Multiperiod corporate default prediction –A domain knowledge-tailored neural network approach

Wei-Lun Luo, Chuan-Ju Wang
Research Center for Information Technology Innovation, Academia Sinica, Taiwan, awilliea@citi.sinica.edu.tw, cjwang@citi.sinica.edu.tw

Jin-Chuan Duan
Asian Institute of Digital Finance National University of Singapore, bizdjc@nus.edu.sg

Ming-Feng Tsai
Department of Computer Science, National Chengchi University, Taiwan, mftsai@nccu.edu.tw

A domain knowledge-tailored neural network for multiperiod default prediction is proposed. Instead of using conventional neural networks, we customize them using economic domain knowledge. This allows us to regulate the model's performance and avoid overfitting. To attest the effectiveness of our approach, experiments are conducted from 1994-2021 on a sizable US corporate default dataset. The results demonstrate that our model performs considerably better than the state-of-the-art econometric model, and it is more robust than conventional neural networks. In addition, we find our model performs well on several dataset subgroups. Although its predictive power degrades in the high credit risk years for the long-term horizons, its performance still can be reasoned. Our proposed method can be applied to most neural networks, providing inspiration for current machine-learning research on financial applications.

Key words: default analysis, machine learning, neural network, deep learning, economic domain knowledge

1. Introduction

Default risk is a lender taking its risk on a borrower who may unable to meet its obligation under the agreed terms Lando (2009). Thus, credit risk management and the development of financial policies all depend on an understanding of the factors that cause a default. Due to their different debt structures, firms may have different risk profiles for long-term debts and short-term debts. Under this phenomenon, it is critical to consider the term structure in detail for credit analysis. Although the previous research can provide short-term and long-term risk rates, it still does not comprehensively consider the term structure of default probabilities. Structured and reduced-form techniques are the two main categories into which credit risk modeling may be divided. In this study, we use a reduced-form technique to predict the default term structures.

Altman (1968) was the earliest research that uses a reduced-form technique to address the task of default prediction. The primary methodology used in this research was discriminant analysis, which

produces credit scores that only provide ordinal ranks. Later, Ohlson (1980), Zmijewski (1984) change the methods from discriminant analysis to binary response models, such as logit regression, but their methods can only generate a default probability for a given period. To address this issue, Campbell et al. (2008) used a multiple logit model, which can generate the default probabilities for different prediction horizons. Recently, Duffie et al. (2007) (DSW) used a doubly stochastic Poisson intensity model to model default occurrences. In their model, in order to generate the term structure of default probabilities for multiperiod default prediction, it is necessary to specify time dynamics of the state variables, i.e. modeling a high-dimensional time series, which is hard to implement in real-world applications. Duan et al. (2012) (FIM), applied a forward intensity model to the task of default prediction. Like DSW, it can predict both default and other kinds of corporate exits like mergers and acquisitions, but its approach did not need to explicitly model and estimate the high-dimensional state variable process. It is now viewed as a state-of-the-art econometric approach.

Compared with econometric approaches, machine-learning-based approaches have more complex functional forms, enabling them can have a chance to obtain a better performance. However, most of the machine-learning research (Yeh et al. (2014, 2015), Huang et al. (2004), Ribeiro et al. (2012), Alakar et al. (2018), Sarkar et al. (2018), Sirignano et al. (2016), Eom et al. (2020)) failed to generate the term structure of default probabilities because they viewed default prediction as a ranking task. Recently, there are two research solve the previous problem. Luo et al. (2022) used a carefully designed neural network to generate a consistent term structure of cumulative default probabilities. Although it can obtain a good performance, it still lacks the information for default events because it only models cumulative default probabilities instead of the forward intensities of default events. Divernois (2020) has addressed the previous issue by using neural networks to predict the forward intensities through the framework in Duan et al. (2012). Nevertheless, its approach does not consider the overfitting issue of its neural networks, which is an important problem existing in the current machine-learning research. Besides, its method is not evaluated in overtime experiments, which may not be applicable in real-world applications.

In this paper, we propose a domain knowledge-tailored neural network to address the task of multiperiod corporate default prediction. Like Divernois (2020), we follow the framework in Duan et al. (2012) that can generate the forward intensities for default and the other-exits events. We can also ensure that our model can obtain a consistent cumulative default term structure with this framework. Compared with Divernois (2020), our main difference is that we use economic domain knowledge to tailor our neural model. With this design, our model can have fewer parameters and still maintain strong predictive power for default prediction, which can be proved in our empirical results.

Our experiments are conducted in a large panel dataset covering most of the companies in the US from 1994 to 2021. We analyze our model through both a cross-sectional experiment and an overtime experiment. With the results of these two kinds of experiments, it is obvious how a machine-learning model, especially neural networks, can be easily overfitting. Based on these results, our domain knowledge-tailored (DKT) approach outperforms the econometric approach and the traditional neural networks. We can further describe the superiority of our DKT approach from two aspects. Compared with the econometric approach, our DKT model has more complex functional forms, indicating it can achieve better predictive power. Compared with the traditional neural networks, our DKT approach can regulate the model to gain better performance and prevent it from overfitting. For a more comprehensive analysis, we also conduct overtime experiments to analyze the model's performance on different sub-groups of our dataset.

2. Methodology

In this section, we describe the details of our methodology, including the framework we use to formulate the multi-period default prediction, how to apply machine learning methods to it, and our proposed domain knowledge-tailored methods.

2.1. Framework: A Forward-Intensity Approach

We follow the forward-intensity approach of Duan et al. (2012) to model the corporate default events for companies. Under this framework, a firm's default probability is generated by two doubly stochastic Poisson processes with one governing default and the other forming of corporate exit, for example, a merger/acquisition. In contrast to the typical approach of modeling through spot intensities, forward intensities which depend on the known values of feature variables at the time of prediction would describe the characteristic of the future events and estimated by the values of the input variables at the prediction time. With this method, we can forecast the default probabilities for any prediction horizon (e.g., 1 month, 12 months, and 5 years) without knowing the variables in the future. Another critical issue in default prediction is that a company can exit the market for other reasons (e.g., mergers and acquisitions) than default or bankruptcy. To comprehensively address the default prediction task, we need to take these " other events" into account by using another independent Poisson process to model them. Besides, except for the default events and the other-exit events, we also use a Poisson process to model the survival events for a company, which represents that a company is still surviving in the market. In this way, since a company can only default, other-exit, or survive in the market, these three independent Poisson processes are mutually exclusive.

To formulate the above description mathematically, we use $f_m(X_{i,t})$ and $q_m(X_{i,t})$ to represent the forward intensity of default and other-exit Poisson processes respectively for the time interval

between m to $m + \Delta t$, where $X_{i,t}$ denotes the covariates of the ith company at prediction time t, Δt denotes the minimum unit of time interval (which is one month in this paper), and m denotes the prediction horizon. In other words, these two forward intensities are generated by the f_m and q_m functions with the same input variable $X_{i,t}$. Because we can only use the past information to generate the forward intensity of a Poisson process, m is greater than or equal to t.

Based on these two kinds of forward intensities and the definition of a Poisson process, we can first formulate the survival probability as Equation 1, calculating the probability that there are no default and other-exit events. Secondly, since a default event is signaled by a jump in a Poisson process, its probability is defined by Equation 2 as a function of its forward intensity. Thirdly, because these three independent Poisson processes are mutually exclusive, we can calculate the other-exit forward intensity by Equation 3, which is 1 minus the survival and default probability.

$$p_s(X_{i,t}; m) = e^{-(f_m(X_{i,t}) + q_m(X_{i,t})) * \Delta t} \tag{1}$$

$$p_d(X_{i,t}; m) = 1 - e^{-f_m(X_{i,t}) * \Delta t} \tag{2}$$

$$\begin{aligned} p_o(X_{i,t}; m) &= 1 - p_s(X_{i,t}; m) - p_d(X_{i,t}; m) \\ &= e^{-f_m(X_{i,t})} \left(1 - e^{-q_m(X_{i,t}) * \Delta t}\right) \end{aligned} \tag{3}$$

With the above forward default and survival probabilities, the cumulative default probabilities of a given company from time t to $t + n * \Delta t$ can be computed by Equation 4, which is a sum of conditional default probabilities. To be more specific, the conditional default probability for the time interval between m to $m + \Delta t$ is calculated by the forward default probability times the product of the survival probabilities before time m.

$$Prob[X_{i,t}, n; \Delta t] = \sum_{m=0}^{n-1} \left[p_d(X_{i,t}; m) \prod_{j=0}^{m-1} p_s(X_{i,t}; m) \right] \tag{4}$$

While the above equations can help us model the three types of forward probabilities and the cumulative default probability, how to generate the forward intensities is a critical issue. In Duan et al. (2012), it uses linear regressions to generate the forward intensities, the process of which can be seen in Equations 5 and 6. In Equation 5, the default forward intensity for the prediction horizon m is computed by an inner product between $\beta(m)$ and $X_{i,t}$, where $\beta(m)$ is the coefficient vector of linear regression and $X_{i,t}$ is the covariate vector, and then be exponentialized to ensure to obtain a positive intensity. The superscript "FIM" denotes that the forward intensity is computed by the method used in Duan et al. (2012). Likewise, the other-exit forward intensity is calculated in the same way by Equation 6, where $\bar{\beta}(m)$ is the coefficient vector of linear regression specifically for other-exit events. Based on the estimated forward intensities, cumulative default probabilities

for different time intervals can be computed by Equation 4. The term structure of the cumulative default probabilities can be further analyzed for the task of multi-period default prediction.

$$\begin{aligned} f_m^{\text{FIM}}(X_{i,t}) &= \exp\left(\beta_0(m) + \beta_1(m)x_{i,t,1} + ... + \beta_k(m)x_{i,t,k}\right) \\ &= \exp\left(\beta(m) \cdot X_{i,t}\right) \end{aligned} \tag{5}$$

$$\begin{aligned} q_m^{\text{FIM}}(X_{i,t}) &= \exp\left(\bar{\beta}_0(m) + \bar{\beta}_1(m)x_{i,t,1} + ... + \bar{\beta}_k(m)x_{i,t,k}\right) \\ &= \exp\left(\bar{\beta}(m) \cdot X_{i,t}\right) \end{aligned} \tag{6}$$

2.2. Conventional Machine Learning Approaches

2.2.1. Multi-layer Perceptron Although FIM can generate the forward intensities and probabilities in a more simple and explainable way, its limited functional form constrains its performance. To obtain a better performance, Divernois (2020) replaced the simple linear regression with MLP(multi-layer perceptron), one of the machine learning-based models, which has more functional complexity, to generate forward intensities. For simplicity, in the setting of our experiments, our MLP would generate the two types of forward intensities for all prediction horizons at once. The process of the generation can be formulated in Equation 7, where the right arrow is a mapping operator, indicating that the function on the right side can map its arguments and parameters to the outputs on the left side. To be more specific, at the right side of the arrow, Θ^{MLP} is the function representing the multi-layer perceptron, in which $X_{i,t}$ is the input variables, θ_{MLP} is the parameters of the MLP, and n is a parameter deciding how many prediction horizons for each forward intensity the MLP would output. On the other side, $f_m^{MLP}(X_{i,t})$ and $q_m^{MLP}(X_{i,t})$ denote the default and other-exit forward intensities generated by MLP. It is worth noting that there are brackets with subscripts outside of the two forward intensities, which is $(f_m^{\text{MLP}}(X_{i,t}), q_m^{\text{MLP}}(X_{i,t}))_{m=0,1,...,n-1}$, meaning that MLP can generate these two kinds of forward intensities for different prediction horizons at once.

$$\left(f_m^{\text{MLP}}(X_{i,t}), q_m^{\text{MLP}}(X_{i,t})\right)_{m=0,1,...,n-1} \rightarrow \Theta^{\text{MLP}}(X_{i,t}; \theta_{MLP}, n) \tag{7}$$

2.2.2. Gated-Recurrent Unit While MLP can have more complex functionality than FIM, a recurrent-based deep learning model may have a better ability to succeed in a time-series-based task. With this thought, we apply Gated-Recurrent Unit(GRU) to the task of multi-period default prediction by Equation 8. This equation is similar to Equation 7, where the superscripts and the subscripts are changed from MLP to GRU, indicating that the values are for different models. The most different part of these two equations is the input variables of the model function. For MLP, it only takes the covariates of a given company at the current time t. However, for GRU, it takes the covariates of a given company in the past 12 months before the current time t. Compared with

MLP, GRU can better capture the time dynamics of the input variables and then obtain a better performance. For example, if our model can capture both the values and the momentum of the covariates of a given company, it can produce more accurate default probabilities theoretically.

$$\left(f_m^{\text{GRU}}(X_{i,t}), q_m^{\text{GRU}}(X_{i,t})\right)_{m=0,1,\ldots,n-1} \rightarrow \Theta^{\text{GRU}}(X_{i,t-11}, ..., X_{i,t}; \theta_{GRU}, n) \tag{8}$$

2.3. Domain Knowledge Tailored (DKT) Approaches

2.3.1. The Basic Unit of DKT Machine learning-based models can obtain better performance via their complex functional forms, but they may more likely be overfitting. That is what we call a bias-variance trade-off. In order to regulate our models, we leverage economic domain knowledge to tailor our neural networks. In other words, we remove some weights of our neural models through insights from economic research. With this design, we can simplify our model and prevent it from overfitting in an appropriate way. To be more specific, we can first focus on how we transform the fully connected layers, a basic unit in deep-learning models, into tailored ones through economic domain knowledge.

To describe how to tailor a fully connected layer, we need to consider its mechanism from a different perspective. Traditionally, a fully connected layer is viewed as matrix multiplication. However, it can also be viewed as a multiple-grouping method. We can take Figure 1 as an example. In this figure, there are three input variables and three output nodes. For each output node, it is computed by a unique linear combination of each input variable based on the links between two nodes (we remove the activation function for simplicity). For example, the blue node may be computed by $1*n_1+2*n_2+3*n_3$, where n_1, n_2, n_3 are the input nodes and $1, 2, 3$ are the coefficient variables for the linear combination. Likewise, the red node may be computed by $2*n_1+3*n_2+1*n_3$, which has different coefficient variables compared to the other output nodes. Thus, in this example, there are three different linear combinations (the blue, green, and red lines), which can be viewed as three different grouping methods for the given input variables.

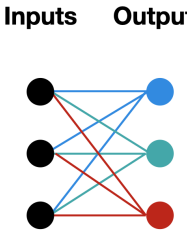

Figure 1 **An example of a fully connected layer with three nodes**

With the grouping perspective about a fully connected layer, we can explain its mechanism for our task in a new way: a fully connected layer can group a firm's covariates and then generate the outputs that are predictive for the default and other-exit forward intensities. However, due to the characteristic of a fully connected layer, the grouping methods of it are decided by the trained weights, some of which may be redundant or even harm the performance of the model. Based on this situation, it is helpful to remove some weights in a fully connected layer by replacing its own grouping methods with the ones that have more economic significance. This method can reduce the parameters of the model to prevent it from overfitting and leverage the economic domain knowledge to improve the predictive power of the model.

To tailor a fully connected layer, we first extract some important factors that cause default and other-exit events based on the findings in the previous economic research. To be more specific, we attribute default events to the financial indicators derived from the aspect of liquidity, profitability, solvency, and others (Zhang et al. (2005), Xie et al. (2011)). Liquidity is the ability of a firm to convert its asset to cash in the short term, which is a critical concern of default events. Compared with liquidity, solvency considers a firm's ability to meet its long-term obligations. Profitability is more intuitive, representing the ability of a firm to earn money. If a covariate does not belong to the previous three factors, we would group it into "others". Based on the definitions of these factors, we can categorize a firm's covariate into four groups. The details of the grouping methods are as follows:[1]

1. Interest rate: Because it can influence the interest that a company needs to pay for its short-term and long-term debts, we group it into "Liquidity" and "Solvency".
2. Return of S&P500 index: Because the return of the S&P500 index represents the overall environment of the financial market in the US, it can influence the liquidity constraints of the market, which is related to the interest rate. Therefore, we group it into "Liquidity" and "Solvency".
3. Financial Aggregate DTD and Non-Financial Aggregate DTD: These two covariates are like the return of the S&P 500 index which can reflect the situation of the financial market, so we also group them into "Liquidity" and "Solvency".
4. DTD: The calculation of it includes the short-term and long-term debts of a firm, so we group it into "Liquidity" and "Solvency".
5. CA/CL and CASH/TA: We group them into "Liquidity" like Ni et al. (2014).
6. NI/TA: The ratio reflects how much net income can a firm earn divided by its total assets, so it is a covariate that belongs to "Profitability".

[1] You can first see section 3.1 for the description of each covariate

7. Size: Because a larger firm can have higher efficiency and market power to generate higher profit(Lee (2009)), we group it into "Profitability".
8. M/B and SIGMA: In our opinion, these two covariates do not belong to the defined three groups, so we put them into "others".

The summary of the grouping can be seen in Table 1, where the row denotes each covariate and the column denote each group. In this table, 1 means the covariate belongs to the given group, and 0 means they do not belong to it. For example, in the first row, we group "Interest rate" to "Liquidity" and "Solvency".

Table 1 Grouping table for default events.

Covariates	Liquidity	Profitability	Solvency	Others
Interest rate	1	0	1	0
Return of the S&P500 index	1	0	1	0
Financial Aggregate DTD	1	0	1	0
Non-Financial Aggregate DTD	1	0	1	0
DTD	1	0	1	0
CASH/TA	1	0	0	0
CA/CL	1	0	0	0
NI/TA	0	1	0	0
Size	0	1	0	0
M/B	0	0	0	1
SIGMA	0	0	0	1

For the other-exit (M&A) events, existing literature (Rodrigues and Stevenson (2013)) shows that they basically follow the hypotheses of efficiency, growth-resource, valuation, and size. The hypothesis of efficiency assumes that the inefficient management of a company will be acquired by a more efficient firm, which can increase its market capitalization. The hypothesis of growth-resource describes that a firm with high growth and low resource or a firm with low growth and high resource has a higher probability to be acquired and vice versa. Compared with the previous two hypotheses, the hypothesis of valuation is quite simple. It states that there is a large likelihood for an undervalued firm to be acquired. For the hypothesis of size, there are two different perspectives. One states that a smaller firm is more likely to be acquired due to its lower acquisition cost. The other states that mergers prefer acquiring a larger firm to increase the size of their companies. Thus, it is not a linear relationship between acquisition and the size of the company, which can be modeled by a non-linear approach, such as machine-learning techniques. Likewise, based on the definitions of these hypotheses, we can categorize a firm's covariate into the four groups as follows:

1. Interest rate and Financial Aggregate DTD: Because the financial aggregate DTD represents the situation of the financial market and then can influence the interest rate which the financial firms announce, these two covariates can yield a great impact on the borrowing cost of a

company, which can be grouped into "Efficiency". Besides, the interest rate can be used as a discount rate for the calculation of the value of a firm, so we also categorize these two covariates into "Valuation".

2. Return of S&P500 index and Non-Financial Aggregate DTD: It is important to analyze a firm's value based on the overall market situation, which can be represented by the return of the S&P500 index and non-Financial aggregate DTD. So, we group these two covariates into "Valuation".
3. M/B: Market-to-book ratio is a classic financial indicator to see whether a stock is overvalued or undervalued. So, we group it into "Valuation".
4. DTD, CASH/TA, and CA/CL: We follow the setting in Rodrigues and Stevenson (2013) to categorize the covariates related to liquidity and solvency into "Growth-resource".
5. SIGMA: SIGMA can represent the deviation between a firm's stock value and the overall market, which indicates how the firm is managed. So, we group it into "Efficiency".
6. NI/TA: In our opinion, this covariate can be grouped into all categories. For example, a company with a high "NI/TA" can reflect that it is a well-managed company and has nice growth. Besides, "NI/TA" can also influence how we value a company and reflect its size.
7. Size: It is intuitive that we group this covariate into "Size".

Table 2 Grouping table for Merge and Acquisition events.

Covariates	Efficiency	Growth-resource	Valuation	Size
Interest rate	1	0	1	0
Financial Aggregate DTD	1	0	1	0
Return of the S&P500 index	0	0	1	0
Non-Financial Aggregate DTD	0	0	1	0
M/B	0	0	1	0
DTD	0	1	0	0
CASH/TA	0	1	0	0
CA/CL	0	1	0	0
SIGMA	1	0	0	0
NI/TA	1	1	1	1
Size	0	0	0	1

Likewise, the summary of the grouping can be seen in Table 2, where the row denotes each covariates and the column denote each group.

With the above grouping methods for the default and M&A events, we can finally transform a normal fully connected layer into a DKT(domain-knowledge tailored) version. A schematic diagram of it can be seen in Figure 2. Compared with the normal one, the DKT version of a fully connected layer has the same input covariates, but it has a fixed size of output nodes and predefined edges between the input nodes and the output nodes. To be more specific, in this figure, the blue nodes

represent the four groups("Liquidity", "Solvency", "Profitability", "Others") for the default events, and the green nodes represent the other four groups("Efficiency", "Growth-resource", "Valuation", "Size") for the M&A events. Besides, for each edge between the input and output nodes, it is followed the criteria we defined previously. For example, the input covariate "Interest rate" can only have the edges which link to the output nodes "Liquidity", "Solvency", "Efficiency", and "Valuation", which is half of all output nodes. With this approach, we can remove some redundant edges in a fully connected layer to regulate our model and obtain a better performance through economic domain knowledge.

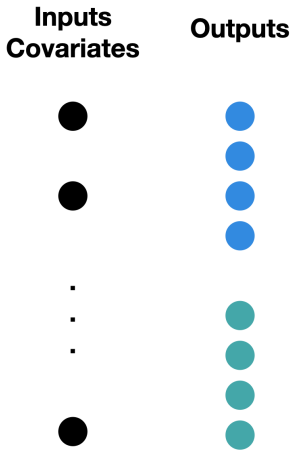

Figure 2 **An example of DKT framework**

2.3.2. GRU In the previous section, we introduce the basic unit of our DKT approach, which is a revised version of a fully connected layer. In this section, we will further describe how we apply our DKT approach to transform a normal GRU into a DKT one.

Before applying our DKT approach to GRU, it is necessary to see how a normal GRU works by Equations 9 to 12. Here, $X_{i,t}$ denotes the input covariates at time t, b_* (b_r ,b_z ,and b_h) are the bias vectors for different calculations, and h_{t-1} is the hidden vector at time $t-1$. Noted that $X_{i,t}$ represents the information at the current time and h_{t-1} represents the information from the past. With these definitions, we can first calculate the reset gate and the update gate for GRU in Equations 9 and 10. The calculations of the two gates are similar. They first apply a linear transformation to the input $X_{i,t}$ and the hidden vector h_{t-1} by an inner-product operation with different matrices W and U. The subscripts of each matrix denote that they are used for different calculations. For example, U_r and W_r are used to calculate the reset gate. Secondly, they sum each linear transformation together with the bias vectors b_r and b_z. Thirdly, due to the characteristic of a gate, they use a sigmoid function to transform the sum into a value between 0 and 1.

$$r_t = \sigma(\boldsymbol{U_r} \cdot h_{t-1} + \boldsymbol{W_r} \cdot X_{i,t} + b_r) \tag{9}$$

$$z_t = \sigma(\boldsymbol{U_z} \cdot h_{t-1} + \boldsymbol{W_z} \cdot X_{i,t} + b_z) \tag{10}$$

For $\hat{h_t}$, its calculation (Equation 11) is also similar to the reset gate and the update gate. The main difference in its calculation is that it uses a $tanh$ function instead of a sigmoid function, indicating that it is not a gate. Besides, before it uses a linear transformation on h_{t-1}, it first applies an element-wise product on reset gate r_t and h_{t-1}. The motivation for this operation is to filter the unimportant information in the past and retain the useful one for the current time.

$$\hat{h_t} = tanh\left(\boldsymbol{U_h} \cdot (r_t \odot h_{t-1}) + \boldsymbol{W_h} \cdot X_{i,t} + b_h\right) \tag{11}$$

After obtaining the above three vectors, GRU can calculate the hidden vector at time t h_t by Equation 12. It is a linear combination for $\hat{h_t}$ and h_{t-1}, and the weights for them are the update gate z_t and the opposite vector of it $(1 - z_t)$, indicating that GRU can consider updating its hidden vector by the past information and the current information.

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \hat{h_t} \tag{12}$$

For implementation, it is worth noting that the linear transformation is implemented by a fully connected layer. For example, $U_r \cdot h_{t+1}$ is conducted by a fully connected layer, where U_r is its parameters. Generally speaking, U_* and W_* are both the parameters for a give fully connected layer.[2]

2.3.3. Domain Knowledge Tailored GRU (DKT GRU) To apply our DKT approach, based on what we introduce in section 2.3.1, we first replace the matrices W_* from fully connected layers to our DKT layers. With this replacement, we can ensure that the input variables would follow the predefined grouping criteria by economic domain knowledge. Besides, because the hidden vectors h_t are generated by DKT layers, the value of each dimension of them has its own economic implication. To maintain the concept of the DKT approach (remaining the meaningful hidden vectors), we also replace the matrices U_* from fully connected layers with diagonal matrices. In other words, with these diagonal matrices, the hidden vectors would only be stretched instead of being grouped, indicating that there is no interaction between the values from each dimension of the hidden vectors. After replacing these two kinds of matrices, we can get a new version of GRU, which is what we called DKT_GRU. Compared with a normal GRU, our DKT_GRU has fewer parameters, which can prevent it from overfitting. Besides, although DKT_GRU is much simpler than a normal GRU, it can still catch useful economic dynamics through our well-designed DKT layers.

[2] U_* represent the set $\{U_r, U_z, U_h\}$ and W_* represent the set $\{W_r, W_z, W_h\}$.

Table 3 Data description

Year	Active Firms	Default/bankruptcies	(%)	Other exit	(%)
1994	6915	17	0.25	223	3.22
1995	7395	16	0.22	362	4.90
1996	7947	17	0.21	401	5.05
1997	8305	48	0.58	568	6.84
1998	8270	75	0.91	891	10.77
1999	7961	85	1.07	921	11.57
2000	7624	106	1.39	782	10.26
2001	6930	174	2.51	757	10.92
2002	6229	118	1.89	533	8.56
2003	5825	80	1.37	472	8.10
2004	5664	37	0.65	371	6.55
2005	5649	35	0.62	384	6.80
2006	5591	21	0.38	382	6.83
2007	5611	23	0.41	463	8.25
2008	5275	58	1.10	382	7.24
2009	4983	105	2.11	322	6.46
2010	4855	29	0.60	313	6.45
2011	4704	32	0.68	304	6.46
2012	4591	39	0.85	262	5.71
2013	4621	28	0.61	239	5.17
2014	4772	27	0.57	212	4.44
2015	4858	40	0.82	275	5.66
2016	4802	65	1.35	362	7.54
2017	4710	42	0.89	311	6.60
2018	4737	20	0.42	262	5.53
2019	4772	33	0.69	292	6.12
2020	4967	70	1.41	238	4.79
2021	5785	17	0.29	242	4.18

3. Data and Performance metrics

In this section, we describe the details of our data and performance metrics, including the statistics of our dataset, the descriptions of each variable, and the math formulas for our performance metrics.

3.1. Data

We conducted our experiments on the Credit Research Initiative (CRI) database, which is maintained by the Asian Institute of Digital Finance (AIDF) of the National University of Singapore. This dataset contains 17,560 public firms in the US and has given rise to 1,833,106 firm-month observations over the period from 1994 to 2021. We can see a summary of the dataset in Table 3. This table shows the number of active firms, defaults, and other exits for each year. As you can see, the overall default rate is ranging from 0.21% to 2.51% in each year, and the rate of other exits is much higher, which ranges between 3.22% and 11.57%.

In this dataset, there are 16 variables for each firm-month observation, containing 4 common variables and 12 firm-specific variables. As the description in the technical report of Credit Research Initiative (2020), these variables are chosen as having a great power to predict the corporate defaults in the US. The descriptions of each covariate are as follows:[3]

1. Common variables
   - Interest rate: 3-month short-term US Treasury bill rate.
   - Stock index return: the trailing one-year return on the S&P500 index.
   - Financial Aggregate DTD: median DTD of financial firms in the US.
   - Non-Financial Aggregate DTD: median DTD of non-financial firms in the US.
2. Firm-specific variables
   - DTD: firms' distance to default, which is used to measure volatility-adjusted leverage based on Merton (1974). The calculation of DTD for financial firms follows the setting in Duan et al. (2012).
   - NI/TA: the ratio of net income to the total assets, which is used to measure the profitability of a company.
   - CASH/TA: logarithm of the ratio of the sum of cash and short-term investments to the total assets, which is used to measure the liquidity of a financial firm.
   - CA/CL: logarithm of the ratio of the current assets to the current liabilities, which is used to measure the liquidity of a non-financial firm.
   - Size: logarithm of the ratio of a firm's market capitalization to the median market capitalization of the firms in the US over the past year.
   - M/B: a firm's market-to-book asset ratio divided by the median market-to-book ratio of the firms in the US.
   - SIGMA: 1-year idiosyncratic volatility, calculated following Shumway (2001). It is computed by regressing the daily return of a firm's market capitalization against the daily return of the S&P500 index. SIGMA is defined to be the standard deviation of the residuals from this regression.

For the first five firm-specific variables, we follow the setting in Duan et al. (2012) to transform them into the level and trend versions of the measures. The level is computed as the one-year average of the measure, and the trend is computed as the current value of the measure minus the one-year average of the measure. Duan et al. (2012) has proved that the usage of level and trend significantly improves the predictive power of the model in the short-term prediction horizons.

Compared to the previous research, among these variables, 12 of them are used in Duan et al. (2012) and 14 of them are used in Luo et al. (2022). The summary statistics of each covariate can be seen in Table 4.

[3] Please see the technical report Credit Research Initiative (2020) for more details.

Table 4 Covariates statistics

	Mean	Std.	Min	25%ptcl	Median	75%ptcl	Max
Interest rate	-0.080	0.852	-1.195	-1.052	-0.202	0.819	1.506
Return of S&P500	0.101	0.162	-0.488	0.031	0.115	0.201	0.668
Financial Aggregate DTD	0.591	1.257	0.000	0.000	0.000	0.000	5.119
Non-financial Aggregate DTD	2.758	1.627	0.000	2.172	3.122	3.961	5.427
DTD_{level}	3.949	3.460	-1.603	1.970	3.332	5.151	53.653
DTD_{trend}	-0.060	1.180	-11.698	-0.558	-0.012	0.496	6.095
$CASH/TA_{level}$	-0.004	0.252	-3.311	0.000	0.000	0.000	3.224
$CASH/TA_{level}$	-0.684	1.464	-9.617	0.000	0.000	0.000	0.000
NI/TA_{level}	-0.004	0.032	-0.916	-0.003	0.001	0.005	0.201
NI/TA_{trend}	-0.000	0.025	-0.491	-0.002	0.000	0.002	0.460
$Size_{level}$	0.088	1.999	-5.912	-1.345	-0.018	1.411	6.673
$Size_{trend}$	-0.016	0.340	-1.905	-0.161	-0.006	0.145	1.985
M/B	1.590	2.793	0.157	0.760	0.986	1.571	75.891
SIGMA	0.170	0.117	-0.072	0.088	0.138	0.217	1.106
CA/CL_{level}	-0.012	0.293	-2.472	-0.074	0.000	0.049	2.584
CA/CL_{trend}	0.655	0.814	-3.817	0.000	0.547	1.090	4.781

3.2. Performance Metrics

To comprehensively evaluate the performance, we measure our model in two different perspectives: the discriminatory power of risk ranking among companies and the matching ability between actual default rates and estimated ones.

3.2.1. Accuracy Ratio To evaluate the discriminatory power of risk ranking among companies for our model, we employ the cumulative accuracy profile (CAP) and its associated accuracy ratio (AR). The accuracy ratio (AR) is defined as the ratio of the area between model CAP and random CAP, which can summarize the cumulative accuracy profile (CAP) in a quantitative way. For a good model, its AR would range from zero to one. The higher the score is, the better the model. Additionally, there exists a relation between the AR and the area under the receiver operating characteristic (ROC) curve: $\mathrm{AR} = 2\mathrm{AUC} - 1$.[4]

3.2.2. R-square(compared with FIM) Except for the power of risk ranking, the matching ability of a model between actual default rates and estimated ones is also important. For this purpose, we followed Duan et al. (2012) in employing the convolution-based default aggregation algorithm(Duan (2010)) to estimate the default rates from the predicted probability for a given prediction horizon (1, 3, . . . , or 60 months). Specifically, at each month-end, we obtain the predicted default rates for all active firms in this period for a prediction horizon. Then, we compare them

[4] Both CAP and ROC are commonly applied by banks and regulators to analyze the discriminatory ability of rating systems that evaluate credit risk Crosbie and Bohn (2003), Vassalou and Xing (2004).

with the observed default rates in the given prediction horizon. We repeat this process for all of our test samples and different prediction horizons. (Recall that Figure(c) plots the comparison for 1- and 60-month prediction horizons, where the bars depict the actual default rates and the lines correspond to the model estimations.) To measure the distance between these two default rates, the intuitive thought is to use R^2 (Eq. 15, where y_i is the default rate of an observation and $f_{i,m}$ is the estimated one of the model) with a trailing realized default rate over the same duration ($\bar{y}$ in Eq. 13), i.e., for k-period forward, we use k-period trailing realized default as a naïve prediction.

$$SS_{total} = \sum_{i=1} (y_i - \bar{y})^2. \tag{13}$$

$$SS_{res} = \sum_{i=1} (y_i - f_{i,m})^2 = \sum_{i=1} e_i^2 \tag{14}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}. \tag{15}$$

However, the above practice would shorten the number of data points in the evaluation. Therefore, we use FIM as the benchmark to compute the R^2 (Eq. 17), i.e., 1 minus the ratio of the sum of squared residuals (Eq. 14) from one model over the same from FIM (Eq. 16, where $f_{i,FIM}$ is the estimated default rate by FIM). If the value of this metric is higher than 0, it means that our model performs better than FIM and vice versa.

$$SS_{FIM} = \sum_{i=1} (y_i - f_{i,FIM})^2 = \sum_{i=1} e_i^2 \tag{16}$$

$$R^2_{FIM} = 1 - \frac{SS_{res}}{SS_{FIM}}. \tag{17}$$

4. Empirical Results

In this section, we describe the details of our empirical results. First, we introduce how we conduct our cross-sectional experiment and overtime experiment. Second, we describe our results for these two experiments, including the value of each performance metric and the graph of the aggregate default rates. Third, we describe the results for different sub-groups of our dataset in overtime experiments.

4.1. Experiments

To evaluate our model, we used two different settings when splitting the data into training, validation, and testing sets. The first experimental setting is referred to as the "cross-sectional experiment," in which we mix 1.8 million monthly samples and separate them into training and testing

sets with a ratio of 9:1. For the training dataset, we further partition it with the 9 to 1 ratio (cross-sectionally) into the core training and validation sub-samples to decide the training epochs for our model. Finally, we would use the decided training epochs to train our model in the whole training sample. Note that in this setting, the data samples from different periods are mixed, which is a commonly adopted approach in literature to attest to model capacity and compare model performance.

The second setting is referred to as the "overtime experiment," for which we use an expanding window setting along the time axis to conduct the experiments. Note that this setting is a commonly used and practical setting for scenarios involving time effects. Specifically, the sample period for our dataset is Jan 1994 to Nov 2021, totaling about 28 years, and the training sample size is first set to 10 years and then uses only the data available at the time for estimation, which is Jan 1994 to Dec 2003. For this training sample, like what we do in the cross-sectional experiment, we would partition the dataset into sub-training and validation (also a 9:1 ratio) and then build our model. For every month over the next year, i.e., 2004, we would perform 1-month to 5-year predictions based on our model and record the results. Advance one year to Dec 2004, we re-train the model using the expanded dataset from Jan 1994 to Dec 2004. Again for every month over 2005, we perform 1-month to 5-year predictions and record the results. We would repeat this process all the way to the end. Finally, we can have aggregate out-of-sample predictions of 18 years (2004 to 2021) and then use them to evaluate the performance of our models. It is worth noting that for such a prediction task involving data across a very long time period, the data distributions are by nature extremely volatile across different time periods; therefore, the purpose of the overtime experimental setting is to evaluate the model's ability to react to new, incoming data, which is more correspond to the real-world application.

4.2. Cross-Sectional Results

Table 5 shows the quantitative results of the cross-sectional experiments for each prediction horizon in AR and R-square (compared with FIM), in which "red color" denotes that the model performs worse than FIM and "bold text" means that the model yields the best performance among all models. Besides, the improvement is between the metrics obtained from the best models and FIM. According to the table, firstly, there are no "red color" at all, which means that all the neural models outperform FIM in all prediction horizons. Specifically, the AR increases from 1.994% to 21.655% for the 60-month prediction horizon; the R-square(compared with FIM) from each prediction horizon ranges from 0.04 to 0.583, which is also commendable progress for corporate default prediction. This situation demonstrates the great potential of neural networks applied to the task of multiperiod default prediction for the settings of the cross-sectional experiment. Secondly,

Table 5 Results of cross-sectional experiments

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	95.443	93.337	91.178	86.746	86.192	76.925	69.649	64.687	60.070
MLP	96.144	94.317	92.538	89.174	88.693	81.771	75.783	70.794	66.418
GRU	**97.346**	**95.025**	**93.787**	**91.591**	**91.302**	**86.342**	**81.375**	**76.863**	**73.079**
DKT_GRU	97.330	94.912	93.364	90.645	90.311	84.844	79.678	74.807	70.666
Improvement (%)	1.994	1.809	2.861	5.585	5.928	12.241	16.837	18.822	21.655
Panel B	R-square (compared with FIM)								
MLP	0.037	0.059	0.096	0.176	0.193	0.280	0.354	0.320	0.273
GRU	0.025	**0.205**	**0.231**	**0.360**	**0.402**	**0.578**	0.579	0.479	0.431
DKT_GRU	**0.040**	0.177	0.223	0.332	0.379	0.553	**0.583**	**0.496**	**0.446**

GRU-based models perform better than MLP(Divernois (2020)) in the most cases, indicating that capturing economic dynamics is helpful for corporate default prediction. Thirdly, among the three neural models, GRU yields the best performance. The reason behind this is that when there is little difference between the label distribution of the training dataset and the one of the testing dataset, a more complicated model(GRU) would capture the relation between the firms' covariates and the default events better.

4.3. Overtime Results

The results of the overtime experiments are listed in Table 6. Noted that these results are evaluated in the aggregate out-of-sample predictions for 18 years. From this table, we observe that the MLP performs worse than FIM in most cases (15 over 18), showing that directly adding functional flexibility may not work in the overtime experiment. In addition, although the GRU performs better than the MLP and FIM in AR, it performs worse than MLP and FIM in R-square (compared with FIM). This situation tells us that capturing time dynamics without filtering noisy covariates seems not good enough. Compared with the cross-sectional experiment, the results in overtime experiments are quite different, indicating that a different distribution between training and testing datasets would significantly influence the performance of normal neural models. This difference also implies that the neural-based models without modification are not qualified to be applied in real-world applications. However, according to the results, you can see that our proposed DKT_GRU dominates the other models. More importantly, the results demonstrate that our DKT approach can regulate the model to gain better performance both for risk ranking and in terms of matching the aggregate default distribution for new incoming data, especially in long prediction horizons; for example, for the 60-month default prediction, the improvements on AR is 14.837% and the R-square (compared with FIM) is 0.757.

Table 6 Results of overtime experiments

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	93.538	92.191	90.040	86.383	85.619	76.410	68.086	60.356	53.915
MLP	93.445	92.195	89.856	85.830	85.000	74.169	65.814	58.851	52.765
GRU	94.268	93.143	91.515	88.667	88.018	78.472	70.856	64.483	59.294
DKT_GRU	**94.767**	**93.559**	**92.000**	**89.301**	**88.693**	**80.379**	**73.681**	**67.330**	**61.914**
Improvement (%)	1.314	1.483	2.177	3.378	3.591	5.193	8.218	11.556	14.837
Panel B	R-square (compared with FIM)								
MLP	0.110	0.123	−0.001	−0.046	−0.036	−0.101	−0.144	−0.092	0.053
GRU	−0.470	−0.486	−0.770	−0.594	−0.557	−0.475	−0.329	−0.243	−0.081
DKT_GRU	**0.156**	**0.315**	**0.279**	**0.160**	**0.155**	**0.098**	**0.370**	**0.554**	**0.757**

4.3.1. Aggregate default rates Figure 3 shows the aggregate default rates of FIM and our proposed DKT_GRU, which are related to how we calculate the R-square (compared with FIM). The blue bars in the figure indicate the actual default rates and the curves correspond to the estimation of different models. Due to how we split the data in overtime experiments, the result shown in the figure is concatenated from each testing year. For example, the default rates in the year 2004 of each subfigure come from the first testing fold; similarly, the ones in the year 2005 come from the second fold. For shorter horizons, the predicted default rates from the models match the observations quite well. Compared with FIM, DKT_GRU can provide a wider range of default rates, indicating that it can catch the peak and low of the observation better. However, as the prediction horizon increases, there is more difference between the predicted default rates and the observed ones, showing a worse performance. As you can see, some of the difference comes from the "jumps" for the predicted default rates. The reason for these jumps is that we expand our training dataset and retrain our models every year. Compared with the default rates predicted by FIM, the ones predicted by DKT_GRU are more stable, especially for the periods around 2004-2005 and 2010-2012. These observations suggest that our DKT_GRU can regulate the model to generate stable predictions and then make for a better estimator of future uncertainty.

4.3.2. Sub-samples analysis Except for the overall performance, we also evaluate the performance of our models for two sub-samples. One sub-sample contains only financial firms (SIC between 6000 and 6999) and the other comprises all the non-financial firms. The results for the financial firms can be seen in Table 7. From this table, we can observe that all models have worse accuracy ratios compared with the ones of the full sample, indicating that the default rates of financial companies are hard to predict. However, our DKT_GRU still performs well in this sub-sample, which achieves 27.982% improvement in AR and has a positive R-square (compared with FIM) in all prediction horizons.

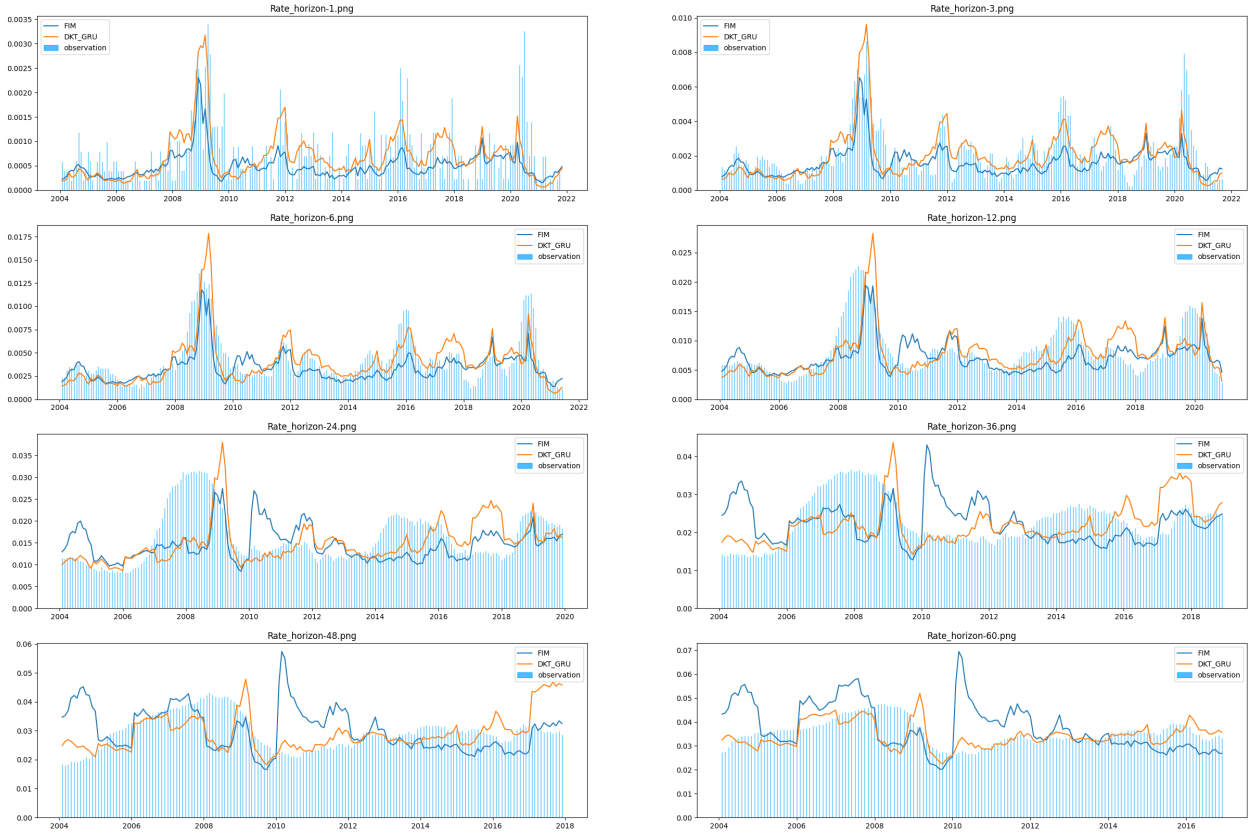

Figure 3 Aggregate default distributions

On the other hand, the performance of the non-financial firms can be seen in Table 8. From this table, we observe that the accuracy ratios for the non-financial sub-sample are similar to the ones of the full sample. However, the R-squares (compared with FIM) are quite different from the ones of the full sample in two parts. One is that MLP outperforms FIM in all prediction horizons. The other part is that although DKT_GRU performs the best in most cases, its performance is worse than MLP in the prediction horizons of 9, 12, and 24.

In summary, although our DKT_GRU generates the best performance in most cases for these two sub-samples, it still needs to be improved in some specific prediction horizons.

4.3.3. Time-Varying Performance Except for analyzing our models' performance for different industries, we further evaluate our models' time-varying performance. For this purpose, we group our test dataset by time into two sub-samples - high credit risk years and low credit risk years. We use the annual realized default rate as our grouping criteria. To be more specific, we

Table 7 **Results of overtime experiments for financial companies**

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	89.640	89.311	88.742	85.936	85.150	72.360	59.179	47.430	39.949
MLP	90.111	89.104	86.752	81.772	80.722	62.602	49.679	41.198	35.558
GRU	**90.741**	89.806	89.240	89.222	88.825	74.271	60.278	48.634	38.806
DKT_GRU	90.377	**89.904**	**89.528**	**89.415**	**89.116**	**78.685**	**69.092**	**59.183**	**51.127**
Improvement (%)	0.928	10.664	0.885	4.048	4.659	8.741	16.750	24.782	27.982
Panel B	R-square (compared with FIM)								
MLP	−0.020	−0.067	−0.155	−0.271	−0.272	−0.321	−0.324	−0.370	−0.465
GRU	−1.409	−2.171	−2.065	−1.515	−1.494	−1.365	−1.325	−1.481	−1.826
DKT_GRU	**0.016**	**0.109**	**0.215**	**0.253**	**0.239**	**0.175**	**0.195**	**0.123**	**0.001**

Table 8 **Results of overtime experiments for non-financial companies**

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	93.805	92.313	89.922	86.019	85.215	76.164	68.473	61.408	55.079
MLP	93.706	92.422	90.015	85.914	85.078	74.804	66.995	60.370	54.377
GRU	94.633	93.502	91.699	88.480	87.798	78.709	71.734	66.173	61.882
DKT_GRU	**95.110**	**93.848**	**92.088**	**88.997**	**88.343**	**80.054**	**73.727**	**68.147**	**63.132**
Improvement (%)	1.391	1.663	2.408	3.462	3.671	5.108	7.674	10.973	14.620
Panel B	R-square (compared with FIM)								
MLP	0.134	0.195	0.132	**0.136**	**0.142**	**0.104**	0.101	0.225	0.452
GRU	−0.112	−0.005	−0.207	−0.221	−0.218	−0.298	−0.110	0.078	0.425
DKT_GRU	**0.153**	**0.295**	**0.222**	0.110	0.102	−0.021	**0.302**	**0.499**	**0.681**

calculate the realized default rate for each year in the test dataset (2004 - 2021) and sort them into three parts from high to low: the top, the middle, and the bottom, where there are different 6 years in each part. We only use the top six years as the high credit risk years(2008, 2009, 2015, 2016, 2017, 2020)[5] and the bottom six years(2006, 2007, 2004, 2014, 2018, 2021) as the low credit risk years to analyze our model. It is worth noting that when we analyze our model in these sub-samples, we use them as our estimation time. For example, to analyze the model's performance in 2008, our model will use the information before 2008/12/31 to generate the default probabilities for all prediction horizons.

After grouping, we evaluate our models on these two sub-samples. In Table 9, we can observe the results for the test data in the low credit risk years. From this table, our DKT_GRU achieves the best performance in most cases, especially for the longer prediction horizons. For example, it

[5] 2008 and 2009 denote the period for the financial crisis, 2015-2017 denote the period for the prolonged energy stress and the economic crisis in other countries, and 2020 denotes the period for COVID-19

can outperform FIM 20% in accuracy ratio and obtain 0.898 R-square(compared with FIM) in the 60-month prediction horizon. Although DKT_GRU is not the best model for some prediction horizons, it still can achieve comparable performance. These results show that our DKT_GRU works well in the low credit risk years.

Table 9 Results of overtime experiments for low credit risk years

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	90.667	88.527	87.382	83.908	82.599	69.853	59.282	51.071	43.899
MLP	90.414	88.995	88.165	84.177	82.633	65.481	56.324	50.834	44.894
GRU	90.551	89.752	**89.840**	87.234	86.312	72.827	63.498	56.396	50.027
DKT_GRU	**91.405**	**89.785**	89.801	**87.458**	**86.433**	**74.373**	**66.578**	**59.592**	**52.805**
Improvement (%)	0.814	1.422	2.813	4.231	4.643	6.470	12.308	16.686	20.286
Panel B	R-square (compared with FIM)								
MLP	−0.315	−0.506	−0.610	−0.701	−0.692	−0.011	−0.105	−0.123	0.103
GRU	−0.225	−0.173	−0.150	0.022	0.045	0.023	−0.015	0.158	0.282
DKT_GRU	**−0.093**	**−0.045**	**−0.008**	**0.109**	**0.112**	**0.083**	**0.427**	**0.825**	**0.898**

On the other hand, we can see the models' performance for the high credit risk years in Table 10. Although DKT_GRU still obtains a good performance in most cases, it achieves worse in the longer prediction horizons compared to the results for the low credit risk years, especially in R-square(compared with FIM). However, FIM and MLP perform much better. The reason behind these results is that the predictions generated by the models are influenced by the current information they have in the high credit risk years. Given such information, the models can probably generate a higher cumulative default rate for all prediction horizons. The longer the prediction horizons, the large difference between the estimated cumulative default probabilities and the realized ones. Besides, the more complicated model is more influenced by this situation, such as GRU.

5. Conclusion

We have developed a domain knowledge-tailored neural network to address the task of multiperiod corporate prediction. Our method follows the FIM's framework and then can generate the forward intensities for default/bankruptcies and the other-exit events, enabling it to output a consistent cumulative default term structure. Depart from traditional neural networks, we leverage economic domain knowledge to tailor the networks, regulating the model to gain better performance and preventing it from overfitting. Experiments on a large real-world corporate default dataset over a lengthy period of time are used to demonstrate the efficacy of our proposed approach. The findings show that our model produces a performance that is noticeably better than the most

Table 10 Results of overtime experiments for high credit risk years

Horizons (months)	1	3	6	9	12	24	36	48	60
Panel A	Accuracy ratio (AR) (%)								
FIM	92.856	92.213	90.279	86.565	85.993	80.187	75.694	70.958	**68.698**
MLP	92.668	91.730	89.190	82.251	84.641	77.795	72.358	65.642	61.987
GRU	93.786	93.019	90.893	87.461	86.839	79.832	74.525	68.667	63.809
DKT_GRU	**94.089**	**93.390**	**91.358**	**88.223**	**87.729**	**81.724**	**76.848**	**71.027**	67.872
Improvement (%)	1.328	1.277	1.194	1.916	2.018	1.916	1.524	0.097	−1.289
Panel B	R-square (compared with FIM)								
MLP	0.218	0.288	0.214	0.128	**0.132**	**0.004**	**−0.014**	**−0.096**	**0.230**
GRU	−0.483	−0.390	−0.715	−0.777	−0.767	−1.155	−1.633	−1.795	−1.570
DKT_GRU	**0.230**	**0.406**	**0.362**	**0.149**	0.127	−0.089	−0.164	−0.338	0.177

advanced statistical model. Besides, it also displays that our DKT method can obtain a more robust performance than traditional neural networks. To further analyze our model, we also evaluate it on different sub-groups of our dataset. For financial firms, non-financial firms, and the low credit risk periods, our model performs very well in most cases. For high credit risk periods, our model's performance degrades in long-term prediction horizons in terms of R-square(compared with FIM) because the model is influenced by the information obtained by the estimation time. With more investigation, we believe that this problem can be solved in future research.

References

Hafiz Alakar, Lukumon O. Oyedele, Hakeem Owolabi, Vikas Kumar, Saheed Ajayi, Olúgbénga O. Akinadé, and Muhammad Bilal. Systematic Review of Bankruptcy Prediction Models. *Expert Systems With Applications*, 94:164–184, 2018.

Edward I. Altman. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, 23(4):589–609, 1968.

John Campbell, Jens Hilscher, and JAN SZILAGYI. In search of distress risk. *Journal of Finance*, 63: 2899–2939, 02 2008. doi: 10.2139/ssrn.770805.

Credit Research Initiative. NUS Credit Research Initiative Technical Report. Technical report, Credit Research Initiative, National University of Singapore, 07 2020. URL `https://d.rmicri.org/static/pdf/Technical%20report_2020.pdf`.

Peter Crosbie and Jeffrey Bohn. Modeling Default Risk. *Moody'S KMV White Paper*, 2003.

Marc Divernois. A deep learning approach to estimate forward default intensities. *SSRN Electronic Journal*, 01 2020. doi: 10.2139/ssrn.3657019.

Jin-Chuan Duan. Clustered Defaults. *National University of Singapore Working Paper*, 2010.

Jin-Chuan Duan, Jie Sun, and Tao Wang. Multiperiod Corporate Default Prediction—A forward Intensity Approach. *Journal of Econometrics*, 170(1)(1):191–209, 2012.

Darrell Duffie, Leandro Saita, and Ke Wang. Multi-Period Corporate Default Prediction with Stochastic Covariates. *Journal of Financial Economics*, 83(3)(3):635–665, 2007.

Eom, Haneul, Jaeseong Kim, and Sangok Choi. Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble model. *Journal of Intelligence and Information Systems*, 26(2):105–129, 06 2020. doi: doi:10.13088/JIIS.2020.26.2.105.

Zan Huang, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study. *Decision Support Systems*, 37(4):543–558, 2004.

David Lando. *Credit Risk Modeling*, pages 787–798. 2009.

Jim Lee. Does size matter in firm performance? evidence from us public firms. *international Journal of the economics of Business*, 16(2):189–203, 2009.

Wei-Lun Luo, Yu-Ming Lu, Jheng-Hong Yang, Jin-Chuan Duan, and Chuan-Ju Wang. Multiperiod corporate default prediction through neural parametric family learning. In *SIAM International Conference on Data Mining (SDM22)*, pages 316–324, 01 2022. ISBN 978-1-61197-717-2. doi: 10.1137/1.9781611977172.36.

Robert C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, 29(2):449–470, 1974. ISSN 00221082, 15406261. URL `http://www.jstor.org/stable/2978814`.

Jinlan Ni, Wikil Kwak, Xiaoyan Cheng, and Guan Gong. The determinants of bankruptcy for chinese firms. *Review of Pacific Basin Financial Markets and Policies*, 17(02):1450012, 2014.

Ja Ohlson. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980.

Bernardete Ribeiro, Catarina Silva, Ning Chen, Armando Vieira, and João Carvalho das Neves. Enhanced Default Risk Models with SVM+. *Expert Systems With Applications*, 39(11):10140–10152, 2012.

Bruno Dore Rodrigues and Maxwell J Stevenson. Takeover prediction using forecast combinations. *International Journal of Forecasting*, 29(4):628–641, 2013.

Suproteem K. Sarkar, Kojin Oshiba, Daniel Giebisch, and Yaron Singer. Robust Classification of Financial Risk. *arXiv preprint arXiv:1811.11079*, 2018.

Tyler Shumway. Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*, 74(1):101–124, 2001. ISSN 00219398, 15375374. URL `http://www.jstor.org/stable/10.1086/209665`.

Justin Sirignano, Apaar Sadhwani, and Kay Giesecke. Deep Learning for Mortgage Risk. *arXiv preprint arXiv:1607.02470*, 2016.

Maria Vassalou and Yuhang Xing. Default Risk in Equity Returns. *Journal of Finance*, 59(2):831–868, 2004.

Chi Xie, Changqing Luo, and Xiang Yu. Financial distress prediction based on svm and mda methods: the case of chinese listed companies. *Quality & Quantity*, 45(3):671–686, 2011.

Shu-Hao Yeh, Chuan-Ju Wang, and Ming-Feng Tsai. Corporate Default Prediction via Deep Learning. 01 2014.

Shu-Hao Yeh, Chuan-Ju Wang, and Ming-Feng Tsai. Deep belief networks for predicting corporate defaults. In *2015 24th Wireless and Optical Communication Conference (WOCC)*, pages 159–163, 2015. doi: 10.1109/WOCC.2015.7346197.

Ling Zhang, Shou Chen, and Xin Zhang. Financial distress early warning based on mda and ann technique. *Systems Engineering*, 11:50–58, 2005.

Me Zmijewski. Methodological Issues Related To the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, 22:59–82, 1984.