

Multi-Modal Deep Learning for Credit Rating Prediction Using Text and Numerical Data Streams *

Mahsa Tavakoli¹, Rohitash Chandra², Fengrui Tian¹, and Cristián Bravo¹

¹Department of Statistical and Actuarial Science, University of Western Ontario, London, Ontario, Canada, N6A 3K7.

²Transitional Artificial Intelligence Research Group, School of Mathematics and Statistics, University of New South Wales, Sydney, Australia.

Abstract

Knowing which factors are significant in credit rating assignment leads to better decision-making. However, the focus of the literature thus far has been mostly on structured data, and fewer studies have addressed unstructured or multi-modal datasets. In this paper, we present an analysis of the most effective architectures for the fusion of deep learning models for the prediction of company credit rating classes, by using structured and unstructured datasets of different types. In these models, we tested different combinations of fusion strategies with different deep learning models, including CNN, LSTM, GRU, and BERT. We studied data fusion strategies in terms of level (including early and intermediate fusion) and techniques (including concatenation and cross-attention). Our results show that a CNN-based multi-modal model with two fusion strategies outperformed other multi-modal techniques. In addition, by comparing simple architectures with more complex ones, we found that more sophisticated deep learning models do not necessarily produce the highest performance; however, if attention-based models are producing the best results, cross-attention is necessary as a fusion strategy. Finally, our comparison of rating agencies on short-, medium-, and long-term performance shows that Moody's credit ratings outperform those of other agencies like Standard & Poor's and Fitch Ratings.

*NOTICE: This is a preprint of a published work. Changes resulting from the publishing process, such as editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this version of the document. This work has been submitted for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible. This work is made available under a Creative Commons BY-NC-ND license. 

Keywords—

Fusion strategies, deep learning, credit ratings, multi-modality, BERT, CNN, cross-attention, earning call transcripts.

1 Introduction

The issue of credit rating prediction has received interest from many researchers due to its profit potential in the financial market Löffler (2020), including portfolio risk management Yijun et al. (2009). This interest has been driven by the importance of credit scoring for accurate assessment of credit risk, which is crucial for banks and other financial institutions to make informed decisions regarding lending money to corporations He et al. (2010). In addition, credit scoring is important for predicting the probability of default of a corporate borrower, which is critical to properly assess risk for what are usually large loans Liu et al. (2022). Related studies have focused on two major approaches, namely linear and non-linear models. Linear approaches have mostly involved regression models Orgler (1971), in which, for instance, linear regression has been applied for constructing a scorecard for outstanding loan evaluation, and logistic regression has been extensively used for credit scoring Thomas et al. (2017). In addition to these traditional approaches, machine learning models such as k-nearest neighbor, decision trees, random forests, support vector machines, and multi-layer perceptrons (MLPs, or simple neural networks) have been employed for credit rating, as well as in other domains in financial applications Zhu (2021).

Researchers have also explored the potential of integrating different models together to form hybrid models, which are heavily applied in financial scenarios like credit rating Jain and Kumar (2007). Complementing the above, deep learning models LeCun et al. (2015) are prominent machine learning models that have the capacity to work with more complex data, and have shown great potential in multimedia applications such as image processing and speech recognition Deng (2014) with applications in finance Ozbayoglu et al. (2020) and other areas Cai et al. (2020). Convolutional Neural Networks (CNN) O’Shea and Nash (2015) are deep learning models suited for image processing tasks Khodaei et al. (2022), but are also used for time series prediction Chandra et al. (2021). It has been shown that CNN outperforms other machine learning techniques including Multi-Layer Perceptron (MLP), support vector machine, random forests, and XGBoost for credit rating prediction Feng et al. (2020). Recurrent Neural Networks (RNNs) Medsker and Jain (2001)

are deep learning models suited for modelling temporal sequences Zhang et al. (2018) and natural language processing (NLP) Chowdhary (2020). Prominent RNN models are Long Short-Term Memory (LSTM) Hochreiter and Schmidhuber (1997) networks, Gated Recurrent Unit (GRU) Cho et al. (2014), and Transformers Vaswani et al. (2017), that are based on novel LSTM models equipped with attention mechanism based on cognition. LSTM and GRU have properties of remembering information over long periods of time overcoming the problems of vanishing gradient of canonical RNNs. Shen et al. Shen et al. (2018) applied GRU and its improved version to predict trading signals of stocks part of the Hang Seng and the S&P 500 indices, and compared them with support vector machines and showed that the GRU obtains better performance. NLP tasks such as language (machine) translation Nguyen et al. (2020), text classification Cambria and White (2014) and sentiment analysis Chandra and Krishna (2021) involves processing raw text that features unstructured information using large deep learning models. One of these models is Bidirectional Encoder Representations from Transformers (BERT) Devlin et al. (2018) which was developed by Google and has been used as a state-of-art from a number of NLP tasks including those mentioned earlier. The BERT model overcomes the limitation of unidirectional language models, and outperforms the LSTM-based model in natural language processing tasks like machine translation and speech-related applications Karita et al. (2019).

Deep learning models have been prominent in financial applications Ozbayoglu et al. (2020) such as financial time series forecasting Sezer et al. (2020). We have found that the majority of the applications have utilized structured datasets, including time series and tabular data with numerical features Wang et al. (2015). However, studies in which unstructured data such as text has been applied for financial applications have also been performed Li et al. (2020a).

In addition to employing unstructured data, some researchers have also presented models for applying both structured and unstructured data, leading to the presentation of multimodal techniques and application to scenarios such as credit rating Stevenson et al. (2021). In models for unstructured data, mixing information from various channels is known as a *multimodal fusion strategy*, referring to the process of combining information by integrating different channels D’Ulizia (2009). A review by Soujanya et al. Poria et al. (2017) indicates that multimodal models are significantly more accurate than unimodal models. Apart from the type of fusion, the fusion level (fusion order) is another topic in multimodal research. There are three approaches in terms of fusion level: early (or signal) fusion, in which raw channels or modalities are combined at the

earliest stage of the pipeline; the intermediate fusion approach, in which the modalities are mixed together after feature extraction and before classification; and the late or decision level, in which the outputs are combined for the classification decision Zhu et al. (2015). Boulahia et al. Boulahia et al. (2021) studied the impact of fusion level in the multimodal recognition process and found that the intermediate level fusion had the strongest impact compared to early and late fusion.

Lee et al. Lee and Yoo (2020) performed a study of multimodal deep learning for stock market prediction and found that accuracy improved using multimodal deep learning. They concluded that intermediate fusion led to better performance compared to late fusion. Luo et al. Luo and Kay (1988) have shown that multimodal models can mix different types of modalities, often using a concatenation technique as a simple fusion strategy.

To the best of our knowledge, no research has been performed regarding the prediction of credit rating classes using another type of fusion for combining different modalities. An example of a more modern type of fusion technique involves cross-attention as the core fusion strategy Huang et al. (2019); Hou et al. (2019); Lee et al. (2018). In recent years, cross-attention has been used to replace concatenation and improve fusion performance over multi-modal data. Cross-attention is an attention-based mechanism that helps to identify the correlation and interaction of the different modalities of data when combining them Meng et al. (2020). Few studies appear to have investigated the effect of unstructured or a combination of unstructured and structured data on credit rating prediction. Moreover, due to the complexity of the analytical process for unstructured data in comparison to structured data, only recent research has investigated its impact on many applications including financial prediction Pejić Bach et al. (2019).

In this paper, we present a study of the fusion of deep learning models for the prediction of company credit rating classes using structured and unstructured datasets of different types. The structured datasets include market, bond, financial ratios, and covariate (agency types and past rating in each class) information, while the unstructured dataset contains text documents of earnings call transcripts. The earnings call is a periodic conference led by the executives of companies to compile company performance and present it to the market Frankel et al. (2017).

We investigated the effects of different fusion strategies, including concatenation and cross-attention, on prediction performance. Besides the fusion type, we were interested in the impact of the order (level) of fusion in a model on performance. Different combinations of fusion type (concatenation and cross-attention) and order (early or intermediate), could lead to four possi-

ble modes for the model's structure. Each possible mode could also be designed based on four commonly used deep learning models (CNN, LSTM, GRU and BERT). Thus, we compared 16 models that had been divided into four main categories in terms of fusion type and fusion level. Moreover, by activating and inactivating each channel in the best model, the value of each channel of structured data and text could be clearly quantified, further helping to evaluate the usefulness of unstructured data. We have also provided a comparison of the impact of the different rating agencies on the short-, medium-, and long-term performance of the model.

Although we performed a comprehensive evaluation of the respective models, because our data period was related to the time of COVID-19 pandemic, it was also worth evaluating the models by considering fluctuations in the market during the pandemic period. The COVID-19 pandemic had a drastic effect on the economy and financial markets and caused a disruption in international trade Brodeur et al. (2021). Zeren et al. Zeren and Hizarci (2020) have indicated that investing in the stock market may not have been the correct option for investors because of the emerging new and unpredictable situation during the peak of the pandemic. Due to the importance of this subject, we needed to investigate the performance of our proposed model during the COVID-19 pandemic and compare the performance to the situation prior to the pandemic. Hence, we attempted to answer the following research questions:

1. Which combination of fusion type or order and deep learning model (CNN, LSTM, GRU, or BERT) can provide better performance for credit rating prediction?
2. What is the contribution of each modality separately, and which channel plays the most important role?
3. How reliable is the model during an unexpected crisis such as the COVID-19 pandemic?
4. What is the impact of the different rating agencies on short-, medium-, and long-term statistical performance when using the best model?

The remainder of this report is presented as follows: Section 2 provides a literature review of related work. Section 3 presents the methodology for structuring different multimodal models. Section 4 presents a complete description of the dataset and visualizes the text modality in detail. Section 5 illustrates model comparisons and discusses the results.

2 Related Works

2.1 Deep learning for finance problems

In recent years, deep learning models have been applied to many financial prediction problems based on structured data, including early delinquency Chen et al. (2021), profit scoring Fitzpatrick and Mues (2021), credit scoring Gunnarsson et al. (2021); Kvamme et al. (2018), price movement prediction Jabeur et al. (2021), trader classification Kim et al. (2020), fraud detection Seera et al. (2021), and others. In a survey, Huang et al. Huang et al. (2020) compared a wide range of models in the application of finance and banking domains including stock market prediction, banking default risk, and credit rating. They considered deep models including RNN, CNN, MLP, and machine learning models and found that CNN and LSTM-based models had outperformed others in most cases. Neagop et al. Neagoe et al. (2018) used CNN and MLP for credit scoring and found that CNN led to better performance when compared to MLP. Golbayani et al. Golbayani et al. (2020) also found that LSTM models consistently outperform other deep learning architectures, including CNNs, in predicting corporate credit ratings issued by the S&P rating agency. Moreover, Qian et al. Qian et al. (2023) proposed a new CNN architecture with a soft reordering mechanism that could adaptively reorganize tabular data for better CNN learning. Experiments showed that this approach could outperform traditional machine learning algorithms and other deep learning models for credit scoring, achieving higher accuracy and computational efficiency. In addition, Adisa et al. Adisa et al. (2022) explored the effectiveness of combining multiple classifiers into ensembles for credit scoring prediction and proposed an optimization approach using an LSTM deep learning algorithm with a genetic algorithm to determine optimal parameters. Their optimized LSTM model outperformed single classifiers and ensemble models. Besides the use of CNN and RNN-based models in the domain of credit risk, Korangi et al. Korangi et al. (2022) also applied a transformer-based model, transformer encoder for panel-data classification (TEP), for midcap companies to detect complex, non-linear relationships over a long range of time series; TEP produced better performance than traditional models.

In addition to credit applications, deep learning involves other financial predictions. For example, in bankruptcy prediction tasks, LSTM-based models have produced better forecasting performances using sequential data when compared to logistic regression, support vector machine, and random forests Kim et al. (2022). Additionally, Zhuang et al. Zhuang et al. (2022) used deep

learning techniques to predict price trends in the Chinese stock market. They proposed a method that combined the CNN and LSTM neural network models to achieve both forecasting accuracy and risk control. The combination of both models yielded better investment returns than the individual models. Vidya et al. Vidya and Hari (2020) addressed the implementation of LSTM networks for gold price prediction and demonstrated that their proposed LSTM model surpassed traditional methods, including autoregressive integrated moving average, deep regression, support vector regression, and CNN. In addition, Aryal et al. Aryal et al. (2019) investigated the application of deep learning models such as LSTM and CNN in forecasting the exchange rate between the United States Dollar (USD) and the Sri Lankan Rupee (LKR) and found that the CNN model exhibited the highest level of accuracy among the models tested. As a result, the authors have suggested that CNN-based models would be the most suitable for financial time series prediction, owing to their ability to extract relevant features and detect non-linear dependencies. Lin et al. Lin et al. (2021) have proposed a new model called CNN-LS, which combines CNN with LSTM to predict the price of six common stock indices. The model uses two paths of CNN and one path of LSTM to extract features. An experiment using 10 years of historic data showed that CNN-LS outperformed other methods in terms of Mean Squared Error(MSE) and Mean Absolute Error (MAE).

2.2 NLP for finance problems

As outlined above, the datasets commonly used in the credit scoring community are structured in a format featuring a mixture of numerical and categorical attributes Wang et al. (2015). Unstructured data such as raw text Gunnarsson et al. (2021) also has great potential in the credit rating domain; however, it has been explored by very few researchers Shi et al. (2018). Hence, we examined studies in which unstructured data had been used for prediction and classification. As an example, Lee et al. Lee et al. (2014) analyzed the importance of text in the prediction of stock price movement by building a corpus that facilitated the forecasting tasks. They showed that the incorporation of text in the prediction task could improve results significantly. Goldberg et al. Goldberg (2016) has found that CNNs can be particularly beneficial when the useful information is sparse and dispersed in different places in the data, which is typically the case with textual data. Such a claim is further supported by the results of Mai et al. Mai et al. (2019), who applied CNN to textual data with accompanying accounting data and found that CNN could generate

more accurate predictions for corporate bankruptcy. RNN, with the property of modeling complex temporal characteristics, better captured the context information in text processing compared to CNN Vargas et al. (2017). Li et al. Li et al. (2020a) evaluated fundamental data and online news and have proposed an LSTM-based model for stock prediction to address the issues caused by interactions among different modes and heterogeneity of the data.

In recent years, approaches to NLP problems have seen significant improvement using deep learning methods, particularly those incorporating attention mechanisms and transformer models Vaswani et al. (2017). These state-of-the-art models address the problem of long-term dependencies via LSTM and utilize the idea of attention from biological and cognitive systems, in which specific words in a sentence receive more attention than others Vaswani et al. (2017). Transformer-based models feature encoder–decoder LSTM models in which an attention mechanism is utilized for processing sequential data, such as natural language text or speech; several studies have used them for credit risk application Dorfleitner et al. (2016); Chen et al. (2018); Netzer et al. (2019). Kriebel et al. Kriebel and Stitz (2022) evaluated transformer-based deep learning models for text extraction on credit-relevant information and compared them to traditional machine learning models. They found that the performance of these deep learning models was superior to other machine learning models in most cases. In addition, the authors noted that the importance of textual data for credit default prediction was reinforced by the significant decrease in the accuracy of the model when the textual component of the data had been removed. Similarly, regarding the question of whether textual information can provide significant benefits in financial applications, Mai et al. Mai et al. (2019) and Matin et al. Matin et al. (2019) have concluded that the use of textual disclosure and segments is promising for the prediction of financial distress and bankruptcy. However, using different datasets, Dorfleitner et al. Dorfleitner et al. (2016) and Chen et al. Chen et al. (2018) found no clear evidence that text characteristics could predict credit default in peer-to-peer lending; this result contrasts that of Matin et al. Matin et al. (2019), who improved bankruptcy prediction by adding textual information related to the firm’s annual reports. Finally, Stevenson et al. Stevenson et al. (2021) applied the BERT model to predict loan default by small companies using textual loan assessments provided by lenders. They concluded that although the text data was useful for prediction when being assessed alone, the value of text data remained unclear because addition of text data to the already existing structured data produced no improvement in aggregate performance. Together, these studies demonstrate that the addition

of text data to deep learning models can lead to better performance, but these improvements have been context-specific, supporting the need for more detailed measurement in various cases, such as the present study related to corporate credit ratings.

3 Methodology

3.1 Backgroung and motivation

3.1.1 CNN and RNN

Initially developed for computer vision tasks LeCun et al. (1995), the typical architecture of CNNs includes an input layer, usually several convolutional layers, post-processing layers such as pooling layers or batch-normalization layers between the convolutional layers, and fully connected layers accompanied by regularization layers such as dropout operations. CNNs have been applied in a variety of areas such as image and pattern recognition Dhruv and Naskar (2020), NLP Widiastuti (2019), speech recognition Musaev et al. (2019) , and others.

In simple (canonical) RNNs Elman (1990); Werbos (1990), which are alternatives to map-based CNNs, the sequence of data is processed by iterating each element within the sequence and a state (context layer) is used to store the information; hence, the output of an RNN is decided by both its current input and the previous state through recurrence. However, in a long sequence of data, the gradient of RNN can become either too small or too large when training using backpropagation, thus causing problems related to gradient vanishing or exploding Hochreiter (1998) and preventing the model from learning useful information. To address the problem, LSTM networks Hochreiter and Schmidhuber (1997) feature better memory capability through the use of memory cells. An LSTM cell has three gates, including a forget gate, an input gate, and an output gate. The forget gate uses the output from the sigmoid function, which is used to decide how much of the information will be passed through depending on the priority and the usefulness of the information for prediction. LSTM network models have been prominent in time series prediction Ahmed et al. (2022), language modeling tasks such as sentiment analysis and language translation Wang and Cho (2015) before transformer architectures became dominant (see below), handwriting recognition Staudemeyer and Morris (2019), and speech synthesis Yang et al. (2020).

GRUs are special implementations of LSTM networks with simpler architectures, featuring a

reset gate and update gate Cho et al. (2014). The reset gate decides the percentage of previous memory and current input to load into the new memory, and the update gate controls how much of the previous memory to save. Due to a simpler model structure and fewer parameters, GRUs are faster to train when compared to LSTM networks Yang et al. (2020). On the other hand, if a sufficient amount of data is available, LSTM networks may perform better. Applications of GRUs include time series predictions Yamak et al. (2019), image recognition Li et al. (2021), and NLP Cascianelli et al. (2018).

3.1.2 Transformers and BERT

Transformer-based models that feature attention mechanisms Vaswani et al. (2017) have been commonly used in NLP tasks such as translation Nguyen et al. (2020) and question-answering Namazifar et al. (2021). The transformer model is based on an encoder–decoder LSTM network with an attention mechanism that is much greater when compared to conventional LSTM networks in terms of the number of trainable parameters.

In self-attention networks Humphreys and Sui (2016), the model takes input as a sequence of embeddings through input embedding functions. Three abstract vectors of query, key, and value from the same sequence of data are used to connect the encoder and the decoder. Query and key are matched to calculate relationships to each other, while the value vector represents the information to be extracted from the input embeddings. In the later stage, these three vectors are packed correspondingly into matrices Q , K , and V to allow simultaneous computation of the attention function (Equation 1) to generate the output as an attention score Vaswani et al. (2017):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k represents the key dimensionality and T represents the transpose operation applied to the key matrix K . The structure of a self-attention network ensures that within a feature sequence, each data point can be attended to by the other data points Korangi et al. (2022), which helps to analyze the context across the input rather than encoder-decoder order Kriebel and Stitz (2022).

The multiple heads of attention can project the features to different high-dimensional subspaces to extract data correlations and interactions from different perspectives. The multi-head attention mechanism generates multiple representations of Q , K , and V , which then computes their scaled dot-product attention. The process involves combining the results through concatenation

and then using a linear layer to scale the output to the desired dimension. There are numerous applications for transformers, including NLP Wolf et al. (2019), computer vision Khan et al. (2022), time series Wen et al. (2022), and others.

The BERT model Devlin et al. (2018) is a transformer-based model that learns bidirectional representations from both the left and the right context of a word within a sentence to overcome the limitation of unidirectional language models. Google has made available a BERT model pre-trained on a large corpus that includes the Wikipedia English dataset along with the BookCorpus, which encompasses over 11,000 books, complemented with Common Crawl to create the final corpus that was used to train the model Devlin et al. (2018). Due to its state-of-the-art performance in NLP tasks such as corpus classification Devlin et al. (2018) and text classification Sun et al. (2019), we included BERT-base, with all its trainable parameters frozen, as one of the models for processing the textual channel of the credit rating data in our proposed framework.

In this study, we employed an attentive CNN, which combines a CNN model for numerical channels with an added attention module in its architecture. This approach can improve the prediction result in classification applications Neumann and Vu (2017).

3.1.3 Information Fusion Strategies

In addition to applying deep learning models to structured and unstructured data, we further investigated the impact of different information fusion strategies in our framework. White et al. White (1991) have defined information fusion as a process that approaches the association, correlation, and combination of data along with information from single and multiple sources.

Fusion techniques refer to the strategies used to combine information from multiple modalities, such as concatenation of information from multiple streams of data. Fusion level, on the other hand, refers to the stage in the pipeline at which the information is merged. Figure 1 shows the three main fusion levels: early fusion, late fusion, and intermediate fusion. The early or signal fusion level combines the raw data directly, preserving as much information as possible and leaving the filtering responsibility fully to the model. Therefore, it has the potential to produce the lowest bias. However, such fusion is limited to data with the same features and patterns and has lower efficiency Luo and Kay (1988), as it comes with a higher variance. A higher-level fusion is intermediate level fusion, in which the process of merging information from different deep learning layers is conducted after it has been first filtered by previous layers Boulahia et al. (2021). The

third and highest level of fusion is late (or decision) level data fusion. At this stage, before being fused together, the information has already been extracted as abstract features or even in a format of decision or prediction generated as the output from previous models. The outputs of late fusion are often preliminary classifications of the tasks Luo and Kay (1988). In this study, we used early and intermediate levels because we were interested how much preprocessing by preceding layers would contribute to distilling knowledge in the context of our problem.

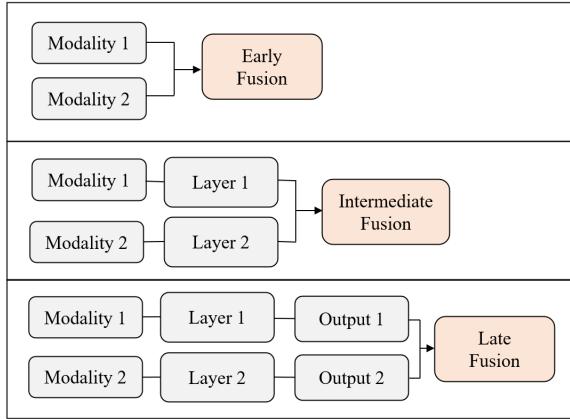


Figure 1: Fusion level strategies for two modalities, including early level, intermediate level, and late level. Only the first two were explored in this study, as we desired to analyze the impact of filtering as each channel moved through the model.

In our framework, we compared different fusion strategies based on various fusion levels and techniques. In terms of technique, we compared the impacts of concatenation and cross-attention fusion, which have been classified as simple operation-based fusion and attention-whiteZhang et al. (2020). As the simplest and the most commonly used fusion method, concatenation is capable of fusing not only lower-level raw data by joining tables of structured data or stacking the data together but also higher-level features by concatenating the features in a format of the output matrix in a certain layer of the neural network. In recent years, the cross-attention mechanism Lee et al. (2018) has been used to replace concatenation and improve fusion performance over multi-modal data. Cross-attention is similar to self-attention Shaw et al. (2018); the only difference is that cross-attention uses input from two different sequences or modalities, with the query and value vectors generated from one modality and the key vector from the other. Such a mechanism is helpful in identifying the correlation and interaction of different modalities of data when combining them.

The combination of fusion strategies in terms of level and technique can be effectively utilized in various scenarios, including 1) signal level fusion by concatenation to fuse different channels of

structured data, 2) intermediate level fusion by concatenation to fuse the layers of different channels of structured and unstructured data, and 3) intermediate level fusion using cross-attention to fuse the layers of structured data and unstructured text data.

3.2 Multi-modal Strategies

We developed a framework that featured appropriate data structures and techniques to address the first research question by investigating which combination of fusion strategies and deep learning models could capture more useful information from the data for model prediction. Hence, we investigated three areas, namely 1) deep learning models, 2) fusion-type strategies, and 3) fusion-level techniques to identify the best combination for our purposes. As mentioned above, the fusion-type strategies included concatenation and cross-attention, and the fusion-level techniques involved early level and intermediate level fusion, which we explain further below. Although applying a fusion-type strategy allowed evaluation of the power of cross-attention versus concatenation, the fusion-level techniques also indicated whether the fusion of data in the early steps of the model (signal fusion) had led to better performance than fusing the channels after the network had applied filters to the data (intermediate fusion). We used CNN, LSTM, GRU, and BERT deep learning models, which we fed into the fusion framework shown in Figure 2.

In each category, we analyzed four deep learning models, which could be combined with four different fusion strategies (Figure 2), thus generating 16 multi-modal models. The five channels of inputs included three semistructured channels (bond, market, financial ratios), one structured channel (covariate, which was previous rating information) and one text channel in each category. Network “A” in Figure 2 represents the network for numerical channels, and Network “B” refers to a network for the text channel. After fusing all modalities (intermediate fusion), we applied a sequence of layers leading the model head. The outputs from the models before the head were flattened, and a sequence of two dense layers with dropout were added. Because these layers were identical for all groups, we have omitted them from Figure 2. In the following text, we explain the fusion characteristics of each group and then illustrate four proposed architectures used in each group.

- GROUP 1: As shown in Figure 2a, signal fusion was used for the structured dataset. Hence, all numerical datasets were concatenated at the beginning, serving as inputs for Network A. On the other hand, the text was trained using Network B. Then intermediate fusion was

performed, and the penultimate layer of different modalities was combined through concatenating the respective output layers.

- GROUP 2: Similar to GROUP 1, as indicated in Figure 2b, we fused all numerical datasets at the beginning (signal fusion) and then concatenated the numerical channel to enable Network A to be trained. The text channel was also sent to Network B, and we fused their output layers using cross-attention.
- GROUP 3: Shown in Figure 2c, each numerical dataset separately passed through network A, and the text dataset entered Network B. We then concatenated their penultimate layers, and only the intermediate level of fusion was used.
- GROUP 4: As indicated in Figure 2d, we first trained four numeric channels separately using Network A and trained the text channel by Network B. After training, we concatenated the penultimate layer of numerical modalities. In the last step, we fused the text layer and the concatenated layer via the cross-attention method.

In designing Network A and Network B, we used four well-known deep learning models, namely CNN, LSTM, GRU, and an attention-based model (attentive CNN for numeric data and BERT for text). (Table 1). Network A’s architecture was designed to process numerical data, while Network B was oriented towards text data (Figure2). According to Table 1, there were four possibilities for Networks A and Network B. The first configuration was a CNN that contained two hidden layers for Network A and three hidden layers for Network B. Between every layer, we added dropout-based regularization Srivastava et al. (2014) (Drop) and max pooling (MaxP) layers, and at the end layer, we used the global average pooling method (GlbAv). In the second configuration, we used LSTM instead to replace the last CNN. The resulting model was a CNN followed by an LSTM layer for Network A and two CNNs and one LSTM for Network B. In the third configuration, we replaced LSTM with GRU and used a CNN and GRU for Network A and two CNNs and GRU for Network B. The final configuration was based on the attention mechanism. Hence, we used CNN and attention layers (attentive CNN) in Network A and BERT for Network B.

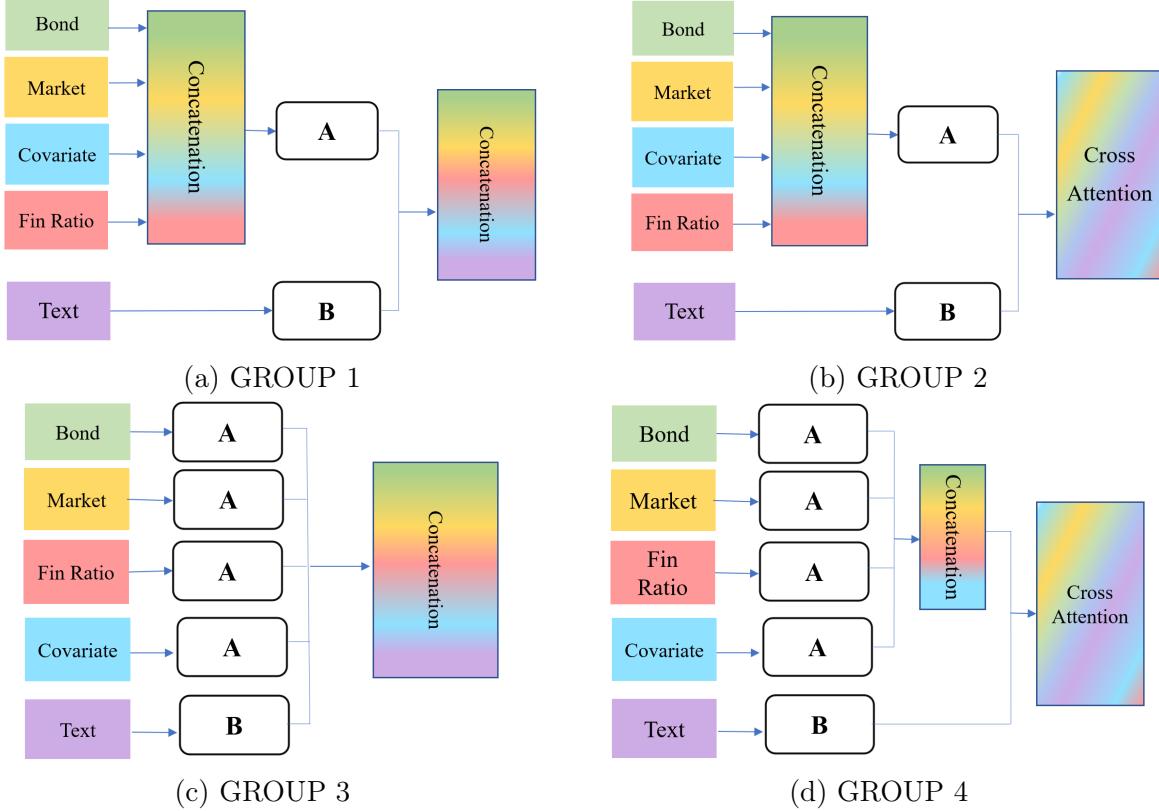


Figure 2: Proposed models, showing different variations in terms of fusion type and fusion level strategies. The labels "A" and "B" indicate two distinct network architectures, each of which corresponds to one of the options presented in Table 1.

As can be seen in Table 1, along with the base models of CNN, LSTM, GRU, and attention (ATT), we used convolutional (Conv) and dropout layers in all models. We examined the different combinations in trial experiments and found that adding one or two CNN and dropout layers led to better performance. The models' parameters such as batch size, learning rate, number of neurons, dropout factor, and number of CNNs in each layer were selected to enable the best performance for the validation set.

4 Experiment Design

4.1 Data

In this research, we used four sources of numerical and textual datasets, matched according to time index, and two identifiers, namely the company's symbol and CUSIP ID. CUSIP refers to the Committee on Uniform Securities Identification Procedures, a 9-digit combination of letters and

Based model	Network A	Network B
CNN	Conv→MaxP→ Conv→GlobAve	Conv→Drop→ Conv→MaxP→ Conv→GlobAve
LSTM	Conv→MaxP→ LSTM→GlobAve	Conv→Drop→ Conv→MaxP→ LSTM→GlobAve
GRU	Conv→MaxP→ GRU→GlobAve	Conv→Drop→ Conv→MaxP→ GRU→GlobAve
ATT	Conv→Drop→ ATT →GlobAve	BERT

Table 1: The architecture of Networks A and Network B for each base model.

numbers that serves as a unique identifier for securities such as stocks, bonds, and other financial products. The dataset contained 27,854 valid records, and the input variables were provided with a minimum lag of two months compared to the target timestamp. The data was split randomly, using 20% as the test set; the remaining data was split into 20% as a validation set and 80% as a training set.

4.1.1 Target Data

The importance of credit rating prediction lies in the crucial information it provides, enabling knowledgeable borrowing and investment decisions as well as market regulation and risk management. Several credit rating agencies exist, but the three largest and best known (Fitch Ratings, Moody’s, and S&P) control approximately 95% of the rating market White (2010). We used a mapping table in which different rating types had equivalence under the European Union Credit Rating Agency Regulation (CRAR) Jewell and Livingston (1999).

Table 2 links letter-based credit ratings with numerical values to provide a means of mapping creditworthiness onto a quantitative scale. Typically, credit rating agencies use a letter-based system to denote credit quality, employing combinations of letters such as AAA, AA, A, BBB, BB, and so on. The highest credit quality is symbolized by the AAA rating. By using the conversion table, users can accurately convert these letter-based ratings into a numerical scale,

which is valuable for performing quantitative analysis or modeling. Based on the table, there are 22 classes of rating, in which class 1 refers to the highest quality and class 22 refers to the defaulted companies. Because the amount of data in some classes was significantly lower than in the other classes (Figure 3), we merged the low-frequency classes (below 5%) with their closest classes. Table 2 indicates which classes were merged with each other.

Code	Moody's	Fitch and S&P	New class	Frequency
1	Aaa	AAA		
2	Aa1	AA+		
3	Aa2	AA	1	9%
4	Aa3	AA-		
5	A1	A+		
6	A2	A	2	9%
7	A3	A-	3	9%
8	Baa1	BBB+	4	13%
9	Baa2	BBB	5	15%
10	Baa3	BBB-		
11	Ba1	BB+	6	23%
12	Ba2	BB		
13	Ba3	BB-		
14	B1	B+	7	14%
15	B2	B		
16	B3	B-		
17		CCC+		
18	Caa	CCC		
19	Ca	CCC-	8	6%
20		CC		
21	C	C		
22	D	D		

Table 2: Table for conversion of letter-based credit ratings to numeric codes. The last two columns indicate the newly merged classes and their frequency (source for first three columns:Jewell and Livingston (1999))

4.1.2 Numerical Channels

In the case of the data that featured the numerical channels, we collected them from 2010 to 2022. The numerical data, which was taken from the source of CRSP/Compustat contained four

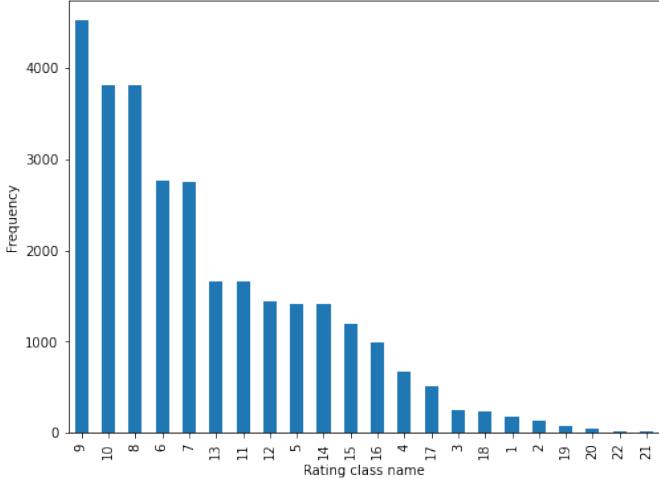


Figure 3: Frequency of ratings in the original dataset. The distribution shows that some classes have a disproportionately low number of ratings, which suggests that they may not contain sufficient information to support accurate classification.

channels, which have been provided in our GitHub repository¹:

1. Bond historical: This dataset had eight features related to the securities such as monthly high price, low price, trading volume, and other trading summaries.
2. Financial ratios: This data included 45 features related to the financial ratios of the companies including valuation, profitability, capitalization, liquidity, and efficiency ratios.
3. The market dataset contained 98 features related to monthly indexes built on market capitalization and portfolios on S&P index, stock market returns, and U.S. Treasury and inflation indicators.
4. The covariate features contained the rating agency type and the last observed rating class at that timestamp.

4.1.3 Text Data

We gathered text documents related to earnings call transcripts provided for each company by the Seeking Alpha website Alpha. Seeking Alpha specializes in financial news and analysis; it covers a wide range of publicly traded companies, stocks, and investments, with specific emphasis on those based in North America, particularly the US and Canada. One of its standout features is its platform for earnings call transcripts French-Marcelin (2020). An earnings call transcript is a

¹<https://github.com/Banking-Analytics-Lab/MultimodalFusionRatings>

written account of a company’s financial conference call with analysts and investors. It summarizes the company’s financial results and provides information regarding its performance, operations, and future plans. During the call, management provides a prepared statement, followed by a question and answer session with analysts and investors. The aim of the call is to offer a deeper understanding of the company’s financial situation and operations, and the transcript serves as a valuable resource for investors and analysts to access for a precise and complete record of the information shared Li et al. (2020b). We used this text as an input channel for our models. To pre-process the text, we fixed various Unicode errors; applied lowercase; replaced all URLs, email addresses and phone numbers with a special token; and finally removed punctuation and stop words. The text was then tokenized into words using the Keras tokenizer `ker` and the BERT encoder tokenizer `ber`, which was used in conjunction with the BERT model. In BERT tokenization, text is split into smaller units called tokens, which consist of words and subwords. Each token is assigned a unique integer identifier for the model to recognize. This approach enhances the model’s understanding of each word’s meaning and the relationships between words in a sentence. The tokenization process also incorporates special characters, such as punctuation, and adds specific tokens to mark the start and end of a sentence Lu et al. (2022).

We used the Keras embedding tokenizer for the other models (CNN, LSTM, and GRU). The Keras tokenizer performs tokenization by transforming words or subwords into integer tokens. It initially converts all text to lowercase then splits the text into separate words. Each word is then assigned a unique integer based on its frequency in the dataset. Once the text is tokenized, it can be used as input to a deep learning model for further analysis in NLP applications Chollet (2021).

5 Results

5.1 Text Visualization

Before training the models, we first explored the documents to understand the relationship of earning calls with risk. There were almost 27,000 documents, in which the frequency of the words varied from 50 to 20,000 words. Figure 4 represents the frequency of the words in the documents for each rating class on average. All classes had a similar word average of almost 4,000 words.

We split the documents by rating class for visualizing common words in the documents. However, for increased transparency, the visualization was performed in two groups rather than eight.

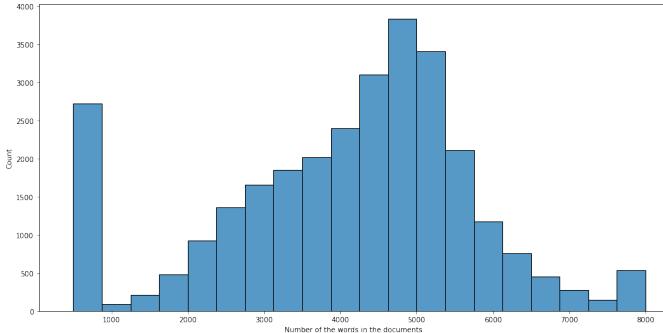
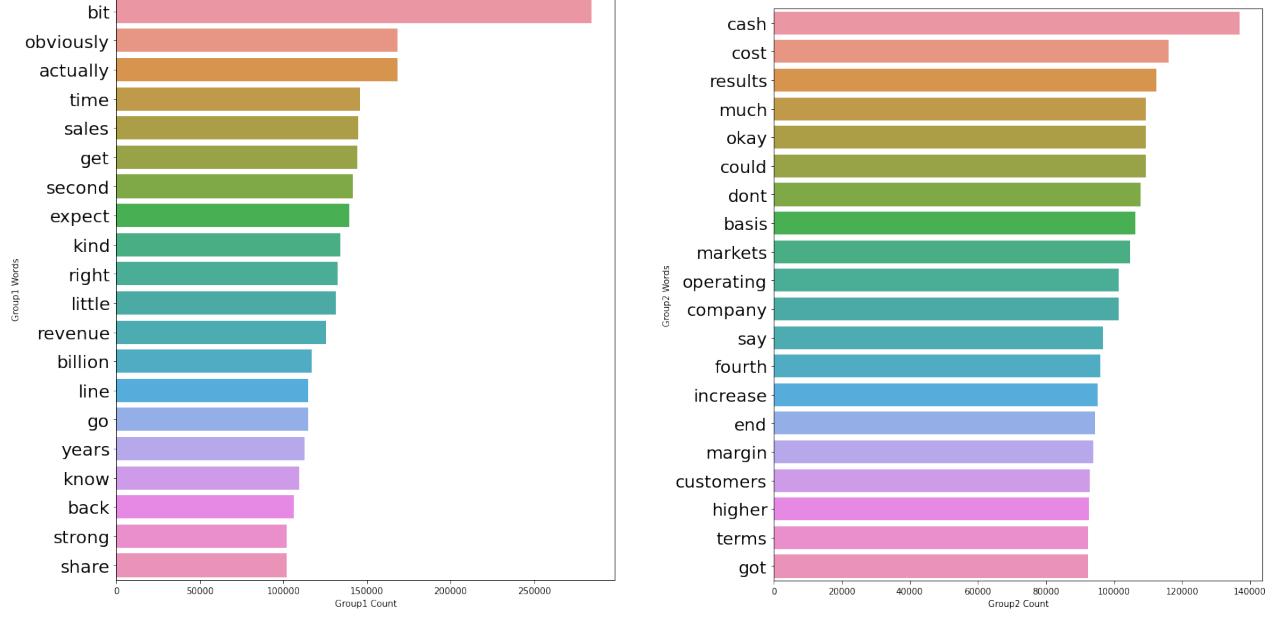


Figure 4: Frequency of the words in the documents.

Thus, we considered two rating groups: high ratings (ratings of 1 to 10, Group 1) and low ratings (ratings of 11 to 22, Group 2), meaning lower than median risk and higher than median risk, respectively. We used n-grams to better visualize the difference between the words in the high- and low-rank classes. An n-gram is a sequence of n consecutive elements taken from a given sample of text or speech Poliak et al. (2017); for example, the one-gram in Figure 5 shows the most frequent words in the high-rating group that did not include the most frequent ones in the low-rating group (and vice versa). As can be seen, the high-rating documents exhibited a wider range of commonly used words than the low-rating group, as evidenced by frequency count. This observation suggests that there may have been a relationship between the diversity of language use and the rating class, which could potentially have been leveraged for prediction purposes. Second, the specific types of words used in each group may also have been relevant. The presence of adverbs such as “obviously” and “actually” in the high-rating documents suggests a potentially positive attitude or confidence in the content being presented. On the other hand, the common use of negative verbs such as “don’t” in the low-rating group may indicate a more critical or negative attitude towards the content. Overall, these observations highlight the potential importance of language use and attitude in predicting the rating class of a document.

5.2 Model Prediction

We used two criteria to assess the performance and the statistical prediction power of each model, namely the weighted area under the receiver operating characteristic (ROC) curve (AUC) and the F-1 score. The AUC measures the entire area under the ROC curve and indicates the performance at different thresholds. Due to the scale-invariant and classification-threshold-invariant properties that it possesses Davis and Goadrich (2006), AUC is a desirable measure for classification per-



(a) Group 1

(b) Group 2

Figure 5: One-gram related to the most frequent words in the high-rank group (Group 1) that does not include the most frequent ones in the low-rank group and vice versa.

formance. Because our problem involved multi-class classification, we used a weighted average of AUCs for each class versus the rest of the classes (one-versus-all approach).

The F-1 score is an appropriate metric that captures a balance of precision and recall in the following way:

$$F - 1 = \frac{2}{recall^{-1} + precision^{-1}} = \frac{2TP}{2TP + FP + FN} \quad (2)$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

Table 3 indicates the performances of the respective models and their groups in terms of AUC and F-1 score, with a 90% bootstrap confidence interval (CI). In this study, we used 10,000 independent experiments for the bootstrap CI.

We found that the best model was associated with Group 1, in which the CNN-based models outperformed the others with AUC values of 0.933 and F-1 score of 0.668. In Group 2, in which cross-attention was used for fusion, the performances of the respective models were close to one another. In this group, GRU-based and CNN-based models led, with AUCs of 0.885 and 0.875, respectively. Similarly, in Groups 3 and 4, in which each channel had trained separately, the CNN- and GRU-based models produced the highest performances, with AUCs of 0.927 and 0.912 for

Group 3 and 0.892 and 0.882 for Group 4, respectively.

Several notable observations from Table 3 are worthy of discussion. First, the attention-based model (BERT) was not the most effective one in all configurations (groups). The reason for this result may be that a more complex model such as BERT has a limited number of accepting tokens. The limitation of tokens refers to the maximum number of tokens that a model can process at a time; in the BERT, this limit is set to 512. This limitation would cause truncation of input sequences in long text and loss of information.

However, the other models did not have such limitation. The average token length in the text documents after cleaning in our application was 4,000, and the other models could process all of the information. Consequently, a more complex model such as BERT would not necessarily produce the highest performance in this case. In addition to producing higher performance, the CNN-based model could be trained in much less time compared to transformer-based models by taking advantage of fewer parameters.

Second, the attention-based models performed better when the fusion strategy was cross-attention rather than simple concatenation (Group 2 vs. Group 1 and Group 4 vs. Group 3). Moreover, maximum performance among all attention-based models occurred when each channel separately trained and cross-attention was then applied to mix the outputs (Group 4). When we used the cross-attention fusion technique (Group 2 and Group 4), the performances of models became close to each other compared to the concatenation technique. The values of AUC in Groups 1 and Group 3 were 0.16 and 0.1, respectively, while in Groups 2 and Group 4, these values were 0.06 and 0.05, respectively.

Finally, although both LSTM and GRU are based on RNNs, the performance of GRU was much better than LSTM in all groups. One possible reason for this effect is that GRUs are simpler models, having only two gates as opposed to LSTMs' three. This simplicity may allow GRUs to more effectively capture dependencies in a dataset that is not excessively large, such as one with 27,000 records. Based on these results, the most successful combination was the CNN-based model combined with early and intermediate fusion strategies in terms of level and with concatenation in terms of fusion technique, answering our first research question. Table 3 also shows that while there was a slight benefit of using both early and intermediate fusion strategies against only intermediate fusion, this difference may not have been statistically significant. Users should test both in real-world applications and use the one that better fits their purposes. In our case, given the slight

positive bias, we would recommend early and intermediate fusion (CNN in Group 1).

Figure 6 shows that the most effective model featured a CNN with three hidden layers (with dropout) for the text channel and another CNN with two hidden layers (with dropout) for the numeric channel, which were concatenated and fed to a dense neural network for computing the output. Each layer’s parameters specified a filter size of 64, kernel size of 2, the ReLU activation function, and strides of 1. We used 100 epochs, a batch size of 100, and a learning rate of 0.0001 for model training.

Group	Metric	Base Model			
		CNN	LSTM	GRU	Attention
1	AUC	0.933±0.004	0.794±0.004	0.864±0.005	0.778±0.006
	F1	0.668±0.011	0.174±0.007	0.516±0.012	0.348±0.011
2	AUC	0.875±0.005	0.822±0.005	0.885±0.005	0.831±0.005
	F1	0.57±0.011	0.459±0.011	0.545±0.011	0.434±0.011
3	AUC	0.927±0.003	0.855±0.005	0.912±0.004	0.823±0.005
	F1	0.661±0.011	0.478±0.011	0.603±0.011	0.437±0.011
4	AUC	0.892±0.005	0.834±0.005	0.882±0.005	0.879±0.004
	F1	0.58±0.012	0.476±0.012	0.557±0.012	0.521±0.012

Table 3: Model performance by base model and Group in terms of weighted-AUC and F-1 score.

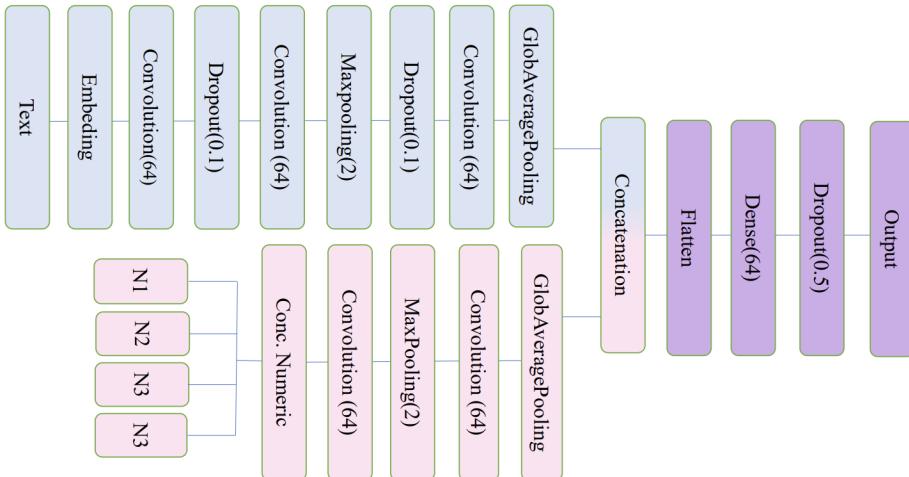


Figure 6: Architecture of the CNN ensemble for the best model, showing the convolution and dropout layers with two streams of data that includes text and numerical data (N1, N2, N3, N4).

The confusion matrix in Figure 7 shows that the most accurate prediction belonged to the two highest and lowest quality bonds, specifically 80% and 73% for the two lowest quality classes

and 76% and 70% for the highest quality “AAA” to “A” (based on Fitch and S&P agency ratings). This indicates that the proposed model can accurately predict credit ratings with low and high credit quality, which are the most critical classes for investors.

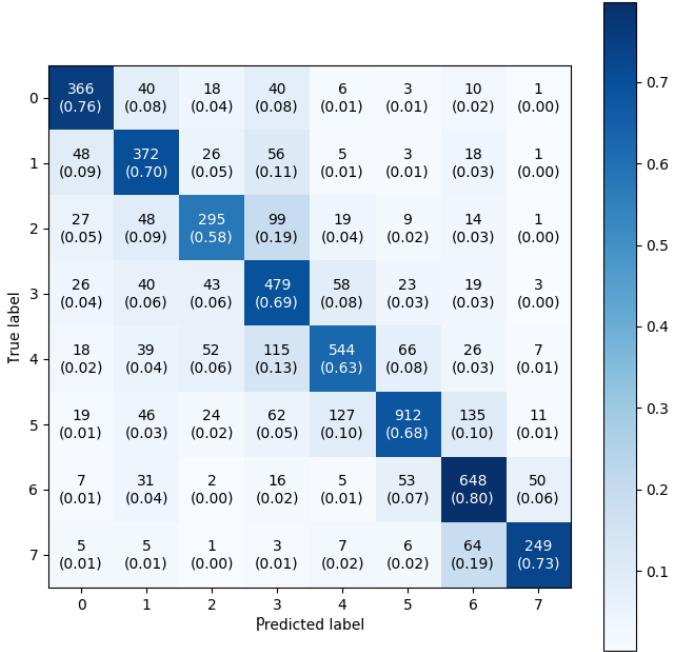


Figure 7: Confusion Matrix for the best model.

5.3 Robustness Test

To ensure the practicality of our model for real-world use, we conducted two robustness evaluations: out-of-time (OOT) and out-of-universe (OOU), also known as out-of-distribution (OOD). These evaluations played a crucial role in analyzing the reliability and generalizability of our model to handle unseen data in real-world scenarios. In this study, we evaluated the OOT performance of the model by splitting the train and test data based on the time index. The results in Table 4 indicate a high level of performance, with an AUC score of 0.935 for the time-based test data, indicating that the model’s performance was comparable to a random train–test split. Furthermore, we assessed the OOU by testing the model on a dataset containing companies that were not present in the training set. The results showed a high level of OOU performance, with an AUC score of 0.929, indicating that the model was able to accurately predict both previously seen and unseen companies and that the patterns found were structural rather than company-specific.

Metric	OOT	OOU	Baseline
AUC	0.935 ± 0.004	0.929 ± 0.004	0.933 ± 0.004
F1	0.689 ± 0.011	0.66 ± 0.011	0.668 ± 0.011

Table 4: Robustness test for the best model showing OOT and OOU and comparing them with the best model performance as a baseline.

5.4 The role of modality in prediction

We evaluated the role of modalities in the prediction and investigated which modality or data channel (text or numerical) had contributed the most. To do so, we retrained the model with each channel as the input. This could be performed by splitting the original dataset into separate datasets. Then, separate models could be trained on each dataset using the same algorithm. Once the models had been trained, we could evaluate their performances on a test set and compare their predictive power. Tables 5 and 6 show that the text channel had the most contribution (AUC of 0.915), which was significantly higher than the numerical channel. In the numerical channel, the highest contribution was related to covariate and financial ratio channels with AUC values of 0.808 and 0.791, respectively, suggesting that the information contained in these channels had been more informative for prediction than that in the other unstructured channels. The covariate channel provided information regarding the rating agency type and the last observed rating class; this information may have been crucial because the rating agency type can affect the reliability and accuracy of the ratings, and the last observed rating class can provide insights into the historical creditworthiness of the company. The financial ratio channel provided a wide range of features related to the financial performance of the companies and could provide insights into their financial health and their ability to repay debts.

Metric	Text contribution
AUC	0.915 ± 0.004
F1	0.654 ± 0.011

Table 5: Performance of text channel for the best model.

Metric	Covariate	Fin. Ratio	Bond	Market
AUC	0.808±0.006	0.791±0.006	0.777±0.006	0.53±0.007
F-1	0.444±0.011	0.35±0.011	0.339±0.011	0.095±0.006

Table 6: Performance of numerical channel for the best model.

5.5 The effect of COVID-19 on model performance

The COVID-19 crisis undoubtedly had an impact on many phenomena, including the financial market and its affected attributes (features) that would be useful for developing prediction models. Thus, it was valuable to examine the effect of this unexpected crisis on credit rating prediction. Table 7 indicates the performance of model prediction before versus during and after the COVID-19 crisis. We found that the AUC was 0.927 before COVID-19. However, after the pandemic, the AUC value increased to 0.964, suggesting that the model’s performance was improved. This improvement may have occurred because many predictable trends in the data became less credible after the crisis. The numerical data that was previously useful in prediction models became less relevant and less reliable. This effect led to a shift in the importance of oral and text reports, as executives could provide useful context to the challenges or competitive advantage each company had in the face of the pandemic, which became more significant in affecting corporate credit ratings.

Metric	Before COVID-19	During and After COVID-19
AUC	0.927±0.004	0.964±0.007
F1	0.657±0.011	0.777±0.026

Table 7: Comparison of the model before versus during and after COVID-19

5.6 Effect of the rating Agency and lag of Prediction

We further investigated the impact of agency rating type and lag of prediction both simultaneously and separately on the performance of the CNN-based model. Lag of prediction refers to the time elapsed between making a prediction and observation of the actual outcome. Examples of agency types include Moody’s Rating (MR)², Standard & Poor’s Rating (SPR)³ and Fitch Rating (FR)⁴

²<https://www.moodys.com/>

³<https://www.spglobal.com/ratings/en/>

⁴<https://www.fitchratings.com/>

White (2010). Table 8 presents the results, from which we can draw at least three conclusions:

Type	All	Short-Term	Medium-Term	Long-Term
All	0.933±0.003	0.936±0.006	0.935±0.005	0.925±0.008
	0.668±0.01	0.691±0.018	0.672±0.015	0.643±0.022
MR	0.945±0.005	0.945±0.009	0.951±0.006	0.938±0.012
	0.719±0.015	0.719±0.026	0.702±0.022	0.652±0.035
SPR	0.909±0.008	0.913±0.013	0.911±0.011	0.898±0.016
	0.604±0.018	0.609±0.034	0.597±0.027	0.595±0.036
FR	0.938±0.008	0.944±0.013	0.933±0.012	0.941±0.015
	0.687±0.024	0.694±0.043	0.683±0.033	0.674±0.045

Table 8: AUC and F1 score by Agency Type & Terms

The first column of Table 8 indicates the impact of rating type without considering lags. The performance of rating based on the “MR” type was the highest, with AUC = 0.945 and F-1 score = 0.719, followed by FR and then SPR. This result shows that for our specific dataset in midcaps, Moody’s provided the most accurate prediction of the point-in-time credit risk for these companies. As each rating was composed both of quantitative and qualitative information, we inferred that Moody’s and Fitch processes had followed the market more closely than S&P, which could either be focusing on through-the-cycle ratings or somehow missing the prediction of default. Nevertheless, all companies demonstrated strong performance overall.

The first row in Table 8 shows the performance of the model according to the lag of the prediction, in which short-term refers to the prediction of rating in less than 4 months and medium-term refers to 5 to 9 months. Finally, long-term prediction refers to predictions of more than 9 months. As expected, the short-term prediction outperformed the median and long-term ones (because credit risk is generally more predictable over shorter time horizons than longer ones), but this difference was very minor. In general, 1-year lookahead periods are very achievable with deep learning models.

Combining the result of agency type with the lag (according to the second row of Table 8), contrary to the expectation that short-term prediction may lead to better performance, medium-term prediction based on MR agency rating outperformed the short-term one. Again, this difference was relatively minor. This contradiction may have arisen because, according to the results of the previous section, the text channel contributes more than numerical channel, such as the previous

rating information, to the prediction. Thus, it would not seem unrealistic that the impact of the transcript would become more apparent over a longer period depending on the agency-type rating of the target data.

6 Conclusions

In this paper, we have presented a framework for the evaluation of 16 multimodal models consisting of a combination of different fusion type and fusion level strategies, combining deep learning models for credit rating prediction. We have shown that a CNN-based model with early and intermediate fusion strategies in terms of level, and concatenation in terms of technique, outperformed others, including those with more complex structures, such as attention-based models. Consequently, the most complex structures did not necessarily perform the best, suggesting that modelers must exercise care in evaluating a large pool of models to find the best-performing ones in credit risk applications. In addition, we conducted a test of robustness to ensure that these models would be applicable when provided with newer information, contributing to the stability of our model. Furthermore, analysis of the contribution of each modality to the prediction showed that the text channel played the most significant role, showcasing the importance of the context provided by managers, particularly during the COVID-19 pandemic. Finally, we investigated the impact of the type of rating agency on short-, medium-, and long-term predictions and discovered that credit ratings generated by Moody’s performed better than the other two rating companies that were evaluated, particularly in the medium term; however, these differences are relatively small.

Future work could include the incorporation of Bayesian models for uncertainty quantification using advanced deep learning models and application of this framework to country-based credit rating prediction.

We believe that this work provides a framework for evaluating different multimodal models for credit rating prediction and supports developments that can leverage a larger amount of available data. The results of this work can be used by rating agencies and other financial institutions to make more informed decisions regarding credit ratings. The finding that text modality significantly outperformed other modalities with respect to prediction can be used by financial institutions to make more informed decisions. In addition, the results regarding the reliability of the model during unexpected crises can be useful in terms of developing more robust and reliable models during such

times.

7 Code

The source code of the framework can be found at the following Github repository:

<https://github.com/Banking-Analytics-Lab/MultimodalFusionRatings>

Acknowledgment

The first, third, and last author acknowledge the support of the NSERC Discovery Grant program [RGPIN-2020-07114]. This research was undertaken, in part, thanks to funding from the Canada Research Chairs program. This work was enabled in part by support provided by Compute Ontario (computeontario.ca), Calcul Québec (calculquebec.ca), and the Digital Research Alliance of Canada (alliancecan.ca).

References

Gunter Löffler. The systemic risk implications of using credit ratings versus quantitative measures to limit bond portfolio risk. *Journal of Financial Services Research*, 58(1):39–57, 2020.

Liu Yijun, Cai Qiuru, Luo Ye, Qian Jin, and Ye Feiyue. Artificial neural networks for corporation credit rating analysis. In *2009 International Conference on Networking and Digital Society*, volume 1, pages 81–84. IEEE, 2009.

Jing He, Yanchun Zhang, Yong Shi, and Guangyan Huang. Domain-driven classification based on multiple criteria and multiple constraint-level programming for intelligent credit scoring. *IEEE transactions on knowledge and data engineering*, 22(6):826–838, 2010.

Qiang Liu, Yingtao Luo, Shu Wu, Zhen Zhang, Xiangnan Yue, Hong Jin, and Liang Wang. Rmt-net: Reject-aware multi-task network for modeling missing-not-at-random data in financial credit scoring. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Yair E Orgler. *Evaluation of bank consumer loans with credit scoring models*. Tel-Aviv University, Department of Environmental Sciences, 1971.

Lyn Thomas, Jonathan Crook, and David Edelman. *Credit scoring and its applications*. SIAM, 2017.

Yuhan Zhu. Research on financial risk control algorithm based on machine learning. In *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLDBI)*, pages 16–19. IEEE, 2021.

Ashu Jain and Avadhnam Madhav Kumar. Hybrid neural network models for hydrologic time series forecasting. *Applied Soft Computing*, 7(2):585–592, 2007.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Li Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA transactions on Signal and Information Processing*, 3, 2014.

Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *Applied Soft Computing*, 93:106384, 2020.

Lei Cai, Jingyang Gao, and Di Zhao. A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, 8(11), 2020.

Keiron O’Shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

Pouya Khodaei, Akbar Esfahanipour, and Hassan Mehtari Taheri. Forecasting turning points in stock price by applying a novel hybrid cnn-lstm-resnet model fed by 2d segmented images. *Engineering Applications of Artificial Intelligence*, 116:105464, 2022.

Rohitash Chandra, Shaurya Goyal, and Rishabh Gupta. Evaluation of deep learning models for multi-step ahead time series prediction. *IEEE Access*, 9:83105–83123, 2021.

Bojing Feng, Wenfang Xue, Bindang Xue, and Zeyu Liu. Every corporation owns its image: Corporate credit ratings via convolutional neural networks. In *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, pages 1578–1583. IEEE, 2020.

Larry R Medsker and LC Jain. Recurrent neural networks. *Design and Applications*, 5:64–67, 2001.

Tong Zhang, Wenming Zheng, Zhen Cui, Yuan Zong, and Yang Li. Spatial-temporal recurrent neural network for emotion recognition. *IEEE transactions on cybernetics*, 49(3):839–847, 2018.

KR1442 Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Guizhu Shen, Qingping Tan, Haoyu Zhang, Ping Zeng, and Jianjun Xu. Deep learning with gated recurrent unit networks for financial sequence predictions. *Procedia computer science*, 131:895–903, 2018.

Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. Neural machine translation with bert for post-ocr error detection and correction. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 333–336, 2020.

Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

Rohitash Chandra and Aswin Krishna. Covid-19 sentiment analysis via deep learning during the rise of novel cases. *PLoS One*, 16(8):e0255615, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019.

Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020.

Xiang Wang, Min Xu, and Özgür Tolga Pusatli. A survey of applying machine learning techniques for credit rating: Existing models and open issues. In *International Conference on Neural Information Processing*, pages 122–132. Springer, 2015.

Qing Li, Jinghua Tan, Jun Wang, and Hsinchun Chen. A multimodal event-driven lstm model for stock prediction using online news. *IEEE Transactions on Knowledge and Data Engineering*, 33(10):3323–3337, 2020a.

Matthew Stevenson, Christophe Mues, and Cristián Bravo. The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, 295(2):758–771, 2021.

Arianna D’Ulizia. Exploring multimodal input fusion strategies. In *Multimodal Human Computer Interaction and Pervasive Services*, pages 34–57. IGI Global, 2009.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

Yu Zhu, Wenbin Chen, and Guodong Guo. Fusing multiple features for depth-based action recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(2):1–20, 2015.

Said Yacine Boulahtia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121, 2021.

Sang Il Lee and Seong Joon Yoo. Multimodal deep learning for finance: integrating and forecasting international stock markets. *The Journal of Supercomputing*, 76(10):8294–8312, 2020.

Ren C Luo and Michael G Kay. Multisensor integration and fusion: issues and approaches. In *Sensor Fusion*, volume 931, pages 42–49. SPIE, 1988.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019.

Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

Tong Meng, Xuyang Jing, Zheng Yan, and Witold Pedrycz. A survey on machine learning for data fusion. *Information Fusion*, 57:115–129, 2020.

Mirjana Pejić Bach, Živko Krstić, Sanja Seljan, and Lejla Turulja. Text mining for big data analysis in financial sector: A literature review. *Sustainability*, 11(5):1277, 2019.

Richard M Frankel, Jared N Jennings, and Joshua A Lee. Using natural language processing to assess text usefulness to readers: The case of conference calls and earnings prediction. *Available at SSRN 3095754*, 2017.

Abel Brodeur, David Gray, Anik Islam, and Suraiya Bhuiyan. A literature review of the economics of covid-19. *Journal of Economic Surveys*, 35(4):1007–1044, 2021.

Feyyaz Zeren and Atike Hizarci. The impact of covid-19 coronavirus on stock markets: evidence from selected countries. *Muhasebe ve Finans İncelemeleri Dergisi*, 3(1):78–84, 2020.

Shunqin Chen, Zhengfeng Guo, and Xinlei Zhao. Predicting mortgage early delinquency with machine learning methods. *European Journal of Operational Research*, 290(1):358–372, 2021.

Trevor Fitzpatrick and Christophe Mues. How can lenders prosper? comparing machine learning approaches to identify profitable peer-to-peer loan investments. *European Journal of Operational Research*, 294(2):711–722, 2021.

Björn Rafn Gunnarsson, Seppe Vanden Broucke, Bart Baesens, María Óskarsdóttir, and Wilfried Lemahieu. Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*, 295(1):292–305, 2021.

Havard Kvamme, Nikolai Sellereite, Kjersti Aas, and Steffen Sjursen. Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217, 2018.

Sami Ben Jabeur, Salma Mefteh-Wali, and Jean-Laurent Viviani. Forecasting gold price with the xgboost algorithm and shap interaction values. *Annals of Operations Research*, pages 1–21, 2021.

Alisa Kim, Y Yang, Stefan Lessmann, Tiejun Ma, M-C Sung, and Johnnie EV Johnson. Can deep learning predict risky retail investors? a case study in financial risk behavior forecasting. *European Journal of Operational Research*, 283(1):217–234, 2020.

Manjeevan Seera, Chee Peng Lim, Ajay Kumar, Lalitha Dhamotharan, and Kim Hua Tan. An intelligent payment card fraud detection system. *Annals of operations research*, pages 1–23, 2021.

Jian Huang, Junyi Chai, and Stella Cho. Deep learning in finance and banking: A literature review and classification. *Frontiers of Business Research in China*, 14(1):1–24, 2020.

Victor-Emil Neagoe, Adrian-Dumitru Ciote, and George-Sorin Cucu. Deep convolutional neural networks versus multilayer perceptron for financial prediction. In *2018 International Conference on Communications (COMM)*, pages 201–206. IEEE, 2018.

Parisa Golbayani, Dan Wang, and Ionut Florescu. Application of deep neural networks to assess corporate credit rating. *arXiv preprint arXiv:2003.02334*, 2020.

Hongyi Qian, Ping Ma, Songfeng Gao, and You Song. Soft reordering one-dimensional convolutional neural network for credit scoring. *Knowledge-Based Systems*, page 110414, 2023.

Juliana Adisa, Samuel Ojo, Pius Owolawi, Agnieta Pretorius, and Sunday O Ojo. Credit score prediction using genetic algorithm-lstm technique. In *2022 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6. IEEE, 2022.

Kamesh Korangi, Christophe Mues, and Cristián Bravo. A transformer-based model for default prediction in mid-cap corporate markets. *European Journal of Operational Research*, 2022.

Hyeyongjun Kim, Hoon Cho, and Doojin Ryu. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. *Computational Economics*, 59(3):1231–1249, 2022.

Huishan Zhuang, Rui Zhao, Xi Luo, and Chengyu Yang. Research on quantitative stock selection strategy based on cnn-lstm. In *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 1142–1147. IEEE, 2022.

GS Vidya and VS Hari. Gold price prediction and modelling using deep learning techniques. In *2020 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, pages 28–31. IEEE, 2020.

Saugat Aryal, Dheynoshan Nadarajah, Dharshana Kasthurirathna, Lakmal Rupasinghe, and Chandimal Jayawardena. Comparative analysis of the application of deep learning techniques for forex rate prediction. In *2019 international conference on advancements in computing (ICAC)*, pages 329–333. IEEE, 2019.

Yuyang Lin, Qiyin Zhong, Qi Huang, Muyang Li, and Fei Ma. A new convolutional neural network and long short term memory combined model for stock index prediction. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–6. IEEE, 2021.

Lei Shi, Zhiyang Teng, Le Wang, Yue Zhang, and Alexander Binder. Deepclue: visual interpretation of text-based deep stock prediction. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1094–1108, 2018.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In *LREC*, volume 2014, pages 1170–1175, 2014.

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. Deep learning models for bankruptcy prediction using textual disclosures. *European journal of operational research*, 274(2):743–758, 2019.

Manuel R Vargas, Beatriz SLP De Lima, and Alexandre G Evsukoff. Deep learning for stock market prediction from financial news articles. In *2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA)*, pages 60–65. IEEE, 2017.

Gregor Dorfleitner, Christopher Priberny, Stephanie Schuster, Johannes Stoiber, Martina Weber, Ivan de Castro, and Julia Kammler. Description-text related soft information in peer-to-peer lending—evidence from two leading european platforms. *Journal of Banking & Finance*, 64: 169–187, 2016.

Xiao Chen, Bihong Huang, and Dezhu Ye. The role of punctuation in p2p lending: Evidence from china. *Economic Modelling*, 68:634–643, 2018.

Oded Netzer, Alain Lemaire, and Michal Herzenstein. When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research*, 56(6):960–980, 2019.

Johannes Kriebel and Lennart Stitz. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, 302(1):309–323, 2022.

Rastin Matin, Casper Hansen, Christian Hansen, and Pia Mølgaard. Predicting distresses using deep learning of text segments in annual reports. *Expert Systems with Applications*, 132:199–208, 2019.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Patel Dhruv and Subham Naskar. Image classification using convolutional neural network (cnn) and recurrent neural network (rnn): a review. *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*, pages 367–381, 2020.

NI Widiastuti. Convolution neural network for text mining and natural language processing. In *IOP Conference Series: Materials Science and Engineering*, volume 662, page 052010. IOP Publishing, 2019.

Muhammadjon Musaev, Ilyos Khujayorov, and Mannon Ochilov. Image approach to speech recognition on cnn. In *Proceedings of the 2019 3rd International Symposium on Computer Science and Intelligent Control*, pages 1–6, 2019.

Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

Dozdar Mahdi Ahmed, Masoud Muhammed Hassan, and Ramadhan J Mstafa. A review on deep sequential models for forecasting time series data. *Applied Computational Intelligence and Soft Computing*, 2022, 2022.

Tian Wang and Kyunghyun Cho. Larger-context language modelling. *arXiv preprint arXiv:1511.03729*, 2015.

Ralf C Staudemeyer and Eric Rothstein Morris. Understanding lstm-a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

Shudong Yang, Xueying Yu, and Ying Zhou. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International workshop on electronic communication and artificial intelligence (IWECAI)*, pages 98–101. IEEE, 2020.

Peter T Yamak, Li Yujian, and Pius K Gadosey. A comparison between arima, lstm, and gru for time series forecasting. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 49–55, 2019.

Boyang Li, Yulan Guo, Jungang Yang, Longguang Wang, Yingqian Wang, and Wei An. Gated recurrent multiattention network for vhr remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.

Silvia Cascianelli, Gabriele Costante, Thomas A Ciarfuglia, Paolo Valigi, and Mario L Fravolini. Full-gru natural language video description for service robotics applications. *IEEE robotics and automation letters*, 3(2):841–848, 2018.

Mahdi Namazifar, Alexandros Papangelis, Gokhan Tur, and Dilek Hakkani-Tür. Language model is all you need: Natural language understanding as question answering. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7803–7807. IEEE, 2021.

Glyn W Humphreys and Jie Sui. Attentional control and the self: the self-attention network (SAN). *Cognitive neuroscience*, 7(1-4):5–17, 2016.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer, 2019.

Michael Neumann and Ngoc Thang Vu. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*, 2017.

Franklin E White. Data fusion lexicon. Technical report, Joint Directors of Labs Washington DC, 1991.

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Lawrence J White. Markets: The credit rating agencies. *Journal of Economic Perspectives*, 24(2): 211–226, 2010.

Jeff Jewell and Miles Livingston. A comparison of bond ratings from moody's s&p and fitch ibca. *Financial Markets, Institutions & Instruments*, 8(4):1–45, 1999.

Seeking Alpha. Seeking alpha. <https://seekingalpha.com>. Most recent collection as of July 2022.

Megan French-Marcelin. *The Rise of Financial Information Platforms: Markets, Machines, and Open Data*. Routledge, New York, NY, 2020. ISBN 9780367187641. doi: 10.4324/9780429459138.

Qing Li, Hongyu Shan, Yuehua Tang, and Vincent Yao. Corporate climate risk: Measurements and responses. *Available at SSRN 3508497*, 2020b.

Keras tokenizer. https://keras.io/api/keras_nlp/tokenizers/tokenizer/. Accessed: September 2022.

Bert tokenizer. https://huggingface.co/docs/transformers/model_doc/bert. Accessed: September 2022.

Hongxia Lu, Louis Ehwerhemuepha, and Cyril Rakovski. A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. *BMC Medical Research Methodology*, 22(1):1–12, 2022.

Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

Adam Poliak, Pushpendre Rastogi, M Patrick Martin, and Benjamin Van Durme. Efficient, compositional, order-sensitive n-gram embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 503–508, 2017.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.