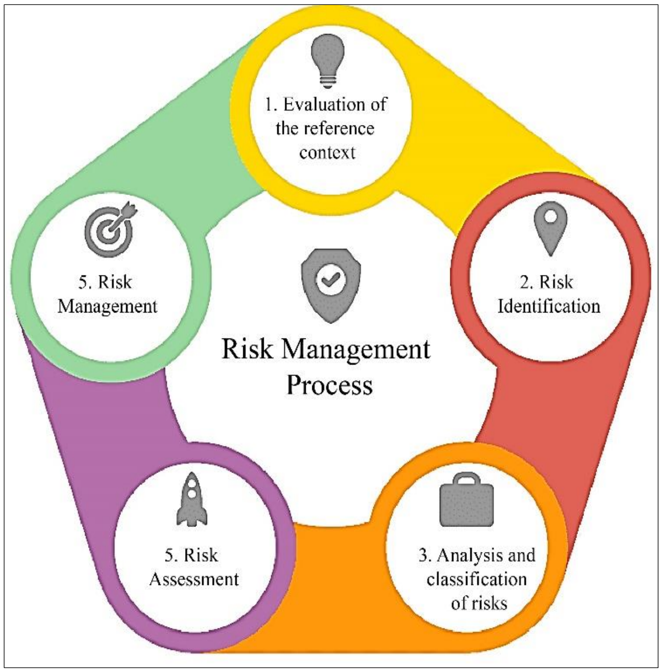


Fig. 6 Risk prediction

The model's adaptability to various domains of financial organizations makes the model highly acceptable among researchers. Working towards this direction, the outcome from the research contribution of M. Cejnek et al. [26] can demonstrate higher adaptability during adaptation.

One of the highly adopted and, at the same time, equally criticised strategy is the automatic regression method. The automatic regression methods can produce the best accuracy during the financial risk predictions and demonstrate the worst time complexity if not designed sensibly. The work by C. Gouriéroux et al. [27] is one such example.

Many researchers have supported the decision to use conditional validation for risk predictions. Conditional validations are one of the most ancient machine learning methods, such as random tree or random forest. R. F. Sproull et al. [28] have showcased an improved strategy based on these methods and assumptions.

The classification and the clustering methods are arguably the most adopted methods for risk analysis, and the two most recent and notable outcomes by T. Zhang et al. [29] and the work by S. V. Stehman et al. [30] have raised more critical reviews than solving this trade-off. In the upcoming sections of this work, a few of these outcomes will be analyzed and validated by testing.

5. Problem Identification

After a detailed understanding of the parallel research outcomes, this section summarises the problems in the existing systems.

The problems are summarized here:

- The risk detection factors are not well defined in terms of pattern detection in data analytics strategies.
- Rule-based analytical policies must be redefined as the financial layouts of enterprise organizations change.
- Data security is a crucial component of this line of research as the financial data is mission critical.
- Also, the volume of the data is growing in all possible directions. Thus, this research direction also demands a reduction in the dimensionality of the data. Hence, newer algorithms for dimensionality reduction must be defined.
- Further, the imputations of the missing values primarily correspond to the mean values of the domains, which is highly incorrect as per the analysis.
- Also, the imputations of the missing values and the outliers do not consider the trend of the data of the analysing domains, which is also an incorrect strategy per this review.

Further, to justify the problems in this research domain, a few parallel research outcomes are analyzed practically with testing on benchmarked datasets in the next section of this work.

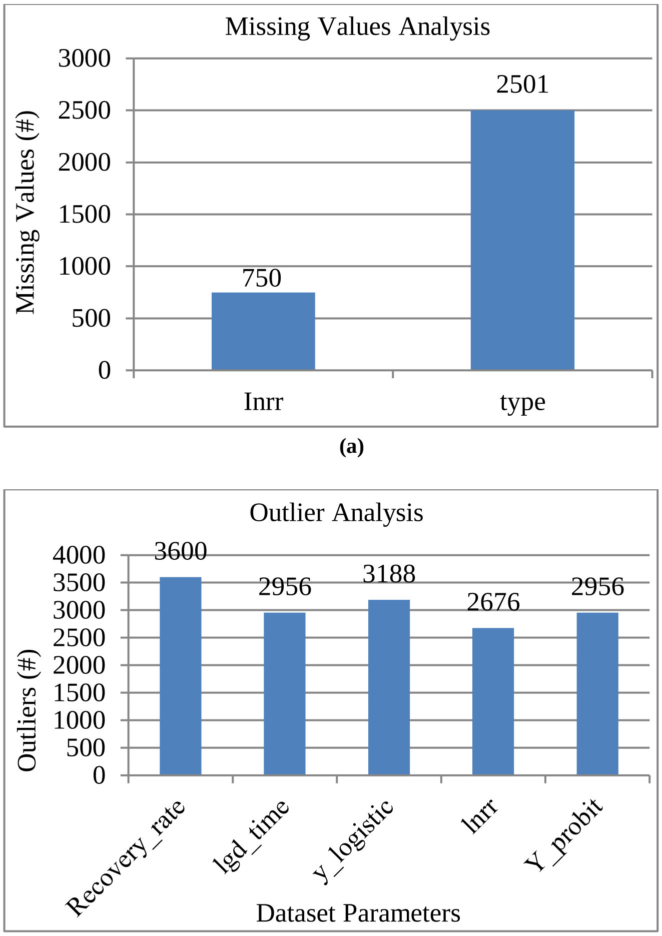
6. Obtained Outcomes and Discussions

Three of the most notable works are further analyzed and tested on the benchmarked dataset presented by Dua, D et al. The description of the dataset is furnished here (Table 1).

Table 1. Dataset description

Dataset Characteristics	Value
Primary Characteristics	Multivariate
Total Instances	690
Total Attributes	14
Domain	Financial
Attribute Types	Real Integer Categorical
Risk Analysis Attribute	class (14th Attribute)
Total Missing Values	750 (lnrr) 2501 (type)
Total Outliers	3600 (Recovery_rate) 2956 (lgd_time) 3188 (y_logistic) 2676 (lnrr) 2956 (Y_probit)

Further, the initial dataset conditions are visualized graphically here (Figure 7).



(a)

(b)

Fig. 7 (a, b) Initial dataset analysis

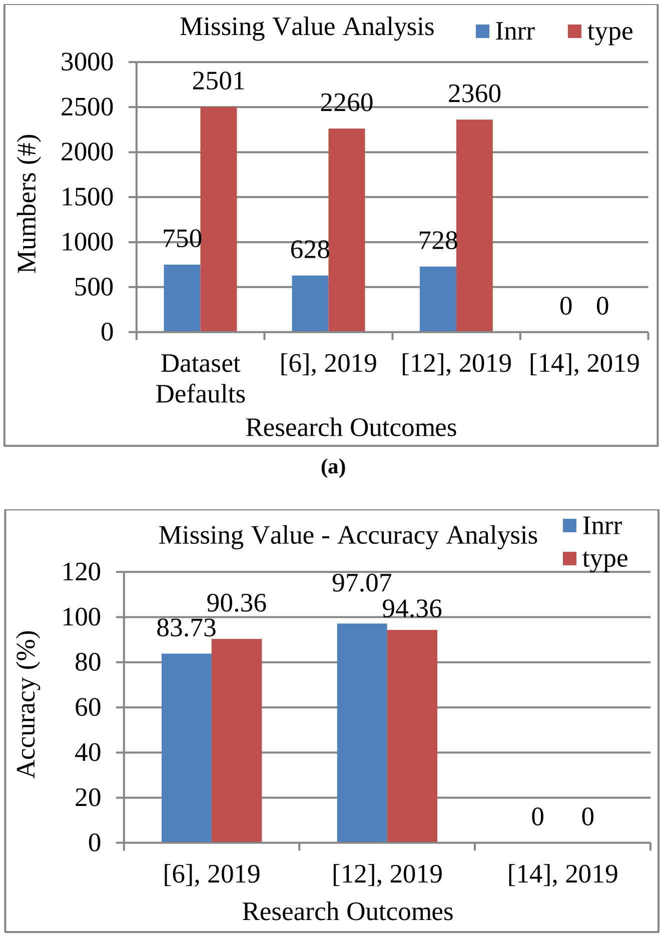
6.1. Missing Value Detection and Imputations

Secondly, the missing value analysis outcomes are furnished here (Table 2).

Table 2. Missing value analysis

Author, Year	Identified Missing Value Count	Actual Missing Value Count	Accuracy (%)
[6], 2019	628 (lnrr) 2260 (type)	750 (lnrr) 2501 (type)	83.73 (lnrr) 90.36 (type)
[12], 2019	728 (lnrr) 2360 (type)	750 (lnrr) 2501 (type)	97.07 (lnrr) 94.36 (type)
[14], 2019	0 (lnrr) 0 (type)	750 (lnrr) 2501 (type)	0 (lnrr) 0 (type)

The results are analysed graphically here (Figure 8). Missing information can happen because of nonresponse: no data is accommodated for at least one thing or an entire unit ("subject"). A few things are bound to create a nonresponse, such as things about private subjects like pay.



(a)

(b)

Fig. 8 (a, b) Missing value analysis

Whittling down is a kind of missingness that can happen in longitudinal examinations—for example, when considering advancement, where estimation is rehashed after a specific timeframe. Missingness happens when members drop out before the test finishes and at least one estimation is missing.

6.2. Outlier Analysis and Imputation

Further, the outlier detection results are analyzed here (Table 3). In information examination, abnormality recognition (additionally outlier location) is the distinguishing proof of uncommon things, occasions or perceptions that raise doubts by contrasting fundamentally from most information. Normally, the atypical things will mean some sort of issue, such as bank extortion, a primary imperfection, clinical issues, or mistakes in a text. Inconsistencies are called outliers, oddities, commotion, deviations and special cases. Specifically, regarding manhandling and organised interruption recognition, the intriguing articles are frequently not uncommon but sudden barges in movement. This example does not cling to the normal measurable meaning of an outlier as an uncommon item, and numerous outlier identification techniques (specifically unaided strategies) will fizzle on such information, except if it has been appropriately amassed. A bunch of investigation calculations might be able to distinguish the miniature groups framed by these examples.

Table 3. Outlier detection analysis

Author, Year	Identified Outlier Count	Actual Outlier Count
[6], 2019	3419 (Recovery_rate) 2856 (lgd_time) 3178 (y_logistic) 2596 (lnrr) 2956 (Y_probit)	3600 (Recovery_rate) 2956 (lgd_time) 3188 (y_logistic) 2676 (lnrr) 2956 (Y_probit)
[12], 2019	3206 (Recovery_rate) 2678 (lgd_time) 2979 (y_logistic) 2434 (lnrr) 2771 (Y_probit)	3600 (Recovery_rate) 2956 (lgd_time) 3188 (y_logistic) 2676 (lnrr) 2956 (Y_probit)
[14], 2019	2564 (Recovery_rate) 2142 (lgd_time) 2383 (y_logistic) 1947 (lnrr) 2217 (Y_probit)	3600 (Recovery_rate) 2956 (lgd_time) 3188 (y_logistic) 2676 (lnrr) 2956 (Y_probit)

The results are analysed graphically here (Figure 9).

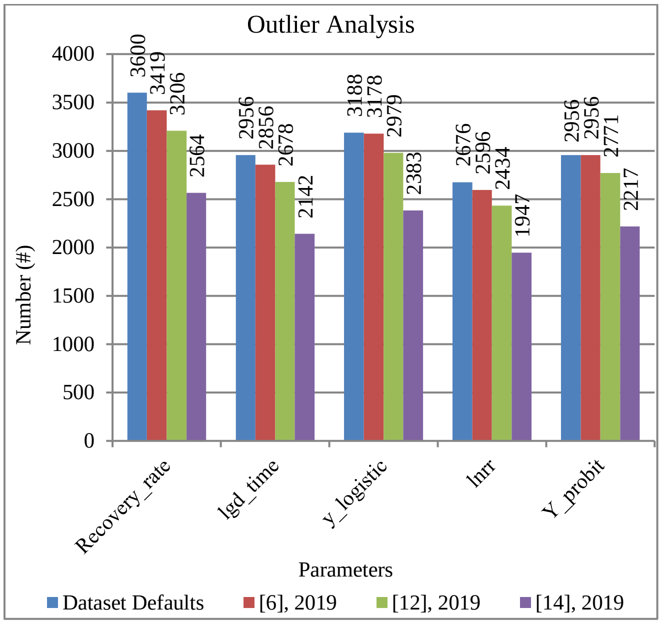


Fig. 9 Outlier detection analysis

6.3. Risk Prediction

Table 4. Risk prediction analysis

Author, Year	Accuracy (%)
[6], 2019	71.3948
[12], 2019	96.0707
[14], 2019	93.2416

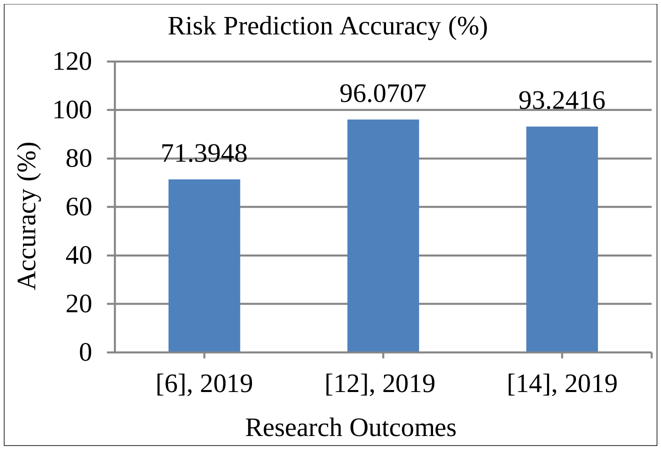


Fig. 10 Risk Prediction Analysis

Finally, the accuracy of the risk prediction is tested for various research works (Table 4). The results are analysed graphically here (Figure 10). Further, the final survey conclusions are presented in the next section of this work.

7. Conclusion

Considering the fissures of the recent research, this work identifies the following major steps to be taken to make a significant contribution towards this research domain. Firstly, A secure framework for managing the queries for risk analysis must be developed considering the highly satisfying time complexity. Secondly, the dimensionality reduction must be done to reduce the data size. Hence reducing the complexity of processing.

Thirdly, the Rule Mining method and proposed Hybrid Apriori algorithm approach will be used to determine the most influential factors for enterprises' financial risk analysis. Finally, a novel framework will be formed with neuro-genetic hybrid algorithms for risk detection and prediction.

References

[1] Natalia Yerashenia, and Alexander Bolotov, “Computational Modelling for Bankruptcy Prediction: Semantic Data Analysis Integrating Graph Database and Financial Ontology,” *2019 IEEE 21st Conference on Business Informatics (CBI)*, Moscow, Russia, vol. 1, pp. 84-93, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] H. Son et al., “Data Analytic Approach for Bankruptcy Prediction,” *Expert Systems with Applications*, vol. 138, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[3] Hafiz A. Alaka et al., “Systematic Review of Bankruptcy Prediction Models: Towards A Framework for Tool Selection,” *Expert Systems with Applications*, vol. 94, pp. 164-184, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[4] Yuji Roh, Geon Heo, and Steven Euijong Whang, “A Survey on Data Collection for Machine Learning: A Big Data-Ai Integration Perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 4, pp.1328-1347, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Tao Huang et al., “Promises and Challenges of Big Data Computing in Health Sciences,” *Big Data Research*, vol. 2, no. 1, pp. 2-11, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[6] TA Lin, “Systematic Risk Measurement of Internet Finance[J],” *Statistics and Decision*, vol. 35, no. 07, pp. 158-161, 2019.

[7] MA Li, and FAN Wei, “Where did the Liquidity Released by the Central Bank Go?–An Empirical Test Based on Micro-level Data[J],” *Modem Economic Science*, vol. 41, no. 03, pp. 39-48, 2019.

[8] Li Yuanyuan, Cui Chenchen, and Liu Siyu, “Financial Ecological Environment Enterprise Risk Commitment and Innovation Efficiency-An Empirical Analysis of Manufacturing Based on Panel VAR,” *Industrial Technology and Economy*, vol. 38, no. 7, pp. 76-87, 2019. [Google Scholar] [Publisher Link]

[9] Yang Songling et al., “Financialization of Entity Enterprises Analyst Coverage and Internal Innovation Driving Force[J],” *Joumal of Management Science*, vol. 32, no. 2, pp. 3-18, 2019. [Google Scholar]

[10] Li Hua, Zhao Shuying, and Sun Qiubai, “Construction and Analysis of Financial Security Indicators Evaluation System Based on the Weighted Principal Component Distance Clustering[J],” Mathematics in Practice and Theory, vol. 48, no. 01, pp. 90-102, 2018. [Google Scholar]

[11] BAI Xue and NIU Feng, “The Measurement Test and Regulation of the Systemic Risk Contribution in Financial Institutions[J],” *Journal of Shanxi Finance and Economics University*, vol. 40, no. 12, pp. 45-59, 2018.

[12] YE Li and WANG Yuanzhe, “CHEN Yongyong Study on the Risk Spillover Effect between Chinese Financial Institutions[J],” *Statistics & Information Forum*, vol. 34, no. 03, pp. 54-63, 2019.

[13] Zhu Xiao-qian et al., “An Indicator of Conditional Probability of Crisis for Systemic Risk Measurement[J],” *Chinese Journal of Management Science*, vol. 26, no. 6, pp. 1-7, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] Liu Zhiyang, “Systemic Risk Contributions?–Evidence from Panel Variable Coefficient Model[J],” *Modem Economic Science*, vol. 41, no. 03, pp. 49-60, 2019.

[15] Documentation, Gapminder, 2018. [Online]. Available: http://www.gapminder.org/downloads/documentation/gd003.

[16] Stanislav I. Koval, “Data Preparation for Neural Network Data Analysis,” *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, Moscow and St. Petersburg, Russia, pp. 898-901, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] Francesco Antoniazzi, and Fabio Viola, “RDF Graph Visualization Tools: A Survey,” *2018 23rd Conference of Open Innovations Association (FRUCT)*, Bologna, Italy, pp. 25-36, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[18] L. Medsker, “Design and Development of Hybrid Neural Network and Expert Systems,” *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*, Orlando, FL, USA, vol. 3, pp. 1470-1474, 1994. [CrossRef] [Google Scholar] [Publisher Link]

[19] Paulo Gil, Hugo Martins, and Fábio Januário, “Outliers Detection Methods in Wireless Sensor Networks,” *Artificial Intelligence Review*, vol. 52, pp. 2411-2436, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[20] Vic Barnett, and Toby Lewis, *Outliers in Statistical Data*, Wiley, New York, USA, 3rd ed., 1994. [Google Scholar] [Publisher Link]

[21] V.L. Brailovsky, “An Approach to Outlier Detection Based on Bayesian Probabilistic Model,” *Proceedings of 13th International Conference on Pattern Recognition*, Vienna, Austria, vol. 2, pp. 70-74, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[22] Hiromasa Kaneko, “Automatic Outlier Sample Detection Based on Regression Analysis and Repeated Ensemble Learning,” *Chemometrics and Intelligent Laboratory Systems*, vol. 177, pp. 74-82, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[23] Edwin M. Knox, and Raymond T. Ng, “Algorithms for Mining Distance-Based Outliers in Large Datasets,” *Proceedings of the 24th VLDB Conference*, New York, USA, pp. 392-403, 1998. [Google Scholar] [Publisher Link]

[24] Markus M. Breunig et al., “LOF: Identifying Density-Based Local Outliers,” *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, New York, United States, pp. 93-104, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[25] Sharmila Subudhi, and Suvasini Panigrahi, “Use of Optimized Fuzzy C-Means Clustering and Supervised Classifiers for Automobile Insurance Fraud Detection,” *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 5, pp. 568-575, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[26] Matous Cejnek, and Ivo Bukovsky, “Concept Drift Robust Adaptive Novelty Detection for Data Streams,” *Neurocomputing*, vol. 309, pp. 46-53, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[27] Christian Gouriéroux, *ARCH Models and Financial Applications*, New York, USA, Springer, 1st ed., 1997. [CrossRef] [Google Scholar] [Publisher Link]

[28] Robert F. Sproull, “Refinements to Nearest-Neighbor Searching in K -Dimensional Trees,” *Algorithmica*, vol. 6, pp. 579-589, 1991. [CrossRef] [Google Scholar] [Publisher Link]

[29] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, “BIRCH: An Efficient Data Clustering Method for Very Large Databases,” *ACM SIGMOD Record*, vol. 25, no. 2, pp. 103-114, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[30] S.V. Stehman, “Selecting and Interpreting Measures of Thematic Classification Accuracy,” *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77-89, 1997. [CrossRef] [Google Scholar] [Publisher Link]