# Using Deep Learning to Interpolate the Missing Data in Time-series for Credit Risks along Supply Chain

Wenfeng Zhang[a,f]; Ming K. Lim[e]; Mei Yang[b]; Xingzhi Li[d]; Du Ni[c*]
[a] College of Mathematics and Statistics, Chongqing University, Chongqing 400044, PR China
[b] School of Economics and Business Administration, Chongqing University, Chongqing, 400044, PR China
[c] School of Management, Nanjing University of Posts and Telecommunications, Jiangsu, 210003, PR China
[d] School of economics and management, Chongqing Jiaotong University, Chongqing, 400074, PR China
[e] Adam Smith Business School, University of Glasgow, Glasgow G14 8QQ, the United Kingdom
[f] College of General Education, Chongqing Polytechnic Institute, Chongqing 401228, PR China

**Corresponding author**

E-mail addresses: mt15nn@mail.wbs.ac.uk (Du Ni)

**Abstract**

**Purpose:** As the supply chain is a highly integrated infrastructure in modern business, the risks in supply chain are also becoming highly contagious among the target company. This motivates researchers to continuously add new features to the datasets for the credit risk prediction (CRP). However, adding new features can easily lead to the missing of the data.

**Design:** Based on the gaps summarized from the literature in CRP, this study first introduces the approaches to the building of datasets and the framing of the algorithmic models. Then, this study tests the interpolation effects of the algorithmic model in three artificial datasets with different missing rates and compares its predictability before and after the interpolation in a real dataset with the missing data in irregular time-series.

**Findings:** The algorithmic model of the time-decayed long short-term memory (TD-LSTM) proposed in this study can monitor the missing data in irregular time-series by capturing more and better time-series information, and interpolating the missing data efficiently. Moreover, the algorithmic model of Deep Neutral Network can be used in the CRP for the datasets with the missing data in irregular time-series after the interpolation by the TD-LSTM.

**Value:** This study fully validates the TD-LSTM interpolation effects and demonstrates that the predictability of the dataset after interpolation is improved. Accurate and timely CRP can undoubtedly assist a target company in avoiding losses. Identifying credit risks and taking preventive measures ahead of time, especially in the case of public emergencies, can help the company minimize losses.

**Keywords**

deep learning; credit risk prediction; interpolation; missing data in irregular time-series; supply chain;

**Declaration**

This work was supported by the 2022 Scientific Research Startup Fund of Chongqing Jiaotong University [Grant No. F1210045]. The authors declare no competing interests.

# 1. Introduction

Credit risk prediction (CRP) refers to the process for predicting whether a company will default in the future through the historical data of a company (Bu *et al.*, 2018; Kousenidis *et al.*, 2019; Ma *et al.*, 2019). Accurate and timely CRP can help companies avoid losses. Early indicators for CRP are obtained mainly from financial reports of the target companies (Tian *et al.*, 2020; Trustorff *et al.*, 2011; Wu *et al.*, 2015), but after the global subprime mortgage crisis in 2008, it has been questioned whether the methods that rely solely on financial data from one company for CRP (Trustorff *et al.*, 2011) are timely (Savitri *et al.*, 2019) and reliable (Ashraf *et al.*, 2020; Holder-Webb *et al.*, 2010). The researchers turn to the information mined from the data indirectly related to the company (Bakoben *et al.*, 2020; Zhang *et al.*, 2022). The data concerning supply chain indirectly play an important role as those direct ones in CRP, for the integrity of the supply chain and the corporate credit risk can easily spread along the supply chain, leading to a chain of credit disasters (Agca *et al.*, 2022; Osadchiy *et al.*, 2016).

In essence, most of the data used in CRP are in time-series with historical information (Chang *et al.*, 2018; Tian *et al.*, 2020), and these data can be collected from different sources. However, some of the CRP indicators are collected in a fixed sampling frequency, while the others are not (Kreindler *et al.*, 2016; Mikalsen *et al.*, 2018), and as a result, it is almost impossible to fit all the indicators for each other when the data of different sources are not collected in a fixed sampling frequency for the same company. Worse still, the missing data will cause a break in the continuity of the data in time-series, which will seriously affect subsequent data analysis and processing of CRP (Tian *et al.*, 2018). Interpolation (Sun *et al.*, 2021; Wubetie, 2017) is the most common way of dealing with the missing data, and recently to interpolate the missing data, the algorithmic models of machine learning such as Deep Neural Networks (DNN), Generative Adversarial Networks (GAN), and Long Short-Term Memory (LSTM) have been employed in time-series (Ma *et al.*, 2020; Yoon *et al.*, 2019), but for irregular time-series, few of the algorithms take the missing data into consideration (Pratama *et al.*, 2016; Wang *et al.*, 2020). As a result, adding the irregular time-series as an independent time-function during interpolation requires more consideration.

To raise the interpolation effects, this study will first summarize the relevant literature of CRP and the interpolation for the missing data in time-series, and then illustrate the approaches to the building of datasets and the framing of the algorithmic model that can interpolate the missing data in irregular time-series with high predictability. Finally, this study will test the interpolation effects of the algorithmic model in artificial datasets with different missing rates and compare its predictability before and after the interpolation in a real dataset with the missing data in irregular time-series.

# 2. Literature Review

This section will review the literature closely related to this study from two aspects: the previous studies in CRP and the methods for interpolating the missing data in time-series.

## 2.1 Credit Risk Prediction

Credit risks speak for the ability of the solvency and possible default of a company, but it has always been difficult to predict the credit risks (Chi *et al.*, 2017; Scannella *et al.*, 2021) with proper timeliness (Savitri *et al.*, 2019) and reliability (Ashraf *et al.*, 2020; Holder-Webb & Sharma, 2010). In this case, many scholars make great efforts in developing powerful algorithms for better CRP. In general, working out CRP mainly involves two types (Chi & Li, 2017) of algorithms: linear discriminant analysis based on statistics (Mahmoudi *et al.*, 2015) and the algorithms of machine learning based on computer programs (Ma & Lv, 2019; Yu *et al.*, 2022). Unlike statistical methods that require researchers to manually estimate the parameters for the CRP, the algorithms of machine learning allow computers to parse the data and grasp useful knowledge from the data by the computers themselves (Bhatore *et al.*, 2020; Pandey *et al.*, 2017). Of the machine learning algorithmic models, the most commonly used include support vector machines (SVM), neural networks (NN), and random forests (RF) (Arora *et al.*, 2020; Teles *et al.*, 2020; Trustorff *et al.*, 2011), and in addition, various variants of deep learning like deep belief networks (DBN) and DNN in CRP, have gained widespread popularity among researchers for their remarkable ability to learn and make informed decisions on its own (LeCun *et al.*, 2015). For example, Yu *et al.* (2018) use the ensemble strategy of the DBN for credit risk classification, finding that the performance of the ensemble DBN model for the classification has been effectively improved. Xie *et al.* (2021) propose a DNN model for CRP to help the banks in China to improve the efficiency of farmers' credit business, claiming that the predictability of the DNN model is better than that of other competing algorithms.

In addition to the use of algorithmic models of machine learning, the use of additional predicting indicators is more seen in CRP (Trustorff *et al.*, 2011). Frequent reports of financial frauds in businesses (Haqq *et al.*, 2019; Reurink, 2018) urge researchers to question the reliability of the core financial data directly obtained from the target company, such as cash liquidity (Wu & Brynjolfsson, 2015), loan yield (Stein, 2005), cost of debt (Mansi *et al.*, 2012). Thus, the data indirectly related to the target company, the data along supply chain in particular, are also traced, mined and employed in CRP (Ashraf *et al.*, 2020; Holder-Webb & Sharma, 2010) to enhance the timeliness and reliability of CRP. Recently, a large body of studies has shown that the data along the supply chain can help protect the upstream and downstream companies in case there would be risk spillovers along the supply chain, thereby adversely affecting the target company (Wu *et al.*, 2022). For instance, Osadchiy *et al.* (2016) assert that credit risk can propagate in supply chain networks and exacerbate supplier production losses; supply chain data are applied to monitoring sudden inventory, sales fluctuations, and thus optimizing the structure of the supply chain (Ni *et al.*, 2022). and also by monitoring data of the supply chain, Lee *et al.* (2021) succeed in timely predicting the credit risk, all of which indicate that the data of the supply chain can serve as a complementary source for CRP.

## 2.2 Interpolation for the missing data in time-series

In CRP, most of the data used are in time-series and often with some data missing (Chang *et al.*, 2018; Tian *et al.*, 2020). These data share rich historical information that is crucial to CRP. Moreover, in the process of data collection and collation, researchers continue to add new indicators, such as those related the supply chain to mine more useful information for high predictability in CRP (Bakoben *et al.*, 2020; Zhang *et al.*, 2022). Usually, these data have to be obtained from different sources. The indicators of the data thus collected may lack certain attributes or specific content, and this shortage may generate a serious information loss--some important information missing. In this case, it is unlikely to effectively learn and build traditional algorithmic models of machine learning, and the output results the models yield are not reliable (Kreindler & Lumsden, 2016; Mikalsen *et al.*, 2018; Tian *et al.*, 2018). Accordingly, it is urgently needed for successful CRP along supply chain to reasonably prepare the missing data, that is, the interpolation for the missing data in time-series

In the past twenty years, a line of research has explored how to deal with the missing data in CRP (Chang *et al.*, 2020; Fouladgar *et al.*, 2020; Guo *et al.*, 2009). Some choose to simply delete the entire row or column containing the missing data out of a dataset. This may lead to the direct loss of valuable information (Ngueilbaye *et al.*, 2021; Stanimirova *et al.*, 2007), and to prevent the loss, some scholars resort to interpolating the missing data. In interpolation, maximum likelihood estimation (Mählmann, 2006), Laplace approximation (Collin-Dufresne *et al.*, 2004), and probabilistic matrix factorization (Fekade *et al.*, 2017) have been traditionally applied. For example, Qu *et al.* (2009) use the probabilistic principal component analysis to predict and interpolate the missing data of the traffic flow datasets; Fekade *et al.* (2017) adopt the probabilistic matrix factorization to interpolate the missing data based on the sensor datasets.

More recently, algorithmic models of machine learning, such as ensemble algorithms or support vector regression algorithms (Appelhans *et al.*, 2015; Li *et al.*, 2011; Wang *et al.*, 2019) are making their way into interpolating the missing data in CRP from the perspective of big data. For example, an ensemble algorithm developed by Chang *et al.* (2020) in their experiment to interpolate the missing data proves its high robustness in both interpolation and prediction. The algorithms of deep learning are also introduced into interpolation. Kaur *et al.* (2019) propose an algorithmic model for interpolation by using GAN, and good results are achieved; Yoon *et al.* (2019) use the algorithmic model of RNN to interpolate within and across data streams, discovering that compared to statistical interpolation methods, RNN offers significant improvement in interpolating the missing data. Ding *et al.* (2020) conduct a multi-dimensional filling experiment on environmental data based on the LSTM model, and they claim that their model makes up for the defect that the interpolation along a single dimension would increase the cumulative errors as the sequence increases, and thus maintains the accuracy in CRP.

As for the missing data in time-series for CRP, scholars have provided different time-functions

based on their own research to detect the locations and the time intervals of the missing data. Zhang *et al.* (2019) propose the time-function of the variable-length sliding window; Baytas *et al.* (2017) work out the time-function of the temporal awareness based on the LSTM; and Tan *et al.* (2020) come up with the time-function of the dual-attention temporal awareness, to name a few. Despite the variation, the time-functions interpolate the missing data merely in regular time-series, except for the one yielded by Baytas *et al.* (2017), who are the pioneers in interpolating the missing data in irregular time-series. Admittedly, with the help of the time-functions, the above-mentioned algorithms have presented effective information, such as the locations and the time intervals of the missing data in the process of interpolation, but these algorithms still have two gaps in interpolating the missing data for CRP. One is that all the time-functions have been put in front of the algorithmic models in which case the real state before and after the interpolation of the missing data cannot be effectively checked by the algorithmic models. The other is the insufficiency of flexibility during the global application of the time-functions in the algorithmic models.

Therefore, to fill the gaps, this study intends to first present a time-function that can check the real state of the missing data in irregular time-series both before and after interpolation, and then to propose an algorithmic model that can apply the time-function globally with high flexibility.

# 3. Material and methods

This section will first introduce the data sources and their preprocessing, and then the approaches to the framing of the algorithmic model for interpolating the missing data in irregular time-series.

## 3.1 Sources and their preprocessing

In this study, 441 companies were selected on the S&P 500 companies list because they had credit rating announcements from any of the three major rating agencies (Moody's, S & P, and Fitch) from January 1, 2009 to December 31, 2019. Their CRP data were of four categories: (1) Corporate Financial Data, collected from the S&P Accounting Database (https://www.spglobal.com/ratings/en/). The indicators involved the weekly, quarterly and annual financial data released, mainly including weekly working capital ratio, interest compensation ratio, retained earnings ratio, return on assets, tax leverage ratio, cash inverse ratio, debt to capital ratio, and the others. (2) Network Activity Data, involving search trends, website visits, and other data of network activities. These data were believed to have strong timeliness and could be used as a supplementary source to corporate financial data of weak timeliness (Moat *et al.*, 2016). (3) Supply Chain Data. This group of data was obtained from the Bloomberg Database (https://www.bloombergchina.com/solution/data-content/), which contained more than 20,000 pieces of quantitative data in supply chain. (4) Credit Risk Data. This study adopted the bond price to represent the corporate credit risk since the fluctuation of bond price was a main reflection of credit risk (Gilchrist *et al.*, 2018; Zamore *et al.*, 2018), and the Standard & Poor's Accounting Database

(https://www.spglobal.com/ratings/en/) was the major source of the bond prices.

Since the data were collected from different sources and dated from January 1, 2009 to December 31, 2019. The mismatch of indicators in time-series would surely lead to data missing, although the loss varies in degree. The indicators' mismatch in the dataset of this study was sorted out and graded into four types (1) No Missing: such indicators as Google trend, rating rank, and upgrade in time window of the target company, the rate of missing data is as low as 0.0%; (2) Low Missing: most of the indicators typical of low missing belong to the core financial data of the target company, such as return on assets, capital turnover, and debt-equity ratio; (3) Moderate Missing: the indicators in this middle-level group are mostly related to suppliers along the supply chain; (4) High Missing: the core financial indicators concerning customers, such as customers' working capital ratio, customers' debt to equity ratio, and customers' capital turnover ratio, share a high rate of absence (as high as 50% on average).

To further illustrate the missing indicators in the dataset of this study, Python 3.0 was employed to visualize the dataset in terms of each single indicator, and some typical cases are presented in Figure 1.

As shown in the figure, there is a specific name list of the indicators along the vertical axis and the actual number of the missing data illustrated by those black horizontal bars. And out of the cases of missing indicators and their actual rates of data missing, we see indicators with similar missing rate predominantly come from the same sources. For instance, the indicators of high missing rate, whose items are located in the upper part of Figure 1, are mainly related to the customers of the target company; those of low missing rate, which are listed in the lower part of Figure 1, belong to the core financial indicators.

The indicators of similar missing rate mainly belong to the same sources, and considering this homogeneity, the indicators with a missing rate greater than 60% are recommended to be deleted. Therefore, the database of this study obtained 125 indicators with 167,160 pieces from the 441 target companies, and 25 indicators are of corporate financial data, 36 network activity data, 58 supply chain data and 6 credit risk data. The missing rate of all the indicators is averaged around 31%.

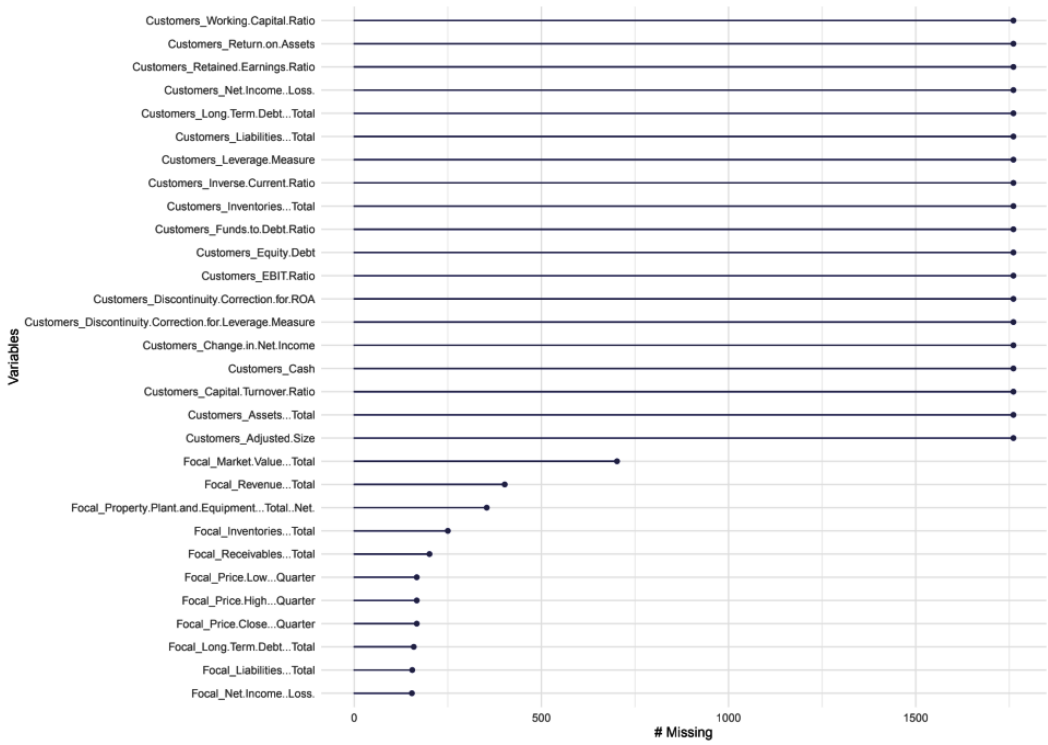


Figure 1. Cases of typical missing indicators

To show the distribution pattern of the missing data in a schematic diagram, we took the same indicator (working capital ratio) of two companies (AA: American airline; ABT: Abbott) over 3-5 years as examples in Figure 2. The upper and lower parts of Figure 2 show the distribution pattern of working capital ratio for AA and ABT (both including the working capital ratio for the target companies and their suppliers & customers along the supply chain), respectively. The two horizontal axes indicate the time-series (the interval of data acquisition is marked as $h_t$), while the figure vertically lists the indicators of the target company, its supplier and customer along the supply chain.

From the distribution pattern of the missing data in Figure 2, it is suggested that the occurrence of the missing data along the time-series for both companies in the past three to five years is irregular. For AA and ABT, the intervals between the two points of the data acquisition ($h_t$) within the past three to five years are quite different for the target companies, their suppliers and customers along the supply chain. Taking AA as an example, although the data were fully acquired at the same time on May 29, 2013 for the target company, its supplier and customer along the supply chain, the missing data (indicated as the hollow boxes in Figure 2) occurred frequently on other days of data acquisition. Therefore, the intervals $d_t$ between two solid boxes for the same indicators are quite different for AA (40 months), its supplier (38 months) and customer (33 months) along the supply chain.

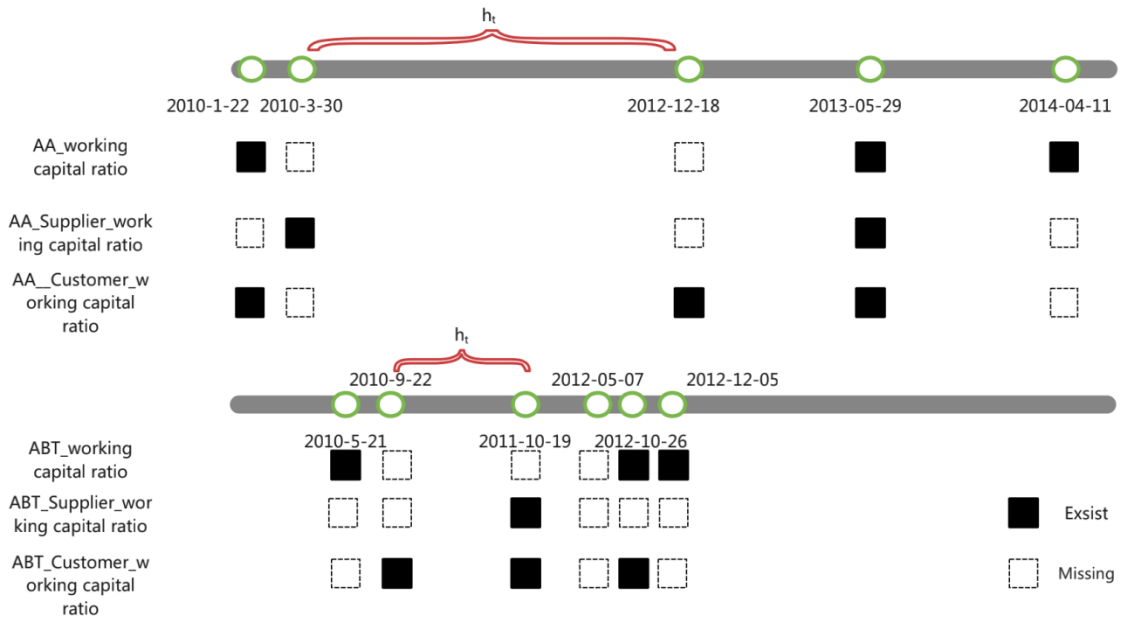


Figure 2 The schematic diagram of missing data in time-series

Therefore, the data collection interval of the datasets in this study is different within the same sample company and across different sample companies, and the distribution pattern of the missing data is also different. The varied distribution of locations and the time intervals of the missing data determined that the original dataset collected in this study was characterized by the missing data in irregular time-series.

### 3.2 Building of the algorithmic models

As indicated in Section 2, various algorithmic models of machine learning such as DNN, GAN, and LSTM have been used in interpolating the missing data for CRP. For the missing data in time-series in particular, several time-functions such as the variable-length sliding window (Zhang *et al.* 2019), the time-function of the temporal awareness (Baytas *et al.* 2017), the dual-attention temporal awareness (Tan *et al.* 2020), have been employed in CRP. It is noteworthy that most of these time-functions are built on the basic model of LSTM in interpolating the missing data in time-series. Therefore, we will build an algorithmic model on LSTM too, but to interpolate the missing data in irregular time-series. And Baytas *et al.* (2017) provide a good reason for doing so, claiming that the LSTM has the potential for flexibility by adding gates and stacking the models for checking the state of the missing data during the interpolation process.

LSTM was initially used in the field of big data based on deep learning by Hochreiter *et al.* (1997) at the end of the last century, and since then it has experienced many improvements, of which the vanilla LSTM proposed by Graves *et al.* (2005) is most widely employed as a model in deep-learning algorithms tradition (Ma *et al.*, 2020; Yoon *et al.*, 2019). Therefore, this study intended to select the

vanilla LSTM as the basic model for interpolating the missing data in irregular time-series, as shown in Figure 3.

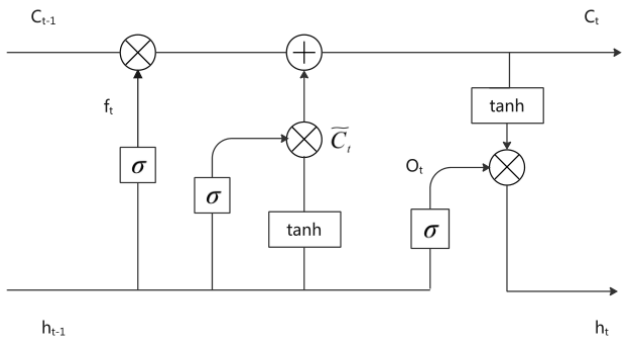


Figure 3 Illustration of traditional LSTM

As shown in Figure 3, the traditional LSTM as the basic model in this study generally contains three inputs and two outputs. The three inputs present the cell state $C_{t-1}$ and hidden state $h_{t-1}$ at time *t-1*, and input $X_t$ at time *t*; the two outputs show the cell state $C_t$ and hidden state $h_t$ at time *t*. Through three different gates (input gate, forgetting gate, and output gate), the memory module is able to control the hidden state of the memory and the time that occurred previously. However, as a basic model, although the traditional LSTM has advantages in detecting the state of the missing data in regular time-series (Ma *et al.*, 2020), it is difficult to deal with the missing data in irregular time-series obtained in this study. Thus, this study intends to improve the traditional LSTM by: (1) adding time-functions representing the locations and the time intervals of the missing data; (2) adding a set of gating units to check the state of the missing data in irregular time-series before and after the interpolation. The improved LSTM based on the interpolation of the missing data in irregular time-series is shown in Figure 4.

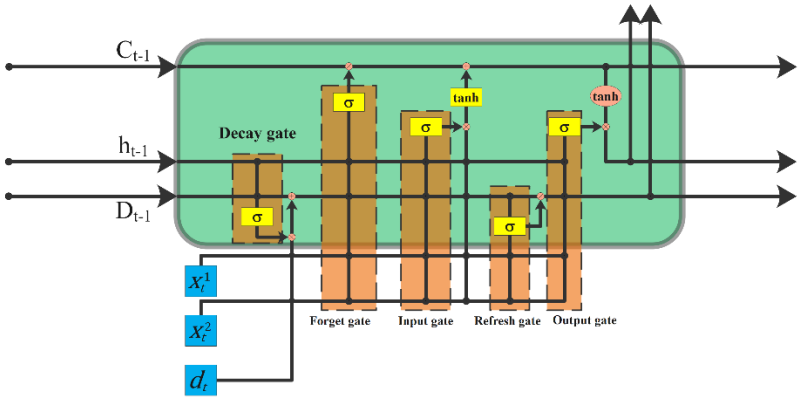


Figure 4 The improved LSTM for interpolation

More specifically, to increase the ability to process the missing data in irregular time-series, this

study needs to first build a time-function indicating the locations and intervals of the missing data. The process for building the function is as follows:

Suppose that a set of data contains $n$ variables and the time interval is $T$, i.e., $X = \{X_1, X_2, X_3, \dots, X_T\} \in \mathcal{R}^{\mathrm{T}\times\mathrm{N}}$, where $X_{\mathrm{t}}$ $(t = 1, 2, 3, \dots, T)$ is the observed value of all variables at time $t$, i.e., $X_{\mathrm{t}} = \{x_t^1, x_t^2, x_t^3, \dots, x_t^N\}^{\mathrm{T}} \in \mathcal{R}^{\mathrm{N}}$; $x_{\mathrm{t}}^{\mathrm{n}}$ is the observed value of the *n-th* variable at time $t$. The timestamp is denoted as $s_t$, and the missing data in $X$ is marked as dummy variable $m$ (Bankó *et al.*, 2012), namely

$$m_t^n = \begin{cases} 0, & if\ x_t^n\ is\ missing \\ 1, & otherwise \end{cases} \tag{1}$$

Define a multi-dimensional variable $d_t = \{d_t^1, d_t^2, \dots, d_t^N\}$ as the time interval of the missing data in the dataset $X$, i.e.,

$$d_t^n = \begin{cases} s_t - s_{t-1} + d_{t-1}^n, & t > 1, m_{t-1}^n = 0 \\ s_t - s_{t-1}, & t > 1, m_{t-1}^n = 1 \\ 0, & t = 1 \end{cases} \tag{2}$$

After the above-mentioned processing for the missing data in irregular time-series, {X, M, Δ} can be obtained, of which:

$$M = [m_t^n]^{n\times T}, t = 1,2,\dots,T; n = 1,2,\dots,N \tag{3}$$

$$\Delta = [d_t^n]^{n\times T}, t = 1,2,\dots,T; n = 1,2,\dots,N \tag{4}$$

Taking the whole process of the interpolation of the missing data in irregular time-series into consideration, we are able to find that the traditional LSTM, when the missing data is in irregular time-series, cannot cope with the situation where the output decreases with the increase of time intervals. Therefore, to the traditional LSTM, this study introduced $h_t^n$ that had an attenuation mechanism by capturing the relationship among input variables, hidden variables, and the corresponding time intervals for the missing data, and assigning the attenuation coefficient with the flexibility specific to a given variable.

It can also be found in the middle unit of Figure 4 that two more gates (the time-decayed gate and time-refreshing gate) were added to the improved LSTM. Therefore, the new LSTM constructed in this study consists of five gates: time-decayed gate, forgetting gate, input gate, refreshing gate, and output gate. As mentioned above, the role of the time-decayed gate is to introduce an attenuation function with time by capturing the relationship among input variables, hidden variables, and the corresponding time intervals for the missing data, and the values of the attenuation function will fluctuate with the specific variables.

As indicated in Figure 4 for the time-decayed gate, Sigmoid $[\mathrm{sigmoid}(W_1 * d_t + W_2 * d_t^2)]$ is used to represent the state of attenuation with time. The attenuation will change with the combination of $W_1$ and $W_2$ in different weights, and with the squared time variable $d_t$. In this way, the sigmoid

function will take more forms and increase the flexibility of time-decayed variables. The calculation formula of time-decayed gate is as follows,

$$D_t' = W_{dt} d_t \odot \sigma(W_{th} h_{t-1} + W_{tt} D_{t-1} + b_t) \tag{5}$$

Where $\odot$ represents the pairwise matrix multiplication; $\sigma$ is the sigmoid function; $\sigma(x) = \frac{1}{1+e^{-x}}$; $W$ and $b$ variants represent the weight matrix and offset, respectively. It can also be seen from Figure 4 that time-decayed gate can be affected by the time interval $d_t$ between the points of two non-missing data. The role of the refreshing gate is to reset the data after the data has been updated, and is mainly controlled by the time-decayed gate. Its calculation formula is as follows,

$$D_t = D_t' \odot \sigma(W_{ru} x_t^1 + W_{ri} x_t^2 + W_{rd} D_t' + b_r) \tag{6}$$

On the basis of the whole process of the interpolation of the missing data in irregular time-series, it can be assumed that the functions representing the locations and time intervals of the missing data can be well established, and the gates added to the model can better detect and handle the state of the missing data in irregular time-series before and after interpolation. However, in order to ensure the generalization of the interpolation in a more stable way, we also add the SENet module (Hu *et al.*, 2018) where stacked autoencoders (SAE) are recommended to enhance the interpolation accuracy. The structure of the stacked model is shown in Figure 5.

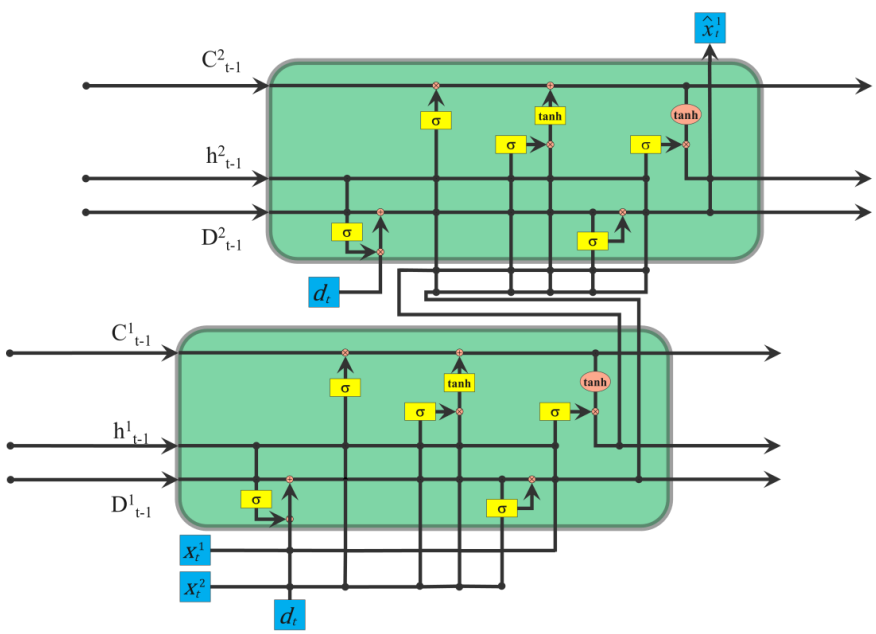


Figure 5 The stacked LSTM for interpolation

As can be seen from the intermediate layer of the two stacked units in Figure 5, the SENet module is added to all the gates of the two stacked units. Its role is to automatically capture the weights in importance of each target channel through learning without altering the structure of the original

algorithms, and to filter the features that are not valuable to the current tasks and boast the valuable features according to the weights in importance of each target channel. In addition, the SAE is used to mine the features more deeply. As claimed by Hu *et al.* (2018), the SAE, as a classical deep learning network structure, can be stacked to form a multi-layer network. Each layer within the network will take the features extracted from the previous layer of the network as an input to further extract more abstract and deeper features for the present layer of the network. Thus, the accuracy of the algorithms is enhanced by mapping the input signs to the hidden layer expressions through the feature extraction of the data.

To sum up, the stacked LSTM for interpolation proposed in this study has the following characteristics: (1) To improve the checking of the states of missing data in irregular time-series before and after interpolation. That is, two gates, the time-decayed gate and the refreshing gate, are added to the traditional LSTM. (2) A time-decayed function is introduced to capture the relationship between input variables, hidden variables, and missing data time intervals. This increases the flexibility of time decayed expressions by allowing them to take more forms in the LSTM. (3) The SENet and SAE modules are added to the LSTM to improve the generalization and accuracy of the algorithms. As a result, the time-decayed LSTM, abbreviated TD-LSTM, is the name given to the stacked LSTM for interpolation presented in this study.

# 4. Results and discussion

This section will show the interpolating effects of the TD-LSTM in artificial datasets with different data missing rates and compare the predictability of DNN with those of the other algorithmic models in a real dataset with the missing data in irregular time-series before and after interpolation by the TD-LSTM.

## 4.1 Interpolating effects in artificial datasets

To examine the interpolating effects of the TD-LSTM proposed in this study, three artificial datasets, out of the database obtained in Section 3, were established for testing and their missing rates were deliberately set at 31%, 40% and 50%, respectively. The lowest missing rate settled at 31% because the missing rate of the database as discussed in Section 3 was averaged at 31%. The highest missing rate was 50% for the reason that the variables whose missing rates were over 60% had been deleted because of their high homogeneity, as is seen in Section 3. For comparison, the RNN, LSTM, and Time-aware LSTM were chosen as the counter-partners of the TD-LSTM. We chose these three models for two reasons. The first is that these three algorithmic models are functionally similar to the TD-LSTM in interpolating missing data; the second is that RNN is one of the most frequently used algorithmic models in interpolating missing data, LSTM is the basic model for the TD-LSTM proposed in this study, and T-LSTM is the first algorithmic model used to interpolate missing data in regular time-series (Baytas *et al.* 2017).

Previous to the process of testing the interpolating effects of the TD-LSTM, the three missing

datasets were first normalized. The normalization is expressed as:

$$x_n' = \frac{x_n - x_n^{min}}{x_n^{max} - x_n^{min}} \tag{7}$$

Where, $x_n^{min}$, $x_n^{max}$ are the minimum and maximum values of the n-*th* variable, respectively. For evaluating the interpolating effects of the algorithmic models, the mean square error (MSE) was used as the index to calculate the error rates between the values interpolated by the models and the original values. The calculation is as follows:

$$MSE = \frac{\sum_{i=1}^{N}\sum_{j=0}^{P_n}(y_j^i - y_j^{i'})^2}{num\_missing} \times 100\% \tag{8}$$

Where, $num\_missing$ is the total number of missing values of each variable in the dataset; n is the number of variables; $P_n$ is the number of missing values of the n-th variable; $y_j^i$ is the original values corresponding to the j-th missing value of the *i*-th variable; and $y_j^{i'}$ is the estimated value of the j-th missing value of the i-th variable.

During the testing, we took the *keras package* of Python to build the algorithmic models. All the four algorithmic models (the RNN, LSTM, T-LSTM, and TD-LSTM) were laid with two layers of networks; each layer had 32 channels. The compression ratio of SENet in TD-LSTM was 1/4 as set by the pioneer of SENet, Hu *et al.* (2018). The initial learning rate was set at 0.02, decaying to 90% after 50 epochs; 3000 epochs were learned in total. The optimization algorithm was done by RMSprop. Although the hyper-parameters had not been carefully tuned since these were not the focus of this study, we tried with some regular settings manually, and yielded reasonable robustness of the results.

After the testing was done, the results were downloaded and presented in Figure 6. As shown there, all the four algorithmic models can effectively interpolate the missing data in irregular time-series. In regard to the three datasets with different missing rates, the interpolating effects generally tend to get weaker as the missing rates increase for all the four models, which is specifically seen in the MSE dropping down from 0.51, 0.51, 0.48 to 0.42 in turn for LSTM, T-LSTM, RNN, and TD-LSTM at the missing rate of 31%, from 0.53, 0.55, 0.50 to 0.49 at 40%, and 0.66, 0.63, 0.57 to 0.53 at 50%. As for an individual algorithmic model, the TD-LSTM can interpolate most effectively (0.42 at the missing rate of 30%, 0.49 at 40%, and 0.53 at 50%) as compared with the other three models (0.51, 0.51, and 0.48 for LSTM, the T-LSTM, the RNN at the missing rate at 31%, 0.53, 0.55, and 0.50 at 40%, and 0.66, 0.63, and 0.57 at 50%) in all the three datasets of different missing rates.

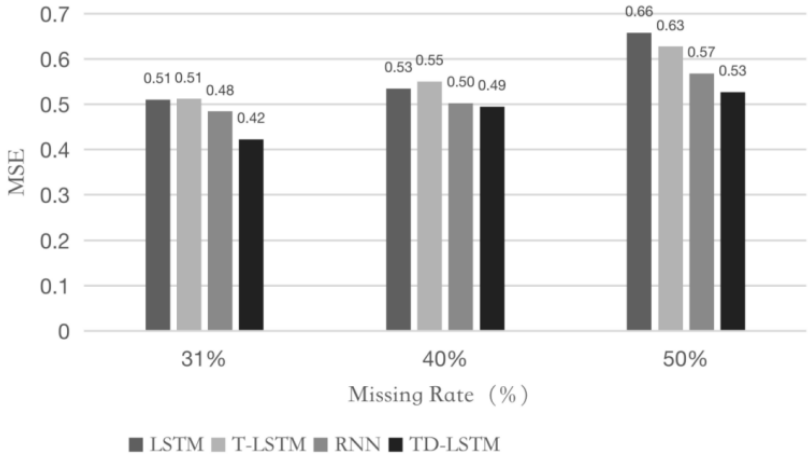


Figure 6 Interpolating effects of four algorithmic models

In summary, the TD-LSTM proposed in this study, like the other algorithmic models such as the RNN, LSTM, and T-LSTM, works in interpolating the missing data at different missing rates in irregular time-series. Moreover, the interpolation effects of the TD-LSTM are the best of the four algorithmic models. In other words, the TD-LSTM is able to monitor the missing data in irregular time-series by capturing more and better time-series information, and thus interpolates the missing data more efficiently.

## 4.2 Predictability in real datasets

Many studies in CRP claim that various algorithms of machine learning, such as DNN, SVM, NN, and RF, can do well in the real datasets with missing data. The truth is that, with the number of the indicators being expanded, not all the algorithms can meet the requirements well under different conditions for CRP. Among the algorithms mentioned above, DNN outperforms the others because it can learn and make informed decisions on its own (LeCun *et al.*, 2015) and has demonstrated state-of-the-art accuracy in the prediction of many intelligent tasks (Dargan *et al.*, 2020; Tian *et al.*, 2020). Thus, this study used DNN to predict the CRP of a real dataset with the missing data in irregular time-series before and after TD-LSTM interpolation. We used stochastic gradient descent as the optimizer and Dropout to prevent overfitting in the DNN algorithmic processing. The specific parameters set for the algorithm of the DNN are shown in Table 1.

Table 1 Parameters set for the DNN

| Parameters | DNN |
|---|---|
| Hidden units | (32, 32, 32,) |
| Activation | (ReLU, ReLU, sigmoid) |
| Dropout rate | (0.5, 0.5, 0,) |
| Optimizer | RMSprop |
| Learning rate | 1e-4 |

Based on the parameters set for DNN in Table 1, we verified the predictability of the TD-LSTM in

the real database that had been obtained in Section 3. The area under the curve (AUC) was used as the index for measuring the predictability of the algorithmic models. As we know, an algorithm having a larger AUC shares better predictability. For the database prepared in Section 3, we split it both before and after interpolation into the training and testing sets in a ratio of 70% to 30%, and input the data into Python for algorithmic processing. The AUC presented by DNN is as high as 84.15% for the dataset after interpolation. In order to compare the predictability of DNN with the other algorithmic models frequently used in CRP, we adopted SVM, RF, and NN with one hidden layer. The parameters of SVM and RF were set as default settings in *scikit-learn package* of Python, while the NN employed an artificial neural network of one hidden layer with 32 hidden channels. Although the parameters of SVM, NN, and RF were not tuned specifically, it was found that the results were quite stable within the settings of regular hyper-parameters. Table 2 shows the results of four different algorithmic models for predicting in datasets with missing data in irregular time-series both before and after interpolation.

Table 2 The predictability of four algorithmic models before and after interpolation

| **Algorithms** | **Area under curve (AUC)** | |
|---|---|---|
| | **Before Interpolation** | **After Interpolation** |
| DNN | 80.13% | 84.15% |
| SVM | 73.51% | 79.40% |
| RF | 76.50% | 78.84% |
| NN | 69.47% | 76.53% |

As shown in Table 2, in regard of the differences of the predictability for the four algorithmic models before and after interpolation, the AUCs of four algorithmic models are all higher after interpolation than those before interpolation (84.15%-80.13%, 79.40%-73.51%, 78.84%-76.50%, and 76.53%-69.47% for DNN, SVM, RF, and NN, respectively). As for the differences of the predictability amongst the four individual algorithmic models, the ranks of the AUC share a similar pattern with slight alteration both before and after interpolation, that is, DNN is the best, and NN is the last (80.13%-69.47% before interpolation, and 84.15%-76.53% after interpolation). The alteration lies in the rank shift between SVM and RF, that is, before interpolation, RF (76.50%) takes the second place, and SVM the third (73.51%); but after interpolation, SVM (79.40%) overtakes RF (78.84%).

To sum up, DNN, like the other algorithmic models such as SVM, NN, and RF can be employed in CRP for the missing data both before and after the interpolation by the TD-LSTM proposed in this study, on top of which DNN outperforms the other three models in interpolation.

# 5. Conclusion

In order to raise the timeliness and reliability of CRP, researchers have tried to mine information from any source that may involve the target company, such as the data offered by the upstream and downstream suppliers and customers along its supply chain. The variety of data sources may help with

CRP in the range of its indicators. However, since the data collected from different sources do not follow a fixed sampling time frequency, the extension of the indicators may, to quite a degree, cause a mismatch of the data, and end up with a more severe situation of data missing in irregular time-series, and in turn impair the CRP effect. To maintain the extension of indicators while overcoming their weaknesses, this study first introduced the methods for building the dataset and an algorithmic model, the TD-LSTM, to interpolate the missing data in irregular time-series. This study then tested the interpolation effects of the TD-LSTM in artificial datasets with varying missing rates and compared the TD-LSTM to other models for predictability before and after interpolation in a real dataset.

The significance of this study in methodology is summarized in three steps: (1) we added two gates (the time-decayed gate and the refreshing gate) to the traditional LSTM for better detecting and handling the state of missing data in irregular time-series before and after interpolation. (2) We added a time-delayed function to the traditional LSTM, which could be used globally in the TD-LSTM with great flexibility. (3) We incorporate SENet and SAE into the traditional LSTM to improve the generalization ability of the algorithm.

Apart from the innovated methodology, the empirical analysis of three artificial datasets with different missing rates and a real dataset with the missing data in irregular time-series yielded the following conclusions: (1) The TD-LSTM proposed in this study can monitor the missing data in irregular time-series by capturing more and better time-series information, and interpolating the missing data efficiently; (2) The DNN can be used in CRP for the dataset with the missing data in irregular time-series after the interpolation by the TD-LSTM. Furthermore, this study provides a good starting point for further research into new algorithmic models for interpolating missing data with specific features.

Admittedly, this study still has some limitations. Firstly, although this study has proved the fusion of different sources along the supply chain is a feasible scheme to improve CRP, the other sources can be introduced into working out better CRP, such as customers' behavior and users' portrait. Secondly, the public emergency can also pull down the credit rating of a company within a short period, for example, the COVID-19 has made the CRP much more difficult (Luo, 2021). In future research, still more sources and powerful algorithmic models should be encouraged to generate an even stronger and more time-friendly CRP.

# References


Agca, S., Babich, V., Birge, J. R., & Wu, J. (2022). Credit shock propagation along supply chains: Evidence from the CDS market. *Management Science, 68*(9), 6506-6538.

Appelhans, T., Mwangomo, E., Hardy, D. R., Hemp, A., & Nauss, T. (2015). Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. *Spatial Statistics, 14*, 91-113.

Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing, 86*, 105936.

Ashraf, M., Michas, P. N., & Russomanno, D. (2020). The impact of audit committee information technology expertise on the reliability and timeliness of financial reporting. *The Accounting Review, 95*(5), 23-56.

Bakoben, M., Bellotti, T., & Adams, N. (2020). Identification of credit risk based on cluster analysis of account behaviours. *Journal of the Operational Research Society, 71*(5), 775-783.

Bankó, Z., & Abonyi, J. (2012). Correlation based dynamic time warping of multivariate time series. *Expert Systems with Applications, 39*(17), 12814-12823.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., & Zhou, J. (2017). *Patient subtyping via time-aware LSTM networks.* Paper presented at the Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology, 4*(1), 111-138.

Bu, D., Kelly, S., Liao, Y., & Zhou, Q. (2018). A hybrid information approach to predict corporate credit risk. *Journal of Futures Markets, 38*(9), 1062-1078.

Chang, C., Deng, Y., Jiang, X., & Long, Q. (2020). Multiple imputation for analysis of incomplete data in distributed health data networks. *Nature Communications, 11*(1), 1-11.

Chang, Y. C., Chang, K. H., & Wu, G. J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing, 73*, 914-920.

Chi, Q., & Li, W. (2017). Economic policy uncertainty, credit risks and banks' lending decisions: Evidence from Chinese commercial banks. *China Journal of Accounting Research, 10*(1), 33-50.

Collin-Dufresne, P., Goldstein, R., & Hugonnier, J. (2004). A general formula for valuing defaultable securities. *Econometrica, 72*(5), 1377-1407.

Ding, Z., Mei, G., Cuomo, S., Li, Y., & Xu, N. (2020). Comparison of estimating missing values in iot time series data using different interpolation algorithms. *International Journal of Parallel Programming, 48*(3), 534-548.

Fekade, B., Maksymyuk, T., Kyryk, M., & Jo, M. (2017). Probabilistic recovery of incomplete sensed data in IoT. *IEEE Internet of Things Journal, 5*(4), 2282-2292.

Fouladgar, N., & Främling, K. (2020). A novel LSTM for multivariate time series with massive missingness. *Sensors, 20*(10), 2832.

Gilchrist, S., & Mojon, B. (2018). Credit risk in the euro area. *The Economic Journal, 128*(608), 118-158.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18*(5-6), 602-610.

Guo, X., Jarrow, R. A., & Zeng, Y. (2009). Credit risk models with incomplete information. *Mathematics of Operations Research, 34*(2), 320-332.

Haqq, A. & Budiwitjaksono, G. S. (2019). Fraud Pentagon for Detecting Financial Statement Fraud. *Journal of Economics, Business, and Accountancy Ventura, 22*(3), 319-332.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8),

1735-1780.
Holder-Webb, L., & Sharma, D. S. (2010). The effect of governance on credit decisions and perceptions of reporting reliability. *Behavioral Research in Accounting, 22*(1), 1-20.
Hu, J., Shen, L., & Sun, G. *et al.* (2018). *Squeeze-and-excitation networks.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Kaur, H., Pham, N., & Fomel, S. (2019). Seismic data interpolation using CycleGAN. In *SEG technical program expanded abstracts 2019* (pp. 2202-2206): Society of Exploration Geophysicists
Kousenidis, D. V., Ladas, A. C., & Negkakis, C. I. (2019). Aggregate accounting data and the prediction of credit risk. *The International Journal of Accounting, 54*(01), 1950001.
Kreindler, D. M., & Lumsden, C. J. (2016). The effects of the irregular sample and missing data in time series analysis. In *Nonlinear Dynamical Systems Analysis for the Behavioral Sciences Using Real Data* (pp. 149-172): CRC Press.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.
Lee, C. H., Yang, H. C., Wei, Y. C., & Hsu, W. K.. (2021). Enabling blockchain based scm systems with a real time event monitoring function for preemptive risk management. *Applied Sciences, 11*(11), 4811.
Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software, 26*(12), 1647-1659.
Luo, J. (2021). Forecasting COVID-19 pandemic: Unknown unknowns and predictive monitoring. *Technological Forecasting and Social Change, 166*, 120602.
Ma, J., Cheng, J. C., Jiang, F., Chen, W., Wang, M., & Zhai, C. (2020). A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings, 216*, 109941.
Ma, X., & Lv, S. (2019). Financial credit risk prediction in internet finance driven by machine learning. *Neural Computing and Applications, 31*(12), 8359-8367.
Mählmann, T. (2006). Estimation of rating class transition probabilities with incomplete data. *Journal of Banking & Finance, 30*(11), 3235-3256.
Mahmoudi, N., & Duman, E. (2015). Detecting credit card fraud by modified Fisher discriminant analysis. *Expert Systems with Applications, 42*(5), 2510-2516.
Mansi, S. A., Maxwell, W. F., & Zhang, A. J. (2012). Bankruptcy prediction models and the cost of debt. *The Journal of Fixed Income, 21*(4), 25-42.
Mikalsen, K. Ø., Bianchi, F. M., Soguero-Ruiz, C., & Jenssen, R. (2018). Time series cluster kernel for learning similarities between multivariate time series with missing data. *Pattern Recognition, 76*, 569-581.
Moat, H. S., Olivola, C. Y., Chater, N., & Preis, T. (2016). Searching choices: Quantifying decision-making processes using search engine data. *Topics in Cognitive Science, 8*(3), 685-696.
Ngueilbaye, A., Wang, H., Mahamat, D. A., & Junaidu, S. B. (2021). Modulo 9 model-based learning for missing data imputation. *Applied Soft Computing, 103*, 107167.
Ni, D., Lim, M. K., Li, X., Qu, Y., & Yang, M. (2022). Monitoring corporate credit risk with multiple data sources. *Industrial Management & Data Systems* (ahead-of-print).
Osadchiy, N., Gaur, V., & Seshadri, S. (2016). Systematic risk in supply chain networks. *Management Science, 62*(6), 1755-1777.
Pandey, T. N., Jagadev, A. K., Mohapatra, S. K., & Dehuri, S. (2017). *Credit risk analysis using machine learning classifiers.* Paper presented at the 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).
Pratama, I., Permanasari, A. E., Ardiyanto, I., & Indrayani, R. (2016). *A review of missing values handling methods on time-series data.* Paper presented at The 2016 International Conference on Information Technology Systems and Innovation (ICITSI).
Qu, L., Li, L., Zhang, Y., & Hu, J. (2009). PPCA-based missing data imputation for traffic flow

volume: A systematical approach. *IEEE Transactions on Intelligent Transportation Systems, 10*(3), 512-522.
Reurink, A. (2018). Financial fraud: a literature review. *Journal of Economic Surveys, 32*(5), 1292-1325.
Savitri, E., Andreas, A., & Surya, R. A. S. (2019). Influencing factors: the timeliness of financial reporting submissions. *Business and Management Studies, 5*(1), 43.
Scannella, E., & Polizzi, S. (2021). How to measure bank credit risk disclosure? Testing a new methodological approach based on the content analysis framework. *Journal of Banking Regulation, 22*(1), 73-95.
Stanimirova, I., Daszykowski, M., & Walczak, B. (2007). Dealing with missing values and outliers in principal component analysis. *Talanta, 72*(1), 172-178.
Stein, R. M. (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking & Finance, 29*(5), 1213-1236.
Sun, C., Chen, Y., & Cheng, C. (2021). Imputation of missing data from offshore wind farms using spatio-temporal correlation and feature correlation. *Energy, 229*, 120777.
Tan, Q., Ye, M., Yang, B., Liu, S., Ma, A. J., Yip, T. C. F., ... & Yuen, P. (2020). *Data-gru: Dual-attention time-aware gated recurrent unit for irregular multivariate time series.* Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.
Teles, G., Rodrigues, J. J. P. C., Rabê, R. A., & Kozlov, S. A. (2020). Artificial neural network and Bayesian network models for credit risk prediction. *Journal of Artificial Intelligence and Systems, 2*(1), 118-132.
Tian, Y., Zhang, K., Li, J., Lin, X., & Yang, B. (2018). LSTM-based traffic flow prediction with missing data. *Neurocomputing, 318*, 297-305.
Tian, Z., Xiao, J., Feng, H., & Wei, Y. (2020). Credit risk assessment based on gradient boosting decision tree. *Procedia Computer Science, 174*, 150-160.
Trustorff, J. H., Konrad, P. M., & Leker, J. (2011). Credit risk prediction using support vector machines. *Review of Quantitative Finance and Accounting, 36*(4), 565-581.
Wang, H. Y., Li, H., & Shen, J. Y. (2019). A novel hybrid fractal interpolation-SVM model for forecasting stock price indexes. *Fractals, 27*(04), 1950055.
Wang, Y., Wang, B., Tu, N., & Geng, J. (2020). Seismic trace interpolation for irregularly spatial sampled data using convolutional autoencoder. *Geophysics, 85*(2), V119-V130.
Wu, J., Zhang, Z., & Zhou, S. X. (2022). Credit rating prediction through supply chains: A machine learning approach. *Production and Operations Management, 31*(4), 1613-1629.
Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic Analysis of the Digital Economy* (pp. 89-118): University of Chicago Press
Wubetie, H. T. (2017). Missing data management and statistical measurement of socio-economic status: application of big data. *Journal of Big Data, 4*(1), 1-44.
Xie, Y., Li, A., Liu, Z., Chen, K., & Tu, H. (2021, 9-11 Oct. 2021). *A Novel Deep Neural Network Model for Credit Risk Prediction of Chinese Farmers.* Paper presented at the 2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC).
Yoon, J., Zame, W. R., & van der Schaar, M. (2019). Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. *IEEE Transactions on Biomedical Engineering, 66*(5), 1477-1490.
Yu, B., Li, C., Mirza, N., & Umar, M. (2022). Forecasting credit ratings of decarbonized firms: Comparative assessment of machine learning models. *Technological Forecasting and Social Change, 174*, 121255.
Yu, L., Zhou, R., Tang, L., & Chen, R. (2018). A DBN-based resampling SVM ensemble learning paradigm for credit classification with imbalanced data. *Applied Soft Computing, 69*, 192-202.
Zamore, S., Ohene Djan, K., Alon, I., & Hobdari, B. (2018). Credit risk research: Review and

agenda. *Emerging Markets Finance and Trade, 54*(4), 811-835.

Zhang, M., Lucas, J., Ba, J., & Hinton, G. E. (2019). Lookahead optimizer: k steps forward, 1 step back. *Advances in Neural Information Processing Systems, 32*.

Zhang, W., Yan, S., Li, J., Tian, X., & Yoshida, T. (2022). Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E: Logistics and Transportation Review, 158*, 102611.