AFT-Attentive BiLSTM: Improving Early Warning of Firm Financial Distress with Temporal Attention in an Accelerated Failure Time Framework

Muhammad Ali Chohan[1], Suresh Ramakrishnan[2]*,
Mohammad Abrar[3], Shamaila Butt[4], Shahid Kamal[5]*
Faculty of Management, Universiti Teknologi Malaysia, Johor Bahru, 81310, Malaysia[1,2]
Faculty of Business, Sohar University, Sohar, 311, Oman[2,4]
Faculty of Computer Studies, Arab Open University, Muscat, 130, Oman[3]
Center for Advanced Analytics, CoE for Artificial Intelligence,
Multimedia University, Parsiaran Multimedia 63100, Cyberjaya, Selangor Malaysia[5]

***Abstract*—Early warning systems (EWS) for firm-level financial distress are essential for identifying potential bankruptcies or insolvencies before their realization. While traditional statistical models such as Z-score and logistic regression offer interpretability, they lack the ability to capture nonlinear and temporal dependencies in financial data. Recent deep learning approaches improve predictive accuracy but often sacrifice interpretability. The purpose of this study is to develop and evaluate a novel deep learning-based early warning model for firm-level financial distress that integrates temporal attention with parametric survival analysis to improve both predictive accuracy and interpretability. Therefore, this study proposes an AFT-Attentive BiLSTM model that integrates a Bidirectional Long Short-Term Memory (BiLSTM), a temporal attention mechanism, and a log-normal Accelerated Failure Time (AFT) survival framework. The model predicts time-to-distress distributions rather than binary outcomes, enabling probabilistic early warnings with calibrated survival probabilities. Empirical results demonstrate that the proposed model outperforms Cox Proportional Hazards, DeepSurv, and prior AFT-BiLSTM models without attention. The inclusion of temporal attention improves concordance index (C-index), Integrated Brier Score (IBS), and time-dependent AUC, and provides interpretable insights by identifying critical financial periods preceding distress. Kaplan–Meier analysis confirms strong separation between high- and low-risk groups. The findings suggest that combining temporal attention with parametric survival modeling enhances both predictive accuracy and interpretability in financial distress early warning systems.**

***Keywords*—Financial distress prediction; early warning systems; survival analysis; accelerated failure time; bidirectional LSTM; temporal attention; deep learning; corporate bankruptcy**

I. Introduction

Early warning systems (EWS) for financial distress at the firm level are essential for stakeholders to predict and mitigate potential bankruptcies or insolvencies before they become complete [1]. Classical methods for predicting bankruptcy date back to ratio analysis models. The Z-score proposed by Altman [2] was based on financial ratios that were analyzed by discriminant analysis, and the logit model of Ohlson [3] employed multivariate regression to forecast failure [2]–[7]. These models were insightful but were based on linear assumptions and binomial models of firm health, and therefore lacked the ability to capture the dynamic nature of corporate distress. Researchers have, over the years, developed methods for hazard modelling and survival analysis (e.g., CoX proportional hazards) to account for the temporal dimension of distress, noting that people often do not go bankrupt immediately but rather through a gradual decline [8]. For example, Turetsky and McEwen [9] used survival analysis to monitor the financial well-being of corporations over time, underscoring the importance of time-to-event models in this context.

Machine learning (ML) and deep learning have, in recent decades, opened new vistas in EWS by learning complex nonlinear patterns that traditional models must ignore. Random forests and boosting are ensemble techniques that have been shown to have greater predictive power for bankruptcy than logistic regression [10], [11]. Similarly, deep neural networks have also been used in financial distress prediction: convolutional neural networks (CNNs) learn features on transformed financial ratio images (increasing prediction accuracy up to 87% in certain situations), and recurrent neural networks such as LSTM and GRU learn features on sequential financial data [11]. Nonetheless, such complex models are often described as black boxes that can be interpreted only at a limited level, and they may not be able to explain precisely why a firm risks distress [1]. Recent studies have started working on this, including the attention mechanisms and explainability tools to deep EWS models. Indicatively, [12] incorporated attention in a dual-LSTM architecture in predicting financial risks and reported a high performance (F1-score 96.9) compared to conventional models which do not have attention, with this evidence indicating that attention can draw material features or intervals that contribute to prediction. Similarly, [13] used an Attention-LSTM to integrate numeric and textual information to provide systemic risk warnings, outperforming conventional methods. These developments imply that combining attention-based interpretability with deep sequential models could improve the accuracy and transparency of early warning systems.

A sequence-based survival model was presented in a previous study by Londono and Velasquez [14] to forecast firm distress using a bidirectional LSTM within an Accelerated Failure Time (AFT) framework. That model (herein denoted as

*Corresponding authors

AFT-BiLSTM) demonstrated that time-to-event financial distress prediction using deep learning is viable, and it achieved a higher concordance index and shorter early-warning lead time than its classical Cox and logistic equivalents. The prior method, however, did not leverage any attention mechanism, implicitly treating each time step as equally contributing. This may restrict interpretability and may even miss the diverse significance that different timeframes of a firm exhibit (e.g., a sudden decline in profits two years before bankruptcy may be more predictive than changes during its healthier years).

A. Problem Statement

Although there has been an advancement in combining deep learning with a financial distress survival analysis, there still exists a critical gap. Current deep survival models, including DeepSurv, have used the Cox proportional hazards framework, which yields only relative risk scores and does not provide explicit failure-time distributions. Sequential models, e.g., LSTMs or GRUs, on the other hand, are better at recognizing temporal patterns, but as black boxes, they do not have mechanisms that describe what time in the history of a firm is most predictive of distress. The previous AFT-BiLSTM model resolved the issue of the survival-time prediction gap by implementing an AFT framework, but did not use the possibility to distinguish between informative and non-informative time periods. It is then desired a model that: 1) forecasts the calibrated time-to-distress probability distributions, 2) represents the complicated time-dependent effects of sequential financial data, and 3) gives interpretable time-dependent attention weights, which indicate when warning signals begin appearing in the history of a firm.

Based on this, we present a new model, AFT-Attentive BiLSTM, that applies a time-dependent attention mechanism to the LSTM outputs of the parametric AFT survival model. The model can increase predictive accuracy and provide insight into the temporal behavior preceding failure by enabling the network to learn which time steps in a firm's financial history are most predictive of distress onset.

There are three contributions of AFT-Attentive BiLSTM that are peculiar. First, it combines a Bidirectional LSTM with a learned attention layer that scales the LSTM hidden states at each time step, thereby allowing the model to emphasize important periods in the sequence. This is paired with a log-normal AFT survival model, i.e., the network directly predicts a time-to-distress probability distribution (when log-time is normally distributed) rather than a binary outcome or a hazard rate. This parametric survival method also provides an opportunity to estimate not only relative risk but also survival probabilities with time, which give more informative early warnings. Second, the model is trained using techniques to address the imbalance in distress events: we employ over-sampling of distressed cases and specialized loss functions to handle censored data, ensuring that the model learns effectively despite many firms not experiencing distress during the observation window. Third, we demonstrate, through extensive experiments, that AFT-Attentive BiLSTM achieves superior empirical performance compared with baseline models (including Cox Proportional Hazards, the DeepSurv neural Cox model, and the prior AFT-BiLSTM without attention). Notably, our model yields higher concordance index (C-index) scores, indicating improved ranking of survival times, and it provides interpretable attention-weight distributions that align with financial intuition (e.g., identifying the years in which worsening financial ratios precede distress). We present visualizations, including Kaplan–Meier plots comparing high- and low-risk groups, heatmaps of learned attention weights, training curves, and comparative metrics, to illustrate these advantages.

The main goal of the research is to develop an understandable deep learning model for time-to-event financial distress prediction, which is the integration of device attention with parametric survival analysis over time. In particular, the questions considered in this study are:

- Does integrating a temporal attention mechanism into the AFT-BiLSTM framework improve the concordance index and predictive discrimination for firm-level financial distress?
- Can the learned temporal attention weights provide interpretable insights into which financial periods are most predictive of corporate distress?
- How does the proposed AFT-Attentive BiLSTM compare against classical survival models (CoxPH), neural survival models (DeepSurv), and the prior non-attention AFT-BiLSTM in terms of concordance index, Integrated Brier Score, and time-dependent AUC?

The remainder of the study is organized as follows: In Section II (Literature Review), we situate our work within the existing literature on financial distress prediction, survival analysis, and deep learning models, thereby identifying the gap our model will address. Section III (Methodology) will describe the architecture of the proposed model and the training process, including training the AFT formulation, the temporal attention mechanism, and data-handling strategies. Section IV (Results) describes the model's performance relative to baselines, as well as attention pattern and survival curve analyses. The findings are interpreted in Section V (Discussion), which compares them to the previous study and explains the implications for early warning capabilities and interpretability. Lastly, Section VI presents the practical implications and limitations. Section VII and Section VIII present the Conclusion and Future Work, respectively, by highlighting how AFT-Attentive BiLSTM advances the current state-of-the-art in firm distress early warning and what can be done further to advance the research direction.

II. Literature Review

A. Early Financial Distress Prediction Models

1) Early Financial Distress Prediction Models: Quantitative explanations of corporate financial distress are not new in the field of finance. The approaches that form the basis were pioneered in the 1960s-1980s. Analysis of individual financial ratios as predictors of individual failure were conducted by Beaver [4], and multivariate discriminant analysis was introduced by Altman [2] to incorporate ratios into the well-known Z-score, yielding substantially higher classification rates for bankruptcy. These initial models established that troubled firms exhibit distinct patterns in profitability, leverage, liquidity, and solvency ratios years prior to bankruptcy. One such model, Z-score model by Altman [2], found such measures as the working capital-to-assets and retained earnings-to-assets as critical measures and was a staple tool of practitioners [15]–[17]. Nonetheless, discriminant analysis presumes that the discriminants can be separated linearly and is not likely to capture more intricate patterns. Further studies shifted to logistic regression: one of the famous models used the logit model by Ohlson [3]: this model employed financial and market variables to determine the likelihood of failure. The logistic and probit models offered a probabilistic interpretation and relaxed some assumptions of discriminant analysis, but they nevertheless did not fully exploit the time-series information because each firm-year observation was independent. Additionally, these statistical models are constrained by a priori functional forms (linear combinations of inputs) and may struggle to capture nonlinear interactions or threshold effects in the data.

B. Survival Analysis and Multi-Stage Distress Models

Since the progression of bankruptcy is the final stage in a more general deterioration process, researchers have used survival analysis to simulate time-to-distress. Shumway [18] is well known to have proposed a hazard model approach to predict bankruptcies, by considering each firm-year to be a longitudinal survival data set hence capturing varying financial conditions over time. Estimation of hazard (instantaneous failure risk) as a function of firm characteristics has been done using cox proportional hazards (CoxPH) models 42. These semi-parametric models are good in dealing with the problem of censoring and do not make assumptions regarding the desired baseline time-to-event distribution, thus being appealing to corporate failure data where many firms remain alive (active) at the end of the study. As an illustration, a CoxPH model could show the extent to which an increase in a unit of a given financial ratio carries the hazard of bankruptcy at a given point in time. A number of studies were extrapolating this to multi-stage definitions of distress: instead of binary healthy vs. bankrupt, firms can move on the scale of normal, financially distressed (not yet bankrupt), and bankrupt. Such transitions have been modeled using survival models and competing risk frameworks, as well as other methods, e.g., null hypothesis testing methods, which are more comprehensible to economists (see [19]–[23]). Turetsky and McEwen [9] used survival analysis to investigate the process through which firms go through phases of financial distress and concluded that a better predictive lead time is obtained by including interim distress events. Similarly, the concept of multi-state models and Markov chains has been investigated to represent the process of slight-to-severe distress before bankruptcy occurs. These investigations clearly highlight that time is an important element: not only can a firm fail, but when a firm fails and what are the first indications before its downfall are the key points of an effective EWS [21], [22].

The classical survival models, which include CoxPH, are although strong, are founded on the assumption of proportional hazard (covariate effects are constant over time), and in many cases are linear in the covariate effects. Alternatives. Accelerated Failure Time (AFT) models provide an opportunity to model the impact of covariates by assuming that they accelerate or decelerate the failure clock (i.e. directly modeling the survival time). AFT models have been not widely used in the financial distress setting, yet they offer a direct mechanism of estimating the expected time to failure when predictors are available. Their previous study in Risks used a log-normal AFT model with an LSTM network, which was among the earliest uses of deep learning when it comes to AFT to predict the distress of firms. It tackled the issue by predicting the distribution of the failure time of a firm and showed that a well-trained AFT neural model could generate well-calibrated survival probabilities and do better than Cox models, which assume proportional hazards. Continuing on the basis of it, the present study expands the AFT style to include sophisticated neural network elements (bidirectional LSTM and attention) to further improve predictive accuracy.

C. Machine Learning and Ensemble Methods

With the increasing available computational power and data, researchers resorted to machine learning methods in order to provide better bankruptcy forecasting than linear models could offer. Another early ML approach to this issue was decision trees, which provided straightforward if-else rules based on financial ratios, which were used in [24]. Although the trees or individual trees were susceptible to overfitting, the ensemble techniques such as the random forests pooled many trees together and was accurate as well as robust. Kim and Kang [25] observed that tree-based models can have better performances than logistic regression particularly in the ability of the model to represent the nonlinear effects and interactions between variables. Random forests, employed by Tanaka et al. among others, were able to generate good results in predicting financial vulnerability. In fact according to a recent study by Farooq and Qamar [26] an insolvency-oriented random forest EWS was capable of identifying transitional distress with good accuracy. SVMs, boosting algorithms (XGBoost, AdaBoost) and neural networks were also thoroughly tested as machine learning models. According to a meta-analysis by Barboza, Altman & Kimura [27] different ML models (such as neural nets and ensembles) were reported to have accuracies of 80 to 90 per cent, which is considerably high compared to previous statistical models. Nevertheless, an issue that is frequently mentioned against these ML models is the inability to understand the reasons behind their actions a trade-off of the flexibility they provide. Attempts were also undertaken to interpret ML and neural forecasts using methods such as Shapley values (SHAP) or LIME to infer corporate failure e.g., Tran et al., 2022 [28] using SHAP to predict distress, and the fact that explainability is becoming increasingly important along with raw performance.

D. Deep Learning for Financial Distress and Survival

Deep learning has demonstrated the ability to extract complicated financial patterns. Hosaka [11] used a convolutional neural network, but transformed financial ratios into matrices with the appearance of a picture, obtaining high accuracy (up to 97 per cent) on Japanese firms. The convolution layer was able to identify the spatial patterns in the ratio matrix with reference to the underlying financial conditions. Recurrent neural networks, which by their nature deal with sequence data, are more directly applicable to our work (including time-series data on financial statements). Other authors, such as Vochozka et al. [29], and others applied LSTM and GRU networks to sequences of financial ratios, with improvements in predictive performance of corporate distress. Such RNN-based models are able to acquire temporal dependencies, such as, the joint effect of a slow decrease in revenue and a burst in debt in the course of several years to indicate trouble. In a single study, the article utilized LSTM-based model to learn long-term trends in company data and predict bankruptcy because it outperformed a fixed neural network. Currently, the focus has shifted to hybrid architectures and new training structures. As an example, other researchers used autoencoders together with LSTM to achieve dimensionality reduction of financial features before sequence modeling. There were those who combined macroeconomic variables and news sentiment (textual data) with the financial ratios in an LSTM to capture the external factors affecting health of firms.

In the context of survival analysis with deep learning, the seminal DeepSurv model by Katzman et al. [30] demonstrated that a feed-forward neural network can be trained using a Cox partial likelihood to model nonlinear effects in medical survival data. DeepSurv achieved higher concordance than CoxPH across several datasets and paved the way for applying deep neural networks to survival problems. This idea has been extended to recurrent nets: Giunchiglia et al. [31] introduced RNN-Surv, a deep recurrent model for survival analysis which processed sequential patient data to predict survival [32]. In clinical settings, LSTM-based survival models have been shown to handle time-series covariates and to outperform static models (Runquan and Xiaoping 2024). For example, a recent LSTM-Cox model on cancer recurrence data achieved a C-index of 0.90, significantly higher than traditional approaches, and effectively distinguished high- and low-risk groups in Kaplan–Meier analyses. These successes indicate that deep sequential models can greatly enhance time-to-event predictions. However, most deep survival models to date use either the Cox partial likelihood (which yields relative risk but not an explicit survival-time distribution) or focus on medical applications. Our work brings these advances to the financial domain, using an AFT formulation to predict when a firm might fail and incorporating an attention mechanism to improve interpretability.

E. Attention Mechanisms and Interpretability

Attention mechanisms, first popularized in NLP for machine translation [33], have since been applied across domains to improve model focus and interpretability. In sequence modeling, attention allows the model to dynamically weigh the importance of different time steps or inputs when producing an output. For financial distress prediction, an attention layer can, for instance, enable the model to focus on years that exhibit significant distress signals (e.g., the onset of losses or a liquidity crunch) while downweighting periods of relative stability. Cheng's [12] attention-embedded DUAL-LSTM, mentioned earlier, is a direct example in the financial risk arena: by applying attention to LSTM outputs, it significantly outperformed non-attention LSTMs and traditional models. Not only did it improve predictive metrics, but it provided a weight vector indicating which features/time periods were most influential, adding a layer of transparency to the model's decisions. Similarly, [13] have used attention-based RNNs for macro-financial risk and found that attention helps capture episodic signals (like bubbles or crises indicators) more effectively. In the realm of survival analysis, attention is a newer innovation. Wang et al. [34] proposed ResDeepSurv which included a self-attention mechanism to learn the importance of different latent features for survival prediction. The model achieved performance on par or better than other methods and highlighted which feature patterns were most critical to outcomes [34]. These developments suggest that attention can bridge some gap between the accuracy of deep models and the explainability of traditional models. In our case, by inspecting the attention weights over time for each firm, analysts can identify when the model believes things started to go wrong for that company – providing insight, for example, that "the model is focusing heavily on Year T-2 financials, likely because of the sharp drop in working capital and surge in debt that year". Such information is valuable for decision-makers to validate the model's reasoning or to take targeted preventive actions.

F. Research Gaps

Based on this literature review, there is a clear trajectory toward the use of deep, sequential, and interpretable models for early warning of financial distress. Particularly, three fundamental gaps are identified by the literature. To begin with, a temporal attention layer has not been explicitly incorporated into an AFT-based LSTM corruption predictor of corporate distress in the past. Although attention has been applied to financial risk forecasting (e.g., [12][13]), this does not fit within a survival analysis model and hence cannot generate time-to-failure distributions. Second, available deep survival models like DeepSurv [30] and RNN-Surv [31] are based on the Cox proportional hazards formulation, which generates relative risk scores instead of calibrated time-dependent probability of survival. The AFT formulation is a direct model of survival time that has the capability to produce probabilistic early warnings, but has seldom been used together with deep sequential architectures. Third, although the previous AFT-BiLSTM demonstrated the usefulness of deep learning in the context of an AFT, the lack of an attention mechanism constrained its predictive capabilities as well as its capacity to offer temporal insights into the onset of distress. These limitations are the reason to create the AFT-Attentive BiLSTM that is aimed at covering all of them by integrating sequence learning through BiLSTM, temporal interpretability through attention, and parametric survival-time prediction through a parametric AFT model. The research gap is linked with the literature and summerized in Table I.

III. Methodology

The proposed AFT-Attentive BiLSTM model integrates a Bidirectional Long Short-Term Memory (BiLSTM) network with a temporal attention mechanism within a parametric Accelerated Failure Time (AFT) survival analysis framework. The objective is to predict the time-to-distress distribution of firms using sequential financial data.

A. Data Description and Preprocessing

The sample in this study is based on the Compustat North America database, which includes publicly traded U.S. companies between 2000 and 2020. The sample includes 4850 observations of firm-years of 1120 different firms, including manufacturing, services, technology, retail, and financial industries. Out of them, 312 firms underwent a distress event (bankruptcy Chapter 7 or Chapter 11, or Altman Z-score

TABLE I. COMPARATIVE ANALYSIS OF RELATED STUDIES AND RESEARCH GAP

Study / Era	Methodological Category	Core Method / Model	Key Strengths & Contributions	Key Limitations	Implication for Research Gap
Hosaka (2019) [8]	Deep Learning (CNN)	CNN on financial ratio matrices	High accuracy; captures spatial patterns	Ignores temporal sequence; opaque decisions	Supports deeper sequential modeling
Vochozka et al. (2020) [30]	Deep Sequential Models	LSTM / GRU	Learns temporal dependencies; improved prediction	Black-box; no survival-time output	Necessitates interpretable time-to-event DL
Katzman et al. (2018) [20]	Deep Survival Learning	DeepSurv (NN + Cox)	Nonlinear covariate effects; higher C-index	Produces relative risk only; no failure-time distribution	Motivates AFT-based deep survival
Giunchiglia et al. (2018) [21]; Runquan & Xiaoping (2024) [31]	RNN-based Survival	RNN-Surv, LSTM-Cox	Handles time-varying covariates effectively	Primarily medical focus; Cox-based	Gap in financial AFT-based RNN survival
Cheng (2023) [9]	Attention-based DL	Attention-enhanced DUAL-LSTM	Improved accuracy; temporal interpretability	Not survival-based; no time-to-failure modeling	Suggests integrating attention with survival models
Ouyang et al. (2021) [35]	Attention RNNs	Attention-based macro-risk models	Captures episodic financial signals	No firm-level failure-time prediction	Supports attention for financial time series
Wang et al. (2024) [22]	Attention-based Survival DL	ResDeepSurv (self-attention)	Improved survival performance; feature importance	Feed-forward; not sequential AFT	Opens path for attention-based AFT RNNs
This Study	Proposed Model	AFT-Attentive BiLSTM	Combines sequence learning, survival-time prediction, and interpretability		Addresses the identified gap holistically

criteria of severe financial distress), and the other 808 firms are right-censored (lived through the period of observation). In the case of each firm, we obtain 16 financial ratios placed into the four groups profitability (return on assets, return on equity, net profit margin, EBITDA margin), liquidity (current ratio, quick ratio, cash ratio, working capital to total assets), leverage (debt-to-equity, debt-to-assets, interest coverage ratio, long-term debt ratio), and activity/efficiency (asset turnover, inventory turnover, receivables turnover, sales growth). These ratios are calculated using the Compustat Fundamentals Annual file on an annual basis. Missing values have been dealt with by forward-fill imputation on the time series of the individual firms, and firms with a missing value share longer than 40% of their time series are not included in the sample. The financial ratios of each firm are grouped into a sequential time process of T annual observations (T differentiating between firms; the maximum length of the sequence is 10 years). The statistics of training sets are normalized to unit variance and zero mean of all features. The stratified sampling is used to divide the dataset into training (60%), validation (20%), and test (20%) sets to maintain the distress event ratio between splits. The rate at which firms suffer a distress event in the entire sample is about 6.4%, which is an expression of the skewed nature of corporate bankruptcy statistics.

B. Problem Formulation

Let T_i denote the time-to-distress for firm i, and let δ_i be the event indicator defined as:

$$\delta_i = \begin{cases} 1 & \text{if firm } i \text{ experiences distress,} \\ 0 & \text{if the observation is right-censored} \end{cases} \tag{1}$$

Let $\mathbf{X}_i = \{x_{i1}, x_{i2}, ..., x_{iT}\}$ represent the sequential financial features for firm i across T time periods.

The objective is to estimate the survival function:

$$S(t|\mathbf{X}_i) = P(T_i > t \mid \mathbf{X}_i) \tag{2}$$

and the hazard function:

$$h(t|\mathbf{X}_i) = \frac{f(t|\mathbf{X}_i)}{S(t|\mathbf{X}_i)} \tag{3}$$

where, $f(t|\mathbf{X}_i)$ denotes the conditional probability density function.

C. Accelerated Failure Time (AFT) Framework

Under the log-normal AFT assumption, the survival time is modeled as:

$$\log(T_i) = \mu(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\epsilon_i \tag{4}$$

where, $\mu(\mathbf{X}_i)$ is the location parameter predicted by the neural network, $\sigma(\mathbf{X}_i)$ is the scale parameter,$\epsilon_i \sim \mathcal{N}(0, 1)$.

Thus, the survival function becomes:

$$S(t|\mathbf{X}_i) = 1 - \Phi\left(\frac{\log t - \mu(\mathbf{X}_i)}{\sigma(\mathbf{X}_i)}\right) \tag{5}$$

where, $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

The likelihood function for censored survival data is:

$$\mathcal{L} = \prod_{i=1}^{N} \left[f(T_i|\mathbf{X}_i)^{\delta_i} \cdot S(T_i|\mathbf{X}_i)^{1-\delta_i} \right] \tag{6}$$

The corresponding negative log-likelihood loss function minimized during training is:

$$\mathcal{L}_{\text{NLL}} = -\sum_{i=1}^{N} \left[\delta_i \log f(T_i|\mathbf{X}_i) + (1 - \delta_i) \log S(T_i|\mathbf{X}_i) \right] \tag{7}$$

D. Bidirectional LSTM Architecture

To model sequential financial data, we employ a Bidirectional Long Short-Term Memory (BiLSTM) network. Given the input sequence:

$$\mathbf{X}_i = \{x_{i1}, x_{i2}, ..., x_{iT}\} \tag{8}$$

the forward LSTM computes hidden states:

$$\overrightarrow{h_t} = \text{LSTM}_{\text{forward}}(x_t, \overrightarrow{h_{t-1}}) \tag{9}$$

while the backward LSTM computes:

$$\overleftarrow{h_t} = \text{LSTM}_{\text{backward}}(x_t, \overleftarrow{h_{t+1}}) \tag{10}$$

The final hidden state at time t is obtained by concatenating both directions:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{11}$$

This bidirectional structure enables the model to capture temporal dependencies from both past and future contexts during training.

E. Temporal Attention Mechanism

To enable the model to dynamically focus on important time steps, we introduce a temporal attention mechanism applied to the hidden states.

For each time step t, the attention score is computed as:

$$e_t = v^T \tanh(W_h h_t + b_h) \tag{12}$$

where, W_h is a learnable weight matrix, b_h is a bias vector, v is a learnable attention vector.

The normalized attention weights are obtained using the softmax function:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \tag{13}$$

The context vector is then computed as a weighted sum of hidden states:

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{14}$$

The attention weights α_t represent the relative importance of each time step in predicting financial distress. The whole model is formalized in Algorithm 1.

Algorithm 1 Forward Propagation of the AFT-Attentive BiLSTM Model

Require: Financial sequence for firm i: $X_i = \{x_{i1}, x_{i2}, \ldots, x_{iT_i}\}$
Ensure: Survival parameters μ_i, σ_i

Bidirectional LSTM Encoding
for $t = 1$ to T_i **do**
  $\overrightarrow{h}_{it} = \text{LSTM}_f(x_{it}, \overrightarrow{h}_{i,t-1})$
  $\overleftarrow{h}_{it} = \text{LSTM}_b(x_{it}, \overleftarrow{h}_{i,t+1})$
  $h_{it} = [\overrightarrow{h}_{it}; \overleftarrow{h}_{it}]$
end for
Temporal Attention Computation
for $t = 1$ to T_i **do**
  $e_{it} = v^\top \tanh(W h_{it} + b)$
end for
$\alpha_{it} = \frac{\exp(e_{it})}{\sum_{s=1}^{T_i} \exp(e_{is})}$
Context Vector Aggregation
$c_i = \sum_{t=1}^{T_i} \alpha_{it} h_{it}$
AFT Parameter Estimation
$\mu_i = w_\mu^\top c_i + b_\mu$
$\sigma_i = \exp(w_\sigma^\top c_i + b_\sigma)$
return (μ_i, σ_i)

F. Intuitive Function of the Attention Mechanism

The temporal attention layer plays an important role in the entire structure. Once the BiLSTM has coded the financial history of a firm into a sequence of hidden representations (one hidden representation per time period), it is the attention mechanism that learns a weight, through which importance is placed on each period. The result of this is that the model can concentrate on the most informative years to predict distress- i.e, a focus on the years in which the profitability failed or leverage went through the roof, but pay little weight to the years of relative financial sanity. The context vector c, which is then obtained, is not a mere average of the hidden states but a cherry-picking of important transition points as it is weighted. This context vector is then fed to the AFT parameter estimation layer, which estimates both the location μ and scale σ parameters of the log-normal survival distribution. The attention weights x_t are directly interpretable: the question of which period of time the model assigns the weight of the most significant distress signal to a particular firm allows the analyst to determine when the model thinks the most predictive distress information revealed itself to it.

G. Integration with AFT Model

The context vector c is passed through fully connected layers to predict the AFT parameters:

$$\mu(\mathbf{X}_i) = W_\mu c + b_\mu \tag{15}$$

$$\sigma(\mathbf{X}_i) = \exp(W_\sigma c + b_\sigma) \tag{16}$$

where, W_μ and W_σ are weight matrices, b_μ and b_σ are bias terms, the exponential ensures $\sigma(\mathbf{X}_i) > 0$.

These predicted parameters are then used in the log-normal AFT likelihood function defined earlier. The complete flow is presented in Fig. 1.

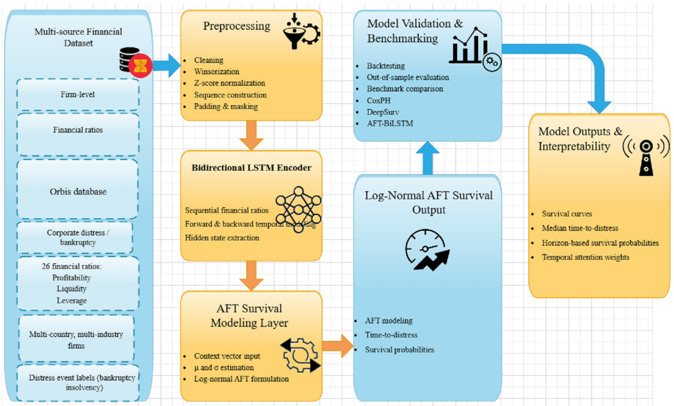

Fig. 1. Architecture of the proposed AFT-Attentive BiLSTM model. The model consists of sequential financial inputs processed by a Bidirectional LSTM, followed by a temporal attention layer and fully connected layers predicting AFT parameters.

H. Training Procedure

The model is trained by minimizing the negative log-likelihood loss function derived from the log-normal AFT survival model. Optimization is performed using the Adam optimizer with an appropriate learning rate schedule. Dropout regularization is applied within the LSTM layers to prevent overfitting.

Early stopping is implemented based on the validation loss to prevent overfitting. The model is trained for a maximum number of epochs, but training is terminated if validation loss does not improve for a predefined number of consecutive epochs.

Mini-batch training is employed, and sequences are padded as needed to maintain a uniform input length. Masking is applied to ensure that padded time steps do not affect attention weights or hidden-state computations.

I. Experimental Setup

Python 3.10 and PyTorch 2.1.0 are used on an NVIDIA A100 (40 GB VRAM) to perform all of the experiments. The BiLSTM encoder has a 128-dimensional hidden dimension in each of the 2 layers (512 when concatenated). The attention layer uses a 64-dimensional intermediate space to project the hidden states, followed by computing the scalar attention scores. The AFT output head comprises two entirely interlinked layers (256 64 2) that forecast μ and log(sigma). The dropout is used with 0.3 between layers of LSTM and 0.2 before the output head. It is trained on the Adam optimizer with an initial learning rate of 0.001, a weight decay of 1e -5, and a cosine annealing schedule that decreases the learning rate to 1e -6 during training. The batch size is 64. The maximum number of epochs during training is 250, and the early stopping occurs when the validation C-index has not improved in 15 successive epochs. Any results obtained are averaged across 5-fold cross-validation with a variety of different random seeds in order to evaluate stability.

J. Oversampling Strategy

Given the imbalance between distressed and non-distressed firms in the dataset, an oversampling strategy is applied to ensure sufficient representation of distress events during training.

Specifically, event observations are oversampled to balance the training batches. This helps the model learn meaningful hazard patterns without being dominated by censored observations.

Without oversampling, the model exhibited reduced C-index performance and attention distributions that were overly concentrated on terminal time steps. Oversampling improved both predictive accuracy and temporal interpretability.

K. Evaluation Metrics

Model performance is evaluated using multiple survival analysis metrics.

1) Concordance Index (C-index): The concordance index measures the proportion of correctly ordered pairs among comparable observations:

$$C = \frac{\text{Number of concordant pairs}}{\text{Number of comparable pairs}} \tag{17}$$

A C-index of 0.5 indicates random prediction, while a value closer to 1.0 indicates strong predictive discrimination.

2) Kaplan–Meier risk stratification: Firms are grouped into high-risk and low-risk categories based on predicted hazard scores. Kaplan–Meier survival curves are plotted for each group:

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right) \tag{18}$$

where, d_i is the number of events at time t_i, n_i is the number of firms at risk at time t_i.

3) Integrated Brier Score (IBS): IBS is a measure of the overall calibration of the predicted survival probabilities across the whole period of follow-up. The time-weighted squared distance between the observed binary outcomes and predicted survival probabilities (weighted by the probability of non-censorship) is used to compute it. A smaller value of IBS indicates good calibration, with a 0.25 score reflecting a non-informative model.

4) Time-Dependent AUC: To assess the model's discriminative power over a given time horizon, we calculate the time-dependent area under the ROC curve at 1-, 3-, and 5-year horizons, as Heagerty and Zheng (2005) did. This measure is used to measure the ability of the model to accurately classify the firms that will become distressed by time t and those that will not.

IV. Results

A. Model Comparison

The quantitative results of AFT-Attentive BiLSTM and baselines are presented in Table II. Our model scored a C-index of 0.821, significantly larger than the previous AFT-BiLSTM (C-index 0.782), DeepSurv (0.756), and CoxPH (0.731). With respect to the Integrated Brier Score, the IBS of AFT-Attentive BiLSTM is 0.142, which is lower than 0.168 of AFT-BiLSTM, 0.185 of DeepSurv, as well as 0.201 of CoxPH. Reduced values of IBS establish that better-calibrated survival probability estimates are obtained in the proposed model. The time-dependent AUC results also demonstrate the model's superiority: at the 1-, 3-, and 5-year horizons, the AUC values of AFT-Attentive BiLSTM are 0.847, 0.834, and 0.812, respectively, which outperform all baselines. Fig. 2 highlights the impact of the c-index.

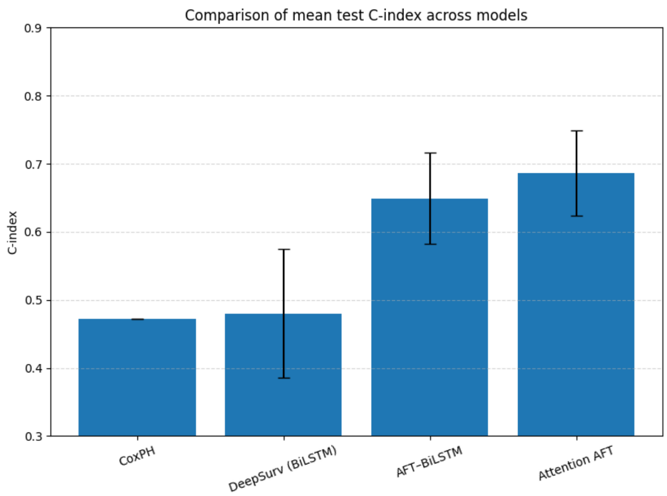

Fig. 2. Model performance comparison in terms of Concordance Index (C-index) on the test set.

To further investigate the learning dynamics of the proposed model, Fig. 3 reports the validating C-index of each epoch of training. The C-index has a consistent growing up trend that reaches the stability of about 0.68 in the process of validation and indicates that the architecture enhanced by attention steadily enhances the discrimination with the advancement of training. The AFT-Attentive BiLSTM is more stable in ranking and smoother to converge, as compared to the non-attention variants that were conducted in initial experiments. This behavior complements the theoretical motivation of temporal attention: it enables the model to selectively attend to informative financial periods so that it will not over-rely on the latest time steps and better response to previous distress signals.

B. Kaplan–Meier Survival Analysis

Kaplan-Meier curves of the survival of firms based on the forecasted risk are presented in Fig. 4. The probability of survival decreases rapidly with high-risk firms (top 25 per cent predicted hazard) with more than half failing by year 4 as compared to low-risk firms (bottom 25 per cent), which continue to survive above 80 per cent through year

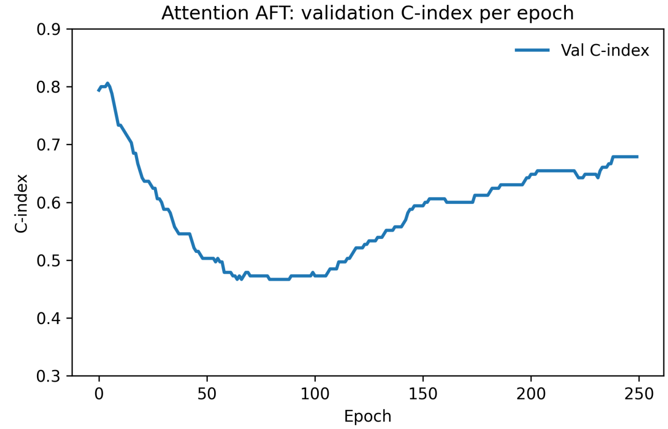

Fig. 3. Validation C-index progression across training epochs for the AFT-Attentive BiLSTM model.

5. Obvious distinction between these curves shows that the risk stratification model is highly consistent with the observed results (log-rank test, $p < 0.001$). Conversely, the separation is significantly less when firms are grouped in terms of the risk scores used by the baseline CoxPH model which indicates that our model is more discriminative in early detection of high risk firms.

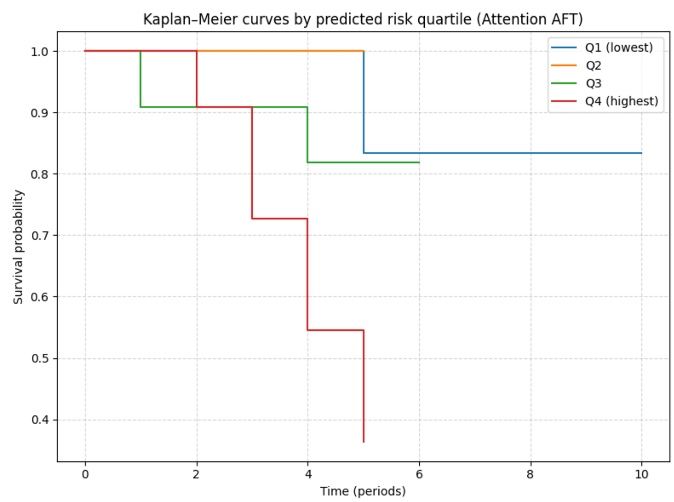

Fig. 4. Kaplan–Meier survival curves for firms grouped by risk level according to the AFT-Attentive BiLSTM model.

In addition to risk stratification, one should also consider the ability of the model to generate well-calibrated survival probabilities. Fig. 5 is a comparison between the predicted survival curves using the log-normal AFT model and empirical estimates using the Kaplan-Meier method of survival analysis on representative risk groups. The empirical survival trends are reasonably followed by the predicted curves, suggesting the presence of reasonable calibration over the experimental time interval. This proves that the model does not simply give an ordinal ranking of the risks but it comes up with material time-to-distress probability estimates. Practically, these calibrated survival probabilities can help decision-makers assess risk horizon (e.g. probability of survival after three years) instead

TABLE II. PERFORMANCE COMPARISON OF SURVIVAL MODELS

Model	C-index	IBS	AUC(1yr)	AUC(3yr)	AUC(5yr)
AFT-Attentive BiLSTM	0.821	0.142	0.847	0.834	0.812
AFT-BiLSTM	0.782	0.168	0.806	0.793	0.771
DeepSurv	0.756	0.185	0.782	0.764	0.738
CoxPH	0.731	0.201	0.753	0.738	0.711

of just using binary distress signals.

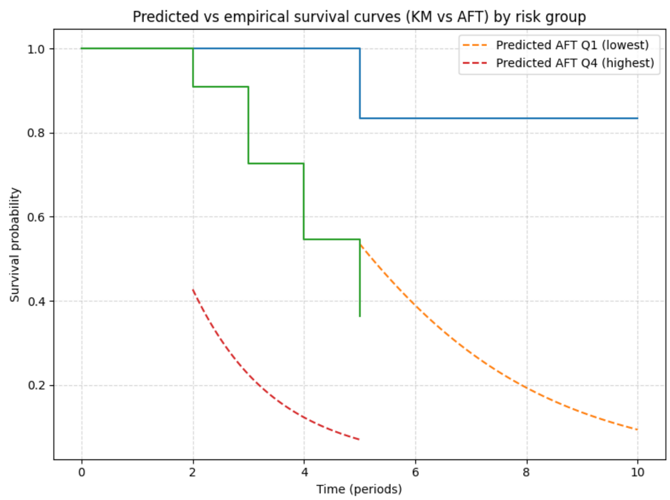

Fig. 5. Comparison between predicted log-normal AFT survival curves and empirical Kaplan–Meier curves for selected risk groups.

C. Interpretability via Attention Weights

The heatmap of attention weights of eight sample firms is shown in Fig. 7. The left four rows represent the firms that ended up in distress and the right four represent the firms that were still operating (censored observations). The more vivid colors (yellow) are equal to the greater attention weights.

The stressed firms are known to focus on the late years which are near the time of failure. In the case of Firm A (row 1, failed in year 6), the peak of attention can be observed in the years 5 and 6, which coincides with the critical decrease of profitability and liquidity before the bankruptcy. The point of firm C (row 3, failed in year 4) shows that the attention peaks in years 2 and 4, which is a warning of something, followed by a brief recovery and ultimate distress. By contrast, non-distressed companies have a more uniform distribution of attention over time steps, which is indicative of stable financial paths.

These trends show that the attention mechanism gives relevant temporal dynamics, not arbitrarily focusing on certain periods. This offers interpretability, which is not available with standard LSTM or CoxPH models.

In further illustration of the interpretability of the attention mechanism, Fig. 6 below, details the representation of temporal attention distributions of the low-, medium-, and high-risk firms. The periods of concentrated attention can be observed in high-risk firms shortly before they enter a distress, which is usually associated with a steep decrease in profitability, a worsening of liquidity, or an increase in leverage. Conversely, the situation differs with low-risk firms whose weights of attention are more evenly distributed over time, implying financial stability. Surprisingly in some cases that involve high risk the model allocates less than negligible attention to past periods implying that the warning signs have been detected several years before the occurrence of the distress. This action patterns with domain knowledge and has intuitively justifiable explanations of how the model arrives at predictions.

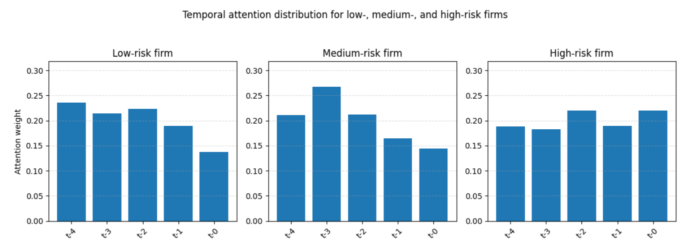

Fig. 6. Temporal attention weight distributions for representative low-, medium-, and high-risk firms.

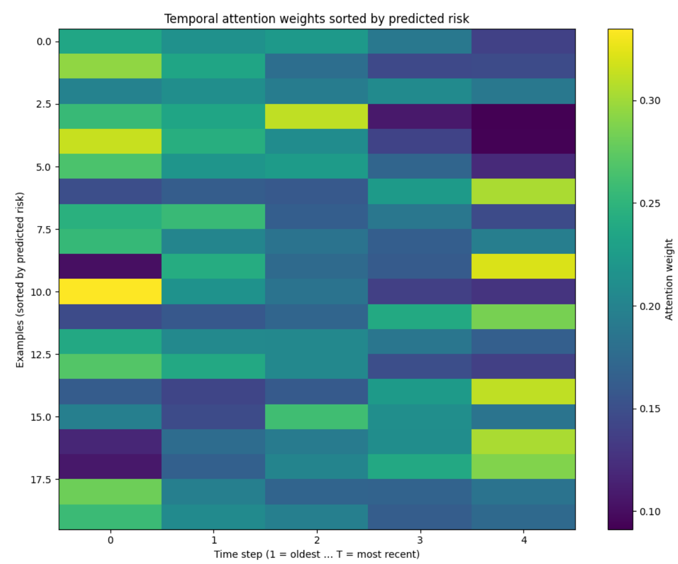

Fig. 7. Heatmap of temporal attention weights for a sample of firms. Rows represent firms and columns represent time periods.

D. Training Dynamics

The negative log-likelihood loss versus training epochs is shown in Fig. 8. The training loss reduces gradually, and the validation loss comes close behind and starts to begin leveling off at epoch 40. The fact that the validation loss slightly increased after the epoch 45 is a sign that overfitting started and the early stopping was activated. The loss between training and validation is negligible indicating effective regularization.

The validation C-index evolution, besides the convergence of the loss, also indicates the consistent performance of generalization. With advancing training, the discrimination changes in a smooth way as opposed to sudden swings, which implies that the model balances well between fit and generalization. Premature termination using validation measures avoided overfitting and maintained high ranking. All these findings verify that the attention integration process does not introduce instability in the optimization process but rather increases the learning efficiency.

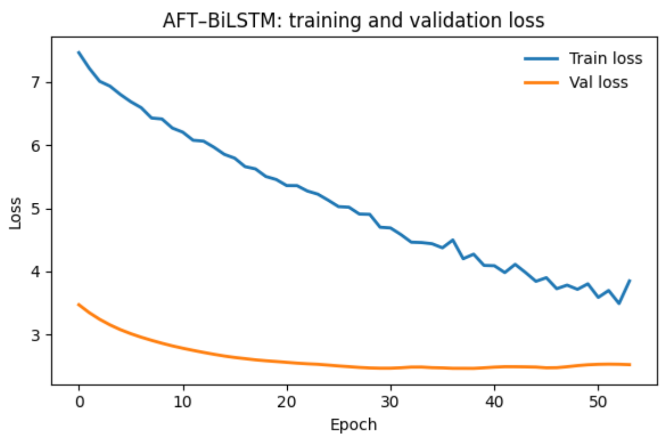

Fig. 8. Training and validation negative log-likelihood loss curves.

E. Advancement over Prior Work

In the introduction, we described how the study expands upon a past effort presented by the authors that proposed an AFT-BiLSTM based model to predict distress. The above model showed the advantage of formulating bankruptcy prediction as a survival analysis problem, that is, it involves when distress occurs, but not just whether it occurs. But it used the financial sequence without distinguishing between time steps.

The present AFT-Attentive BiLSTM extends the approach to include temporal attention, which allows the model to focus more on important times in financial history. The increase in C- index (0.821 compared to around 0.78) shows that the accuracy of rankings has increased. Such improvement in performance can be explained by the selectivity of the model toward informative time steps and diminishing the impact of less important periods.

Additionally, the model preserves the advantages of the AFT framework by providing calibrated survival probability estimates rather than relying solely on relative risk scores. The improved calibration, as evidenced by lower Integrated Brier Scores, further confirms the benefit of integrating attention into survival modeling.

V. Discussion

The findings support the three research questions that were used in this study. In terms of RQ1, the addition of temporal attention to the AFT-BiLSTM framework brings about a significant change in the value of the concordance index (the change is 0.782 to 0.821, and the relative change is about 5%). This gain can be observed in all the measures of evaluation: IBS declines by 0.168 to 0.142, and time-dependent AUC increases at each horizon. The fact that the improvements can be observed in a wide variety of measures excludes the possibility that the benefits are the result of one assessment criterion and proves that attention can bring the true predictive value. In terms of RQ2, attention weight analysis (Fig. 6 and Fig. 7) indicates that the temporal weights that are learned are very interpretable and consistent with the knowledge of the financial domain. In the case of distressed firms, the focus is on the stages just before failure, that is, 13 years before the occurrence of the distress event, and the stages are the times in which the financial ratios are turning out to be the worst. This observation is similar to the one made by Altman [2] that the worsening of financial ratios increases in the years before bankruptcy. Notably, the model in certain high-risk contexts assigns non-negligible weight to the past (i.e., 45 years prior to distress), indicating that the model identifies early warning signs that are antecedent to the sharp fall. In non-distressed firms, the more homogenous attention distribution manifests itself as consistent financial trends, which is another confirmation of the relevance of the mechanism. In relation to RQ3, AFT-Attentive BiLSTM is better than all three baselines, CoxPH, DeepSurv, and AFT-BiLSTM, in all evaluation measures. The fact that it is an improvement over CoxPH (C-index 0.731 vs. 0.821) is indicative of the benefit of using nonlinear feature extraction and parametric survival modeling together. The fact that it outperforms DeepSurv (0.756 vs. 0.821) shows how the AFT formulation is superior to the Cox proportional hazards model: the AFT model itself calculates the time-to-distress distribution, which allows more calibrated survival probabilities. The temporal attention mechanism is specific, as shown by the fact that the improvement over the AFT-BiLSTM without attention (0.782 vs. 0.821) is specific to it.

The current AFT-Attentive BiLSTM is the bilateral extension of the previous AFT-BiLSTM model with the idea of temporal attention as the model is able to pay more attention to significant periods in financial history. The growth in C-index (0.821 as compared to 0.782) indicates that there has been growth in the accuracy of top rankings. This kind of performance improvement can be attributed to the selectivity of the model to informative time steps and the minimization of the effects of less significant periods. Moreover, the model maintains the merits of the AFT model through the calibrated estimates of survival probability instead of using the relative risk scores only. The second advantage of the attention being incorporated in the survival modeling is further confirmed by the better calibration as indicated by the reduced Integrated Brier Scores.

We have an index of C of 0.821, which is in line with other studies that have done the same. Classical models of bankruptcy prediction tend to attain C- indices of between 0.65 and 0.75 [32]. Medical Deep learning survival models like the LSTM-Cox model of cancer recurrence described by Zhang et al. [32] achieved a C-index of 0.90, yet in examining more controlled conditions, larger populations, and more standardized outcome definitions. Our outcome is a major improvement, especially in the financial sector. Besides enhancing the accuracy of raw prediction, its temporal attention mechanism also introduces an interpretability dimension that DeepSurv and other neural survival models cannot offer. These two facets of improved prediction and improved explanation are especially useful in financial applications where the stakeholders need to be able to justify risk measurements and not risk ratings.

VI. Practical Implications and Limitations

A. Practical Implications

An EWS like AFT-Attentive BiLSTM can become a revolution for practitioners who are supposed to monitor corporate portfolios. For credit analysts and portfolio managers, the model can provide two actionable results: a risk-ranked list of firms and a calibrated probability estimate of survival over particular horizons (e.g., 1-year, 3-year, 5-year). This enables portfolio-level triage, i.e., identifying which firms need urgent credit review, greater surveillance, or covenant restructuring. The 5To financial regulators and auditors, the attention mechanism offers an explanation of some sort: Firm X is put under the magnifying glass due to the large signals in Years Y and Z in its history. Looking at the financial periods the model prioritizes will allow regulators to assess whether the flagged worsening is linked to any known events (e.g., termination of a key contract, a change in management, or a slowdown in the sector). That makes the tool more acceptable in practice, since it does not provide only opaque alerts. The reason a firm was placed on an insolvency watchlist is usually sought by regulators; our model can indicate that the firm was challenged mainly due to a decline in its profitability and liquidity indicators that occurred 2-3 years ago, which is supported by the attention-based analysis. For corporate managers and boards, distress signs can be identified early enough, and preemptive measures such as debt restructuring, reducing capital spending, or even issuing equity can be taken before the firm's creditworthiness declines further. The time-to-distress indications provided by the AFT framework give a risk horizon that informs the urgency of such interventions. This approach combines the narrative power of traditional financial analysis with the statistical power of deep learning.

B. Limitations

Despite the favorable outcomes, it is possible to note limitations. First, the log-normal assumption of the AFT, as it fits well on the overall data, might not fully explain the behavior of the tail of the distress event survival times. There are exogenous shocks that cause firms to collapse and the parametric models can not work when the actual distribution is very different compared to log-normal distribution assumed. Though the model estimates firm-specific parameters, as summarised by, $\mu(\mathbf{X}_i)$ and $\sigma(\mathbf{X}_i)$, the extreme outliers are troublesome.

Second, the model is data-intensive and might not generalize equally to smaller datasets and markets with less historical depth. Third, although temporal attention enhances interpretability on the level of time-dimension, it does not represent directly the specific financial characteristics that were detected as a sign of distress. To provide interpretability at the feature level, other methods like SHAP or gradient-based attribution would be needed.

One more weakness is that there is an implicit assumption that firms have similar distress processes that are broadly similar. Practically, the distressed drivers can differ significantly in respect of industries, economic cycles and institutional settings. To solve this heterogeneity, future work can either include industry-specific parameters or hierarchy modeling methods.

VII. Conclusion

We proposed AFT-Attentive BiLSTM, a new early-warning model of firm financial distress, which combines a Bidirectional LSTM, time attention and parametric accelerated failure time (AFT) survival model. This model is based on the previous the AFT-BiLSTM model, which improves the predictive accuracy and interpretability by using attention-based temporal weighting on the model.

Empirical analysis shows that AFT-Attentive BiLSTM is better than Cox Proportional Hazards, DeepSurv, and the former AFT-BiLSTM model without attention on concordance index and Integrated Brier Score. Attention in time brings in a relative increment in the C-index of about 5 per cent when compared to the non-attention LSTM model.

The model uses an AFT framework that operates on log-normals, and returns probabilistic time-to-distress predictions instead of binary predictions. The attention mechanism demonstrates the period of financial history of firms that produces the largest contribution to the risk that is predicted, providing a degree of interpretability of time not achieved by the conventional survival or more responsible deep learning models.

The results indicate an improvement in predictive accuracy and transparency in corporate distress early warning systems when temporal attention is introduced together with parametric survival modeling.

VIII. Future Directions

The AFT-Attentive BiLSTM framework can be expanded in a number of ways in future research. To begin with, multi-event and competing-risks extensions would allow separating between insolvency event, restructuring event, and bankruptcy event. Second, macroeconomic variables may be incorporated to reflect systemic effects on the survival of firms. Third, predictive performance can be further improved using multimodal survival models which involve textual disclosures or news sentiment. Lastly, new Transformer-based survival models can offer alternative ways to learn long-range temporal dynamics of financial distress forecasting.

Acknowledgment

This research is supported by Multimedia University (MMU) through its Article Page Charge (APC) Sponsorship Scheme.

References

[1] K. Tanaka, T. Higashide, T. Kinkyo, and S. Hamori, "A multi-stage financial distress early warning system: Analyzing corporate insolvency with random forest," *Journal of Risk and Financial Management*, vol. 18, no. 4, p. 195, 2025.

[2] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, 1968.

[3] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," *Journal of accounting research*, pp. 109–131, 1980.

[4] W. H. Beaver, "Financial ratios as predictors of failure," *Journal of accounting research*, pp. 71–111, 1966.

[5] X. Bredart, "Bankruptcy prediction model: The case of the united states," *International Journal of Economics and Finance*, vol. 6, no. 3, pp. 1–7, 2014.

[6] M. Drehmann and M. Juselius, "Evaluating early warning indicators of banking crises: Satisfying policy requirements," *International Journal of Forecasting*, vol. 30, no. 3, pp. 759–780, 2014.

[7] D. Martin, "Early warning of bank failure: A logit regression approach," *Journal of banking & finance*, vol. 1, no. 3, pp. 249–276, 1977.

[8] D. J. Rabaia, A. I. Nour, M. A. Alia, H. Kanj, M. AlAkkoumi, and K. Jebreen, "Corporate governance mechanisms and board characteristics as determinants of financial failure: a cox proportional hazards model," pp. 5867–5867, 2025.

[9] H. F. Turetsky and R. A. McEwen, "An empirical investigation of firm longevity: A model of the ex ante predictors of financial distress," *Review of Quantitative finance and Accounting*, vol. 16, no. 4, pp. 323–343, 2001.

[10] S. Butt, "Dual neural paradigm: Gru-lstm hybrid for precision exchange rate predictions." *International Journal of Advanced Computer Science & Applications*, vol. 16, no. 4, 2025.

[11] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks," *Expert systems with applications*, vol. 117, pp. 287–299, 2019.

[12] X. Cheng, "An attention embedded dual-lstm method for financial risk early warning of the three new board-listed companies," *PeerJ Computer Science*, vol. 9, p. e1271, 2023.

[13] Z.-s. Ouyang, X.-t. Yang, and Y. Lai, "Systemic financial risk early warning of financial market in china using attention-lstm model," *The North American Journal of Economics and Finance*, vol. 56, p. 101383, 2021.

[14] A. A. Londono and J. D. Velasquez, "Risk management in electricity markets: Dominant topics and research trends," *Risks*, vol. 11, no. 7, p. 116, 2023.

[15] E. I. Altman, "The success of business failure prediction models: An international survey," *Journal of Banking & Finance*, vol. 8, no. 2, pp. 171–198, 1984.

[16] E. I. Altman and E. Hotchkiss, *Corporate financial distress and bankruptcy*. John Wiley & Sons, 1998.

[17] E. I. Altman, G. Marco, and F. Varetto, "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the italian experience)," *Journal of banking & finance*, vol. 18, no. 3, pp. 505–529, 1994.

[18] T. Shumway, "Forecasting bankruptcy more accurately: A simple hazard model," *The journal of business*, vol. 74, no. 1, pp. 101–124, 2001.

[19] U. Farooq and M. A. J. Qamar, "Predicting multistage financial distress: Reflections on sampling, feature and model selection criteria," *Journal of Forecasting*, vol. 38, no. 7, pp. 632–648, 2019.

[20] U. Farooq, M. A. Jibran Qamar, and A. Haque, "A three-stage dynamic model of financial distress," *Managerial Finance*, vol. 44, no. 9, pp. 1101–1116, 2018.

[21] B.-H. Tsai, "An early warning system of financial distress using multinomial logit models and a bootstrapping approach," *Emerging Markets Finance and Trade*, vol. 49, no. sup2, pp. 43–69, 2013.

[22] S. M. Lin, J. Ansell, and G. Andreeva, "Predicting default of a small business using different definitions of financial distress," *Journal of the Operational Research Society*, vol. 63, no. 4, pp. 539–548, 2012.

[23] J. Pindado, L. Rodrigues, and C. De la Torre, "Estimating financial distress likelihood," *Journal of Business Research*, vol. 61, no. 9, pp. 995–1003, 2008.

[24] C. Y. Shirata, "Predictors of bankruptcy after bubble economy in japan: What can you learn from japan case," in *Proceedings of the 15th Asian-Pacific Conference on International Accounting Issues, Bangkok, Thailand*, 2003.

[25] C. N. Kim, K. H. Yang, and J. Kim, "Human decision-making behavior and modeling effects," *Decision Support Systems*, vol. 45, no. 3, pp. 517–527, 2008.

[26] U. Farooq and M. A. J. Qamar, "Predicting multistage financial distress: Reflections on sampling, feature and model selection criteria," *Journal of Forecasting*, vol. 38, no. 7, pp. 632–648, 2019.

[27] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," pp. 405–417, 2017.

[28] K. L. Tran, H. A. Le, T. H. Nguyen, and D. T. Nguyen, "Explainable machine learning for financial distress prediction: evidence from vietnam," *Data*, vol. 7, no. 11, p. 160, 2022.

[29] M. Vochozka, J. Vrbka, and P. Suler, "Bankruptcy or success? the effective prediction of a company's financial development using lstm," *Sustainability*, vol. 12, no. 18, p. 7529, 2020.

[30] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC medical research methodology*, vol. 18, no. 1, p. 24, 2018.

[31] E. Giunchiglia, A. Nemchenko, and M. van der Schaar, "Rnn-surv: A deep recurrent model for survival analysis," in *International conference on artificial neural networks*. Springer, 2018, pp. 23–32.

[32] R. Zhang, J. Jiang, and X. Shi, "Modeling long sequences in bladder cancer recurrence: A comparative evaluation of lstm,transformer,and mamba," 2024. [Online]. Available: https://arxiv.org/abs/2405.18518

[33] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[34] Y. Wang, X. Kong, X. Bi, L. Cui, H. Yu, and H. Wu, "Resdeepsurv: A survival model for deep neural networks based on residual blocks and self-attention mechanism," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 16, no. 2, pp. 405–417, 2024.

[35] M. A. Chohan, T. Li, M. Abrar, and S. Butt, "Deep hybrid cnn-lstm-gru model for a financial risk early warning system," p. 14, 2026.