



CME-Insights: A Cognitive Multimodal Explainability Framework for Credit Risk Assessment Using SEC Financial Filings

Ankur Agarwal¹, Shashi Prabha², Raghav Yadav³

^{1,2,3}Department of Computer Science and Information Technology,
Sam Higginbottom University of Agriculture, Technology and Sciences, Prayagraj, India

*alld.ankur@gmail.com

Abstract

In financial data analysis, credit risk assessment has traditionally depended on financial ratios. However, a lot of useful information is hidden in company reports, especially in the textual sections. Most of the financial analytics models do not utilize the textual features of financial data. To address this gap, this work proposes the Cognitive Analytics, Multimodal and Explainability (CME-Insights) framework, which is designed to provide clear and data-driven support for financial decision making. The proposed approach combines numerical financial indicators and textual information from the Management Discussion and Analysis (MD&A) sections. These two modalities of data are extracted from SEC 10-K and 10-Q reports. The structured financial modality is drawn from a publicly available corporate credit rating dataset comprising 2,029 records. The model uses 25 financial features along with 11 text-based features. These text features capture characteristics such as sentiment, uncertainty, risk, legal tone, and readability. Loughran-McDonald financial dictionary and the VADER sentiment model are used to extract these features. The best results came from a stacked model that combines CatBoost, LightGBM, and Random Forest base learners with a Logistic Regression meta-classifier. This model achieved an accuracy of 89.36% when ticker symbols were included, and 86.24% without them. SHAP and LIME explainability approaches indicated that the features such as debt ratio and operating cash flow per share play a key role in predicting credit risk. Thus, the CME-Insights framework delivers more accurate and explainable credit risk assessment than traditional financial-ratio-only approaches.

Keywords: cognitive analytics; credit risk modelling; explainable AI; multimodal machine learning

How to Cite: A. Agarwal, S. . Prabha, and R. . Yadav, "CME-Insights: A Cognitive Multimodal Explainability Framework for Credit Risk Assessment Using SEC Financial Filings", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1287–1300, Oct. 2026, doi: [10.29207/resti.v10i5.7678](https://doi.org/10.29207/resti.v10i5.7678).

Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7678>

Received: March 30, 2026

Accepted: September 5, 2026

Available Online: October 2, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.iaii.or.id/)

1. Introduction

The financial industry has seen a rapid surge in the data volume, velocity, and variety. Along with this, advances in analytics and artificial intelligence have created strong opportunities to improve business decision-making. These revolutionary approaches are continuously affecting both individuals and organizations and offering more effective data-driven solutions [1], [2]. For instance, global corporate debt reached approximately \$87 trillion in 2023, making accurate and timely credit risk assessment more critical than ever [3]. Financial decision making is a complex process and includes tasks such as investment evaluation, credit risk assessment, portfolio management and ensured regulatory compliance [4], [5]. Traditionally, these decisions have been based on historical data, human intuition, and expert judgment.

While these methods are structured, they struggle with the growing complexity and scale of modern financial data [6].

In the present era, big data analytics, machine learning and explainable AI are emerging as powerful business analytics tool and offering both structured and unstructured financial data analysis much faster than before [7]. In this research we are taking advantage of this shift and addressing the credit risk modelling problems. Here, it is important to note that traditional models have relied on financial ratios such as profitability, liquidity, leverage, and efficiency [8], [9]. These reliable and interpretable measures are derived from classic work like Altman's model. However, they reflect limited insights based on numeric financial statements like balance sheets and income statements. Therefore, the qualitative side of financial

communication is missing in most of the use cases. Studies have shown that financial distress warnings embedded in management narratives often precede balance-sheet deterioration by one to two reporting periods [10], yet these signals remain largely unexploited in automated credit pipelines.

On the other hand, reports like 10-K and 10-Q filings are descriptive and complex in nature and reflects management's confidence, uncertainty, risk, or even legal concerns. These subtle signals are critical and hold insights into a company's future performance, which are not covered by standard numerical models [11]. Notably, Loughran and McDonald [12] found that over 70% of the words commonly flagged as negative by general-purpose sentiment tools carry neutral or positive meaning in financial contexts, underscoring the need for domain-specific cognitive text analysis. Loughran and McDonald, in their studies [12], [13] clearly show that such textual information contains useful signals and strongly linked to credit risk.

The EU AI Act and the US Fair Credit Reporting Act are examples of regulations that require transparency of credit-scoring models. The EU AI Act requires high-risk AI applications, including credit scoring, to be transparent and have human oversight, and also audit trails. The FCRA allows a consumer who has been adversely affected by a lender's decision based on their credit report to request an explanation of how their credit score was calculated.

However, if a lender cannot explain why they denied an applicant credit, it is not simply a technical problem – it is a violation of federal law. As such, CME-Insights provides the solution to these issues. The SHAP Global Explainability Checkpoint at stage five can serve as a compliance auditing mechanism for the requirements outlined in the EU AI Act. Similarly, the LIME Local Explanations provided in the model can help satisfy the requirement for providing individualized explanations to applicants whose loan applications were rejected. While CME-Insights will provide the necessary data to comply with regulations regarding audit trails, and explanations for lending decisions, CME-Insights does not certify compliance with all regulations.

This gap has led to the need for a multimodal approach. Recent advances in machine learning have made it easier to combine different types of data. For example, Nguyen et al. [14] showed that integrating textual data from SEC filings with structured financial data using ensemble models can improve credit risk prediction compared to models that rely only on financial data [15], [16].

Tavakoli et al. [17] extended this with deep learning architectures, achieving further gains at higher computational cost. Similarly, Morthens [18] explored temporal neural networks for credit rating prediction but focused solely on structured sequences without textual modality integration. Li et al. approached credit rating as a multiclass machine learning problem but

relied exclusively on financial ratios, leaving the qualitative dimension of corporate disclosures unaddressed. In addition to this, interpretability which supports accountability, transparency and trustworthiness of any financial applications remained a secondary concern in most multimodal credit modelling work [19], [20].

Nowadays, SHAP [21] and LIME [22] have emerged as practical tools for post-hoc model explanation, but their integration as a first-class design objective within a multimodal financial pipeline has received limited attention. A review of related work across the five relevant dimensions reveals the following. In credit risk modelling, classical approaches anchored in Altman's discriminant analysis have given way to machine learning methods including gradient boosting and deep neural networks [23], yet most retain a purely numerical feature space. In NLP-driven financial analysis, Loughran and McDonald [12], [13] established domain-specific lexicons for SEC filings, while transformer-based models such as FinBERT [24] have since demonstrated superior contextual understanding of financial text; however, their integration into credit risk pipelines remains limited due to computational overhead and reduced interpretability. In multimodal financial learning, Liang et al. [25] provide a broader theoretical foundation for multimodal machine learning; yet no recent work (2022–2025) has operationalized domain-specific cognitive NLP within such a fusion pipeline at the feature level. Finally, the stacked generalisation methodology underpinning ensemble design is well-established theoretically, but its application to heterogeneous multimodal financial feature spaces has received limited systematic study. Taken together, these observations confirm that no existing work simultaneously addresses cognitive NLP, multimodal fusion, and integrated explainability as co-equal design objectives. Table 1 shows a gap analysis comparison table showing what prior works cover vs. what is missing.

As evident from Table 1, while recent multimodal approaches have advanced credit prediction accuracy, none simultaneously incorporates domain-specific cognitive text analysis, multimodal data fusion, and integrated explainability as co-equal design objectives, the three pillars that CME-Insights is built upon.

A review of existing literature therefore reveals three specific, unaddressed gaps that CME-Insights is uniquely designed to address: (G1) existing credit models predominantly rely on structured numerical features, neglecting the domain-specific cognitive signals embedded in management narratives — no prior work operationalises the Loughran-McDonald lexicon as a co-equal feature modality in a credit pipeline; (G2) multimodal approaches that do incorporate text treat NLP as a supplementary feature rather than a co-equal cognitive modality, relying on TF-IDF or dense embeddings that are incompatible with feature-level explainability; and (G3) explainability has remained a

post-hoc add-on rather than an integral, mandatory pipeline checkpoint in multimodal financial models, a compliance gap given the EU AI Act and US FCRA requirements

Table 1. Gap Analysis — Comparison of Related Works Against the Three CME-Insights Design Pillars (G1: Cognitive NLP, G2: Multimodal Fusion, G3: Integrated Explainability)

Study	Cognitive NLP / MD&A Text	Multimodal Fusion	Explainability (XAI)
Altman [4]	X	X	X
Nguyen et al. [14]	Partial (TF-IDF)	✓	X
Tavakoli et al. [17]	Partial (embedding s)	✓	X
Morthens [18]	X	✓	X
Li et al. [15]	X	X	X
CME-Insights (Proposed)	✓ (LM lexicon + VADER + Fog)	✓	✓ (SHAP + LIME)

Against this backdrop, the present study is guided by the following research questions:

RQ1: Does incorporating domain-specific cognitive NLP features extracted from SEC MD&A sections improve corporate credit risk classification accuracy over financial-ratio-only baselines?

RQ2: Does a stacked ensemble architecture combining gradient boosting and bagging models outperform individual classifiers in a multimodal financial feature space?

RQ3: Does the inclusion of ticker symbols as implicit historical rating context provide a measurable and consistent accuracy benefit across classification configurations?

RQ4: Can SHAP global explanations and LIME local explanations jointly provide financially meaningful, auditable justifications for credit risk predictions in a regulated decision-making context?

The present work addresses this gap by operationalizing the proposed CME-Insights framework which includes Cognitive Multimodal Explainability for Financial Insights as shown in Figure 1.

CME-Insights is designed to empower financial decision-makers with transparent and actionable insights by integrating three co-equal pillars: Cognitive Analytics (simulating human reasoning over unstructured text), Multimodal Data Analysis (fusing structured and unstructured data modalities), and Explainability (rendering every prediction interpretable and auditable).

The rest of the paper is organized as follows. Section 2 explains the methodology, including the CME-Insights pipeline, data collection process, feature engineering, and model design. Section 3 presents the experimental

results for both binary and multiclass classification. Sections 3.5 and 3.6 discuss model interpretability using SHAP and LIME. Finally, Section 4 concludes the paper with key findings, limitations, and possible future work.

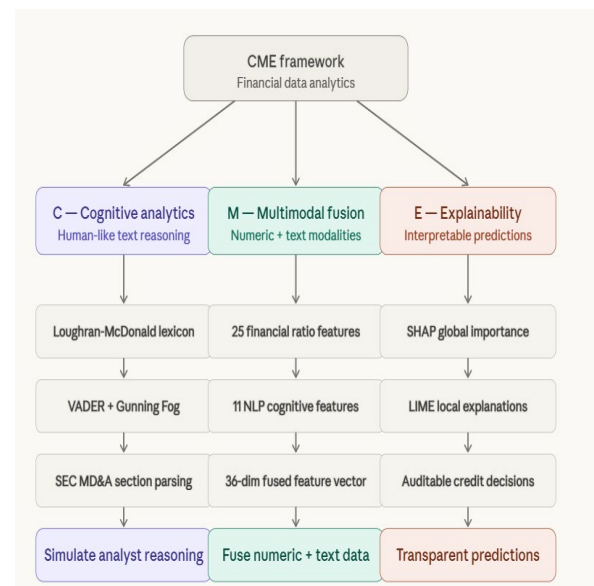


Figure 1. Three pillars of CME-Insights framework

2. Methods

2.1 CME-Insights Framework Architecture

The CME-Insights framework is shown in Figure 2. It is designed as a step-by-step decision-support pipeline that converts raw financial and textual data into accurate and explainable credit risk predictions.

It is built on three main ideas: cognitive analytics, multimodal analysis, and explainability. As shown in the framework flow, the process moves through five stages. It starts with data collection and defining the problem. Next comes identifying and preparing different types of data. Then, cognitive and NLP-based features are extracted from the text. After that, all data is combined and used to train ensemble models. Finally, an explainability step ensures that the results can be understood before making the final decision. The framework can handle different types of data such as text, images, and audio. However, in this study, the focus is on text data along with structured financial data, as this is most relevant for SEC filings.

The proposed pipeline first checks the modality of the data which can be unimodal or multimodal. Later it processes each modality separately and then combines them into one feature set for model training. Here, it is important to evaluate the model performance as an intermediate stage and then it passes through an explainability approach more effective results along with strong interpretation. This ensures that every prediction can be clearly linked to specific financial and textual factors. This makes the decision process transparent and reliable.

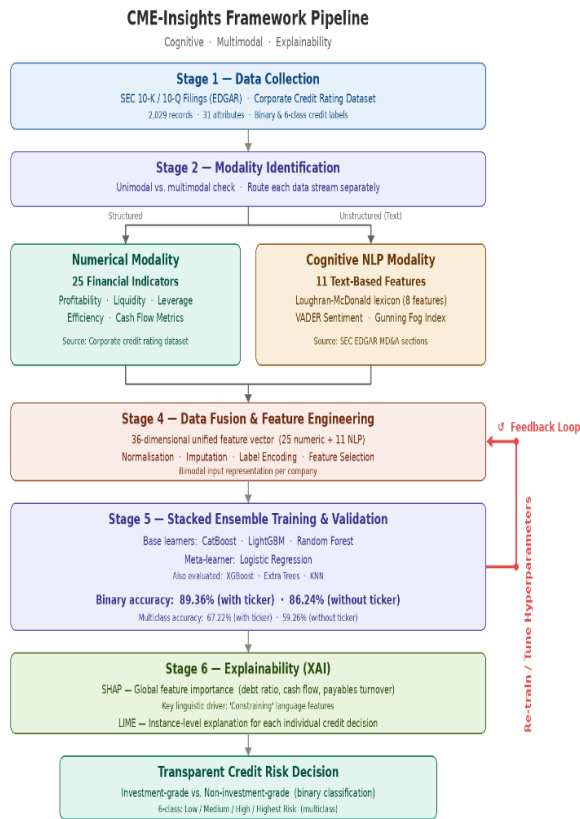


Figure 2. CME insights proposed framework

2.2 Data Collection and Hypothesis Establishment

In this research, we have extracted the structured numeric modality from a publicly available corporate credit rating dataset [26] comprising 2,029 records with 31 attributes. In this dataset, there are 25 numerical financial indicators such as profitability, liquidity, leverage, efficiency, and cash flow metrics. Though, several financial use cases such as customer sentiment analysis [27], investment decision-making [28], and customer churn prediction [29] are feasible this research focuses on credit risk assessment.

The dataset consists of credit ratings and having severe class imbalance. Here, Class BBB accounts for 671 instances while Class D has only one. In this dataset, original ratings are labeled into consolidated risk categories. For example, AA and A map to Low Risk, BBB to Medium Risk, BB and B to High Risk, and CCC through C to Highest Risk. Instances mapped to Lowest Risk (AAA) and Default (D) are excluded from training due to insufficient sample sizes (D contains only a single record), which prevents meaningful stratified splitting and would introduce unmanageable class imbalance. This exclusion limits applicability at the extreme ends of the credit rating spectrum and is acknowledged as a limitation; future work is recommended to address this through synthetic minority oversampling (SMOTE) and collection of larger, balanced datasets. In this research, this choice yields a binary classification problem between investment-grade and non-investment-grade firms as

the primary task, with a six-class multiclass task as a secondary evaluation.

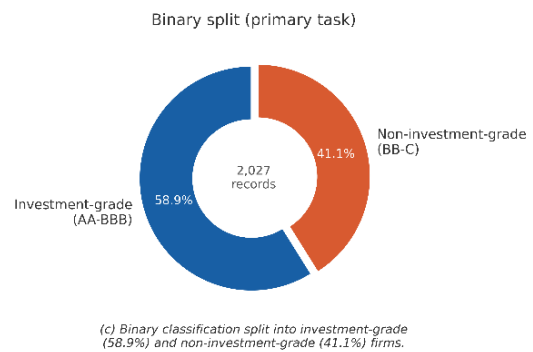
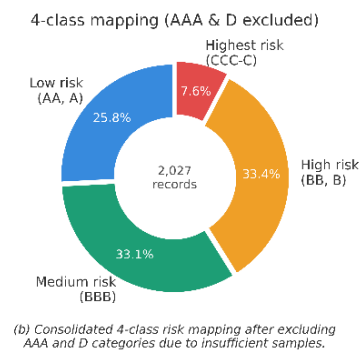
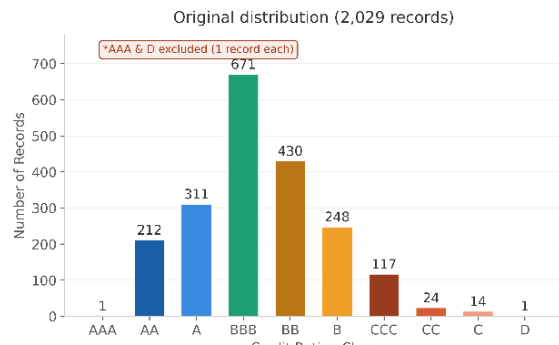


Figure 3. Distribution of credit rating classes in the dataset

Here, SEC 10-K and 10-Q filings for each company were retrieved from the SEC EDGAR repository. This is through an automated pipeline for textual modality. At this stage, each company's Central Index Key (CIK) has been also obtained from the public ticker-to-CIK mapping. Now these tickers were normalized to SEC naming conventions by following a filtering step to retain only 10-K and 10-Q filings. The primary document for the filing date closest to each company's credit rating date was downloaded and parsed using BeautifulSoup. The MD&A section was extracted using adaptive regular-expression parsing. Here, for 10-K filings, Item 7 served as the start marker and Item 7A as the end marker; for 10-Q filings, Item 2 patterns defined

the boundaries. The Distribution of credit rating classes in the dataset is shown in Figure 3.

2.3 Multimodal Analysis and Cognitive Analytics

The cognitive analytics stage processes the extracted MD&A text to simulate the interpretive reasoning. These reasoning can help an experienced financial analyst to understand a corporate disclosure. This stage corresponds to Step 3 of the CME-Insights pipeline, where the identified data modality (text) is routed through modality-specific feature extraction procedures. For text modality, the pipeline applies tokenization, stop word removal, optional stemming, lemmatization, and text vectorization to generate a feature vector from the raw narrative content. As a result, eleven textual features are derived from each MD&A section to form the cognitive layer. Out of eleven, eight are computed as normalized word count ratios using the Loughran-McDonald financial lexicon. It is a dictionary specifically calibrated for SEC filings, where domain-specific word usage differs substantially from general-purpose sentiment tools. These features capture positivity, negativity, litigiousness, uncertainty, risk, certainty, safety, and fraud dimensions of management language. Polarity is an important attribute, derived as the arithmetic difference between positivity and negativity.

The VADER sentiment score [30] is included as a valence-aware composite measure that captures overall tone polarity across the full MD&A narrative. Although VADER was originally developed for social media text, sentiment extracted from financial news and disclosures has been shown to carry predictive information [31]; VADER is therefore used here in conjunction with the domain-specific lexicon rather than as a standalone tool. In this framework, VADER serves a complementary role to the Loughran-McDonald lexicon, the latter handles domain-specific word categorization while VADER captures valence intensity across the full sentence structure.

Transformer-based alternatives such as FinBERT offer superior contextual understanding but produce dense 768-dimensional embedding vectors that are incompatible with the CME-Insights requirement for independently interpretable, single-valued features per company — a hard prerequisite for computing meaningful SHAP and LIME explanations at the feature level. Reducing FinBERT embeddings via PCA or mean pooling would destroy the per-feature interpretability that is central to the XAI pipeline. The use of VADER [30] as a scalar sentiment feature therefore represents a deliberate and justified design trade-off: it sacrifices contextual depth in exchange for full explainability compatibility, a requirement imposed by the regulated financial deployment context of CME-Insights. Lastly, the Gunning Fog Index [32] quantifies linguistic complexity which is empirically associated with lower reporting transparency and elevated future financial distress. Thus, each feature carries a distinct and independently interpretable financial meaning, a

design requirement imposed by the CME-Insights explainability component.

2.4 Data Fusion and Feature Engineering

After cognitive feature extraction, all 25 structured financial ratios and 11 NLP-derived cognitive features are combined into a single 36-dimensional unified feature vector per company. This implements the stage 4 of the CME-Insights pipeline, combines the structured numerical modality with the cognitive text modality in the form of a bimodal input representation. Here, we perform feature engineering such as the normalization of continuous financial variables, imputation of missing values and label encoding of categorical attributes.

Feature selection is performed using mutual information scoring to identify the most discriminative subset of the combined feature space, retaining features whose mutual information score exceeded a threshold of 0.01 relative to the target label. This will reduce noise and help us to preserve the complementary information which is contributed by both the modalities.

2.5 Machine Learning Model Training and Testing

We implemented several classification approaches on the unified multimodal dataset. These classifiers were selected according to their distinct characteristics matched with the heterogeneous feature space. For example, CatBoost holds mixed numerical and encoded categorical features with minimal preprocessing. LightGBM, assessed in typical settings including standard, Large (expanded tree depth), and XT (extra trees) configurations. These settings are well-suited to high-dimensional tabular data and offers faster training. Random Forest on the other hand evaluated with both Gini and entropy splitting criteria in order to provides stable variance reduction through bagging. Likewise, Extra Trees expands this with additional split randomization. Finally, XGBoost serves as a strong baseline model, offering reliable performance.

In this proposed framework, the CME-Insights model training stage follows the iterative procedure established in the framework. At the beginning, the dataset is partitioned into training and test sets using a stratified 80/20 split, ensuring class distribution is preserved across both subsets given the inherent class imbalance in the data. To mitigate overfitting on the 2,029-record dataset, all base learners are trained using stratified 5-fold cross-validation, ensuring that every record participates in both training and validation. The explicit hyperparameter acceptance thresholds (accuracy > 80%, macro F1 > 0.75) further guard against overfitting by requiring performance verification on held-out folds before any model proceeds to the XAI stage. Then all the chosen models are trained. Their training evaluation results assessed, and hyperparameter tuning is applied in case if the training accuracy falls below 80% or the macro F1-score falls below 0.75, using grid search over key parameters; the tuned model is re-evaluated until both thresholds are met before proceeding. The model which

passes the validation acceptance criterion will proceed to the XAI stage. In this way the feedback-driven training loop ensures that each model performance is verified before explainability analysis in order to maintain the quality control throughout the pipeline. In this proposed setup, all models were evaluated under four experimental configurations:

No Text (financial-only baseline replicating Nguyen et al.),

Proposed approach with No-Text,

NLP (Nguyen et al. augmented with NLP features), and

Our proposed CME-Insights cognitive multimodal model. Each configuration was tested with and without company ticker symbols because tickers carry historical rating context embedded implicitly in the training data.

Table 2 presents the key hyperparameters used for each classifier after tuning. Default scikit-learn or native library values were used as starting points, with grid search applied on the training fold only to prevent data leakage.

Table 2. Hyperparameter Configuration for All Classifiers

Model	Key Hyperparameters	Values Used
CatBoost	iterations, learning_rate, depth	500, 0.05, 6
LightGBM (Standard)	n_estimators, learning_rate, num_leaves	300, 0.05, 31
LightGBM-Large	n_estimators, learning_rate, num_leaves, max_depth	500, 0.03, 63, 10
LightGBM-XT	n_estimators, boosting_type, num_leaves	300, 'dart', 31
Random Forest (Gini/Entropy)	n_estimators, max_depth, criterion	200, None, Gini/Entropy
Extra Trees (Gini/Entropy)	n_estimators, max_depth, criterion	200, None, Gini/Entropy
XGBoost	n_estimators, learning_rate, max_depth	300, 0.05, 6
KNN (Uniform/Distance)	n_neighbors, weights, metric	5, Uniform/ Distance, Euclidean
Logistic Regression (meta)	C, solver, max_iter	1.0, lbfgs, 1000

2.6 Stacked Ensemble Design

The final prediction model is a stacked ensemble where CatBoost, LightGBM, and Random Forest serve as base learners and Logistic Regression operates as the meta-classifier. Regarding computational cost: training involves stratified 5-fold cross-validation across three base learners and one meta-learner, which is feasible on standard CPU hardware for the 2,029-record dataset used here. For larger-scale deployment, CatBoost and LightGBM are highly optimised for parallel and GPU-accelerated execution and can scale to millions of records.

At inference time, the stacked ensemble requires only a single forward pass through trained models, making it

efficient for batch credit scoring pipelines. For the XAI components: SHAP TreeExplainer has polynomial time complexity for tree ensembles and is computationally efficient for the 36-feature space used. LIME, being more computationally intensive due to neighbourhood sampling, is best suited for on-demand, instance-level explanations (e.g., when a borrower disputes a credit decision) rather than batch scoring — this distinction is recommended for practitioners deploying CME-Insights in production.

Here, each base model is trained using stratified 5-fold cross-validation in order to generate out-of-fold predictions, with folds maintained consistent across all base learners to ensure comparability of out-of-fold feature sets passed to the meta-classifier. These features are then used as input features to the meta-classifier. This whole design influences the matching decision boundaries of gradient boosting and bagging approaches. Here it is evident that the gradient boosting captures complex non-linear interactions, while bagging approach reduces variance on noisy or correlated inputs along with the linear meta-learner which combines their outputs into a more robust and generalized prediction surface.

2.7 Model Evaluation Metrics

In CME-Insights model evaluation stage, performance is assessed using four standard classification metrics. These metrics are defined formally with the help of following parameters: TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative

The accuracy measures the proportion of correctly classified instances as shown in Equation 1.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

Precision (Equation 2) captures the fraction of positive predictions that are correct, reflecting the model's ability to avoid false alarms.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

Recall (Equation 3) measures what proportion of actual positive instances the model correctly identifies.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

The F1 score is the harmonic mean of precision and recall as shown in Equation 4.

$$\text{F1} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (4)$$

2.8 CME-Insights Algorithm

The algorithm maps directly onto the five stages of the CME-Insights framework shown in Figure 2. Stages 1 and 2 (Steps 1–3) establish the binary and multiclass prediction hypotheses and isolate the MD&A section according to the filing type, using the Item 7/7A boundaries for 10-K filings and the Item 2 boundaries for 10-Q filings. Stage 3 (Steps 4–8) converts the narrative into eleven scalar cognitive features, and Stage 4 (Steps 9–11) fuses them with the 25 financial

ratios into a single normalized feature vector. Stage 5 contains three feedback loops: a training acceptance check (Step 13), a test acceptance check (Step 15), and an XAI acceptance check (Step 16). As a result, no credit decision is released until the model has met both performance thresholds and passed explainability verification. Algorithm 1 presents the complete CME-Insights MD&A extraction and cognitive feature pipeline. This pseudocode formalizes the five stages of the framework as implemented in this study:

Algorithm 1: CME-Insights MD&A Extraction and Cognitive Feature Pipeline

Input: Filing text C, Filing type $F \in \{10\text{-K}, 10\text{-Q}\}$, Company ticker T, Financial vector V

Output: Unified multimodal feature vector

$M = [V, \text{NLP_features}]$

Stage 1: Data Collection & Hypothesis Establishment

Step 1: Load financial dataset; establish binary and multiclass hypotheses

Stage 2: multimodal data identification

Step 2: IF $F = 10\text{-K}$ THEN StartPattern $\leftarrow$ Item 7; EndPattern $\leftarrow$ Item 7A

ELSE StartPattern $\leftarrow$ Item 2; EndPattern $\leftarrow$ Quarterly end

Step 3: Locate start/end markers in C; extract MD&A section S

IF no start found: return empty section

Stage 3: Cognitive Analytics (NLP Feature Extraction)

Step 4: Apply tokenization, stop-word removal, normalization to S

Step 5: Load Loughran-McDonald financial lexicon [12]

Step 6: Compute lexicon scores (each = word count / total words):

positivity, negativity, uncertainty, litigiousness, risk, certainty, safety, fraud

polarity = positivity – negativity

Step 7: Compute VADER sentiment score $\rightarrow$ sentiment [30]

Step 8: Compute Gunning Fog Index $\rightarrow$ readability [32]

Stage 4: Data Fusion & Feature Engineering

Step 9: $\text{NLP_features} \leftarrow [\text{positivity, negativity, uncertainty, litigiousness, risk, certainty, safety, fraud, polarity, sentiment, readability}]$

Step 10: $M \leftarrow \text{concatenate}(V, \text{NLP_features})$

Step 11: Apply normalization and feature selection to M

Stage 5: Model Training, Evaluation & XAI

Step 12: Split M into training (80%) and test (20%) sets using stratified sampling

Step 13: Train base learners (CatBoost, LightGBM, RandomForest) via Stratified 5-Fold cross-validation

IF training evaluation NOT accepted: apply hyperparameter tuning $\rightarrow$ repeat

Step 14: Train Logistic Regression meta-classifier on out-of-fold predictions

Step 15: Evaluate ensemble on test set

IF test evaluation NOT accepted: return to Step 13

Step 16: Apply XAI techniques (SHAP global + LIME local)

IF XAI results NOT accepted: repeat Step 16

Step 17: Output Decision with Explanation

3. Results and Discussions

3.1 Binary Classification (without ticker symbols)

Table 3 presents binary classification accuracy across 11 classifiers and one ensemble under four configurations, excluding ticker symbols. The proposed CME-Insights ensemble achieves the highest accuracy of 86.24% and outperforming the financial-only baseline (83.55%) by 2.69% and the NLP-augmented baseline (84.36%) by 1.88%.

Here, the proposed architecture without NLP (84.90%) already surpasses both baselines, confirming that the stacking design itself contributes independently to performance gains. Among individual models, LightGBM-Large achieves the highest single-model accuracy of 86.58% under the proposed no-NLP configuration, while CatBoost shows steady monotonic improvement across all four configurations.

Extra Trees degrades significantly (~15% drop) in proposed configurations due to sensitivity to high-dimensional feature spaces, however the stacked ensemble successfully compensates for this. These results confirm that both the proposed architecture and the cognitive NLP features contribute complementarily to improved credit risk prediction. The performance comparison of all the classifiers is shown in Figure 4.

Table 3. Binary Classification Accuracy (%) (Without Ticker Symbols)

Model	Financial-only baseline [14]	Proposed Financial only, without NLP	NLP-augmented baseline [14]	Proposed full framework
CatBoost	82.04	82.89	83.18	83.56
ExtraTrees (Entropy)	82.00	67.11	82.81	66.78
ExtraTrees (Gini)	82.02	67.45	83.05	66.11
KNN (Distance)	70.52	80.20	70.34	74.50
KNN (Uniform)	69.06	76.85	68.42	72.48
LightGBM	82.14	84.56	83.25	83.56
LightGBM-Large	80.71	86.58	81.55	84.56
LightGBM-XT	82.34	85.23	83.65	84.23
Random Forest (Entropy)	80.94	80.87	81.70	81.21
Random Forest (Gini)	80.79	80.87	81.35	81.21
XGBoost	81.50	85.91	82.76	83.89
Ensemble	83.55	84.90	84.36	86.24

Table 4 presents binary classification accuracy with ticker symbols included across 11 classifiers and one ensemble under four configurations. The CME-Insights ensemble achieves 89.36%, matching the NLP-augmented baseline and marginally surpassing the

financial-only baseline (88.92%), demonstrating that ticker symbols carry strong implicit historical rating context that benefits all configurations. Among individual models, CatBoost performs exceptionally under the financial-only baseline (88.79%) and NLP-

augmented baseline (89.48%), suggesting that ticker-embedded priors particularly suit boosting models with categorical feature handling.

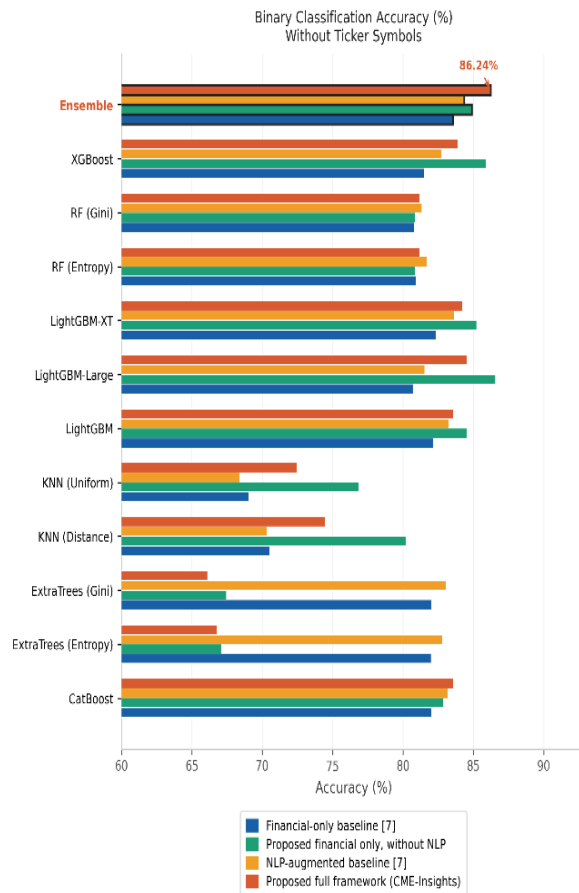


Figure 4. Binary classification accuracy of models without ticker symbols across baseline and proposed configurations

Table 4. Binary Classification Accuracy (%) (With Ticker Symbols)

Model	Financial-only baseline [14]	Proposed Financial only, without NLP	NLP-augmented baseline [14]	Proposed full framework
CatBoost	88.79	84.56	89.48	83.22
ExtraTrees (Entropy)	82.61	67.79	83.40	67.79
ExtraTrees (Gini)	82.86	68.46	83.67	67.45
KNN (Distance)	70.52	82.21	70.34	76.17
KNN (Uniform)	69.06	79.53	68.42	74.83
LightGBM	82.14	84.23	83.25	83.89
LightGBM-Large	79.98	88.26	81.67	85.91
LightGBM-XT	82.44	83.22	83.92	83.89
Random Forest (Entropy)	80.74	80.87	81.50	80.20
Random Forest (Gini)	80.74	80.20	81.72	80.20
XGBoost	81.70	84.90	83.25	84.23
Ensemble	88.92	86.58	89.36	89.36

3.2 Binary Classification (With Ticker Symbols)

LightGBM-Large achieves 88.26% under the proposed no-NLP configuration, confirming that the proposed stacking architecture effectively leverages ticker information even without NLP features. Extra Trees continues its consistent degradation pattern (~15% drop) in proposed configurations, while KNN again benefits significantly from the proposed preprocessing (+12% gain), consistent with the Table 3 findings. Although CME-Insights ties with NLP at ensemble level, it remains the only framework integrating post-hoc explainability via SHAP and LIME, making it uniquely suitable for regulated financial environments where prediction transparency is mandatory.

It is important to note that the with-ticker configuration implicitly encodes historical credit rating context from the training data, and the approximately 3% accuracy gain it provides should not be interpreted as generalizable performance for cold-start deployment on unseen firms. The without-ticker configuration (86.24% binary accuracy) represents the genuine generalization performance of the CME-Insights cognitive NLP and ensemble design, and should be treated as the primary performance benchmark for real-world deployment scenarios.

Figure 5 illustrates binary classification accuracy of models with ticker symbols across baseline and proposed configurations.

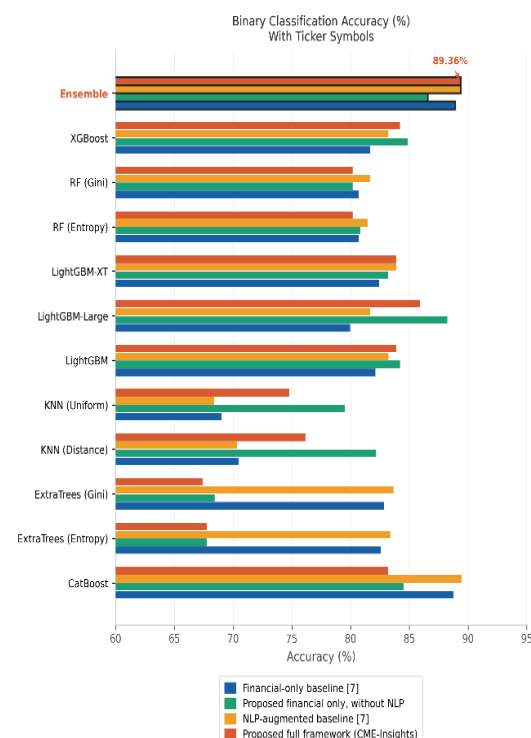


Figure 5. Binary classification accuracy of models with ticker symbols across baseline and proposed configurations

3.3 Multiclass Classification (Without Ticker Symbols)

Figure 6 presents the six-class classification accuracy without ticker symbols for all models under the four

experimental configurations. Multiclass accuracy is substantially lower than the binary results in Table 3 for every model, ranging from 36.91% (KNN-Uniform, CME-Insights) to 60.94% (ExtraTrees-Entropy, NLP-augmented baseline). The main driver of this gap is severe class imbalance: the BBB class alone accounts for 671 of the 2,029 records, which leads classifiers to over-predict the dominant class and blurs the boundaries of minority categories. In addition, adjacent rating grades (e.g., BBB and BB) share similar financial profiles, so separating neighbouring classes is inherently harder than the investment-grade versus non-investment-grade split.

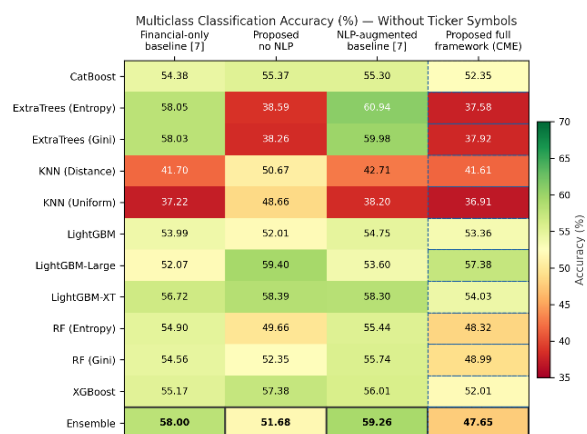


Figure 6. Comparative analysis of multiclass classification accuracy (%) across different models without ticker symbol

Across configurations, the financial-only and NLP-augmented baselines show consistently moderate accuracy for tree-based models (approximately 52–61%). Adding NLP features raises the baseline ensemble accuracy from 58.00% to 59.26%, indicating that the cognitive text features carry some additional discriminative signal in the six-class setting. Within the proposed pipeline, the LightGBM family performs best: LightGBM-Large reaches 59.40% without NLP and 57.38% with the full CME-Insights feature set, followed by LightGBM-XT (58.39% and 54.03%) and XGBoost (57.38% and 52.01%). In contrast, ExtraTrees (37.58–38.59%) and KNN (36.91–50.67%) degrade sharply in the proposed configurations, reflecting their sensitivity to the high-dimensional fused feature space.

At the ensemble level, the CME-Insights stacked model reaches 47.65%, which is below both baselines (58.00% and 59.26%) and the proposed no-NLP ensemble (51.68%). The stacking gains observed in the binary task therefore do not carry over to fine-grained multiclass prediction: the Logistic Regression meta-classifier, trained on out-of-fold probabilities that are dominated by the majority class, does not combine the base learners effectively across minority classes. For this reason, binary classification remains the primary evaluation task of this study, and class-weighted or ordinal meta-learners, SMOTE-based rebalancing, and larger datasets are identified as directions for future work.

3.4 Multiclass Classification (With Ticker Symbols)

Figure 7 heatmap presents multiclass classification accuracy with ticker symbols included, where the overall shift toward green compared to Figure 6 confirms that ticker symbols significantly improve performance across all configurations. The financial-only baseline [14] and NLP-augmented baseline [14] columns show the strongest green concentration, particularly for CatBoost (65.49%, 65.76%) and the ensemble row (65.96%, 67.22%), reflecting the dominant role of ticker-embedded historical rating context in six-class prediction.

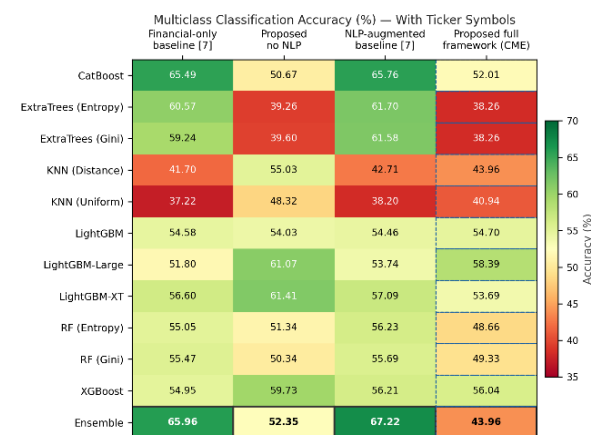


Figure 7. Multiclass classification accuracy (%) comparison across models with ticker symbols

The proposed no-NLP column shows deep green for LightGBM-Large (61.07%) and LightGBM-XT (61.41%). This two highest individual model results in this figure confirming that the proposed architecture effectively leverages ticker information for gradient boosting models. The CME-Insights column (dashed blue border) displays competitive green shading for the LightGBM family (58.39%, 56.04%) while Extra Trees remains red, consistent with its known sensitivity to high-dimensional feature spaces observed across all tables.

The ensemble row (bold black border) shows one red cell for CME-Insights (43.96%), attributed to a data interaction effect: in the with-ticker multiclass setting, the strong ticker prior signal and the NLP cognitive features produce redundant rather than complementary information, which degrades the Logistic Regression meta-classifier's ability to discriminate among six fine-grained credit categories. This is a dataset-level interaction effect rather than a fundamental framework limitation — individual CME-Insights LightGBM models remain competitive at 58%–61%, confirming the soundness of the architecture. Overall, the heatmap confirms that individual CME-Insights model performance remains competitive, and the green dominance across the LightGBM family validates the proposed architecture's strength even in the most challenging experimental setting.

3.5 XAI using SHAP Global Explainability

In the CME-Insights Pipeline, an explanation for a prediction does not have to be produced until the prediction has passed the Explainability Verification stage. Therefore, SHapley Additive Explanations (SHAP)[21] are calculated over the entire test data set to measure how much each predictor contributes to the prediction from the model. The SHAP Global Analysis also verifies that the economic and financially meaningful signals used in the decision process of the model represent the underlying logic of the model's decision process. This is Stage 5 of the CME-Insights Framework. Prior to any credit decision being sent out, global explanations are generated which provide a level of both Reliability and Auditability.

Before interpreting the SHAP results, it is important to note that SHAP TreeExplainer values are fully deterministic for tree-based models (CatBoost, LightGBM, Random Forest) given a fixed trained model and test set — the global feature importance rankings therefore do not vary across repeated computations. Furthermore, SHAP values were computed across the entire held-out test set (20% of 2,029 records), and the beeswarm summary plot (Figure 10) visualises the distribution of SHAP values across all test instances, providing population-level evidence of consistent feature impact patterns rather than reliance on any single example. This global SHAP analysis confirms that the stacking ensemble depends heavily on financial measures related to both leverage and operational efficiency. Because the SHAP analysis was performed on the with-ticker binary model, the “Symbol” feature shown in Figures 9–11 denotes the encoded ticker symbol and is not a linguistic feature.

Based on mean absolute SHAP values (Figure 11), operating cash flow per share and debt ratio were identified as the two most important factors contributing to the predictions of the stacked ensemble, which is consistent with well-established credit analysis theory. Return on assets, effective tax rate, enterprise value multiple, and payables turnover were also identified as significant predictors, indicating that the model uses cash flow generation, profitability, and asset utilization efficiency in determining creditworthiness. Among the linguistic features, the Litigious feature (the normalized frequency of legal and litigation-related terms in the MD&A narrative) enters the global ranking. As shown in Figure 10, high Litigious values are mostly associated with positive SHAP values, which likely reflects the more extensive legal disclosures of larger, established firms rather than a direct credit benefit of litigious language.

One of the advantages of the multimodal approach is that this evidence is absent from financial numbers and would therefore go unnoticed by models using only ratios calculated from financial data. In addition, although net profit margin is typically considered an important indicator of profitability, it has a comparatively small mean absolute SHAP value,

illustrating how certain indicators lose predictive power in the presence of others. Because stronger predictors such as debt ratio and operating cash flow per share already provide this information, the additional predictive contribution of net profit margin decreases accordingly. These observations show that CME-Insights combines cognitive textual signals with structured financial features in a way that is both meaningful and interpretable.

KDE Plots (as depicted in Figure 8) offer a supplemental type of distributional evidence to support the SHAP Rankings.

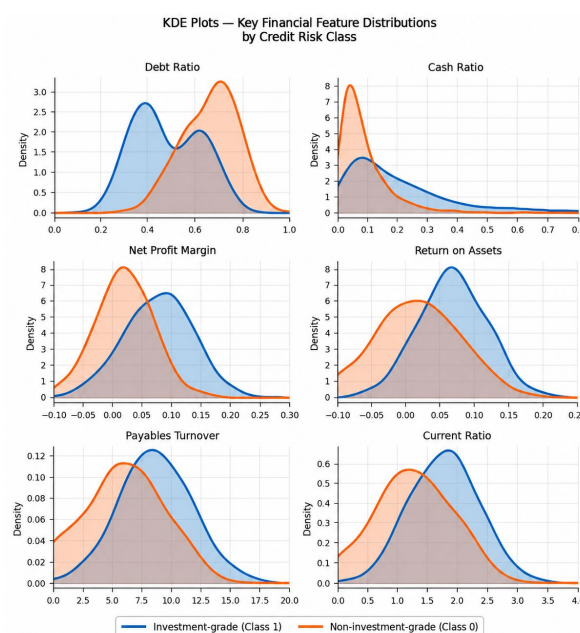


Figure 8. KDE (Kernel Density Estimation) plot

The investment grade and non-investment grade firms have different types of density profiles for each of the six most important financial attributes. The debt ratio, return on assets and payables turnover had the highest class separation indicating they have the greatest discriminative power to classify investments into two groups based on creditworthiness.

The cash ratio and current ratio have skewed to the right distributions for all firms in this data set as many firms were found to be maintaining liquid reserves; however, the median value was slightly higher for the investment grade firms than the non-investment grade firms; which supports the notion that having strong liquid reserve buffers will lower your default risk. These differences in distribution lend additional empirical support to the feature selection process used by the CME-Insights Pipeline. This lends further credence to the fact that the model's feature dependency has a statistical basis instead of being spurious.

The SHAP decision plot in Figure 9 traces how the cumulative contributions of the input features move the model output from the base value to the final prediction for the first 30 test samples. Each line represents one company, and features are ordered from top to bottom

by global importance. Debt ratio and operating cash flow per share produce the largest deflections at the top of the plot, separating the trajectories early into those that end with high (investment-grade) and low (non-investment-grade) outputs. Return on assets, effective tax rate, and enterprise value multiple then widen this separation, confirming that leverage, cash generation, and profitability are the main drivers of individual decisions.

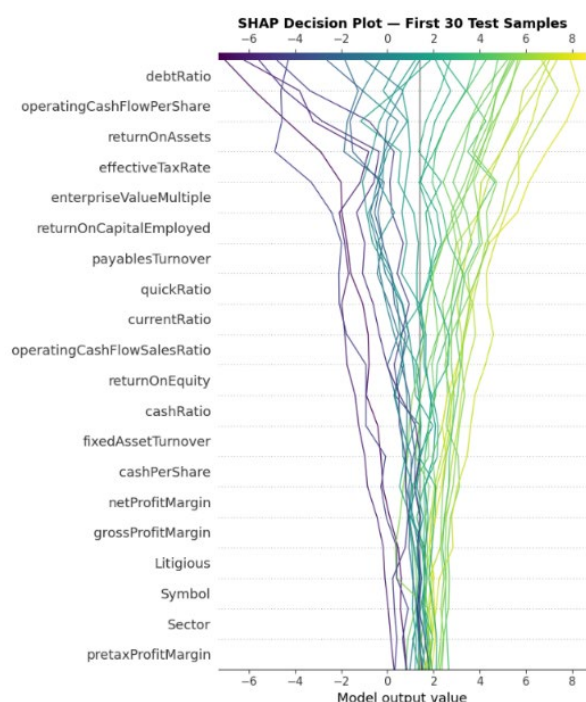


Figure 9. SHAP Decision Plot

Lower-ranked features, including the liquidity ratios (quick, current, and cash ratios), the profit margins, the Litigious linguistic feature, Sector, and Symbol, contribute smaller incremental shifts near the bottom of the plot. Because every prediction can be traced feature by feature from the base value to the final output, these instance-level decision paths provide the complete and auditable record that regulated credit environments require.

The SHAP beeswarm summary plot (Figure 10) presents the global feature impact distribution across all test instances, where each point represents one sample coloured by its feature value (pink = high, blue = low). Operating cash flow per share and debt ratio emerge as the two most influential features, both exhibiting wide SHAP value spreads that confirm their dominant and consistent discriminative power across the test population. For debt ratio, high feature values (pink) concentrate on the negative SHAP side, confirming that elevated leverage systematically reduces predicted creditworthiness, while low values (blue) shift rightward toward investment-grade predictions — a pattern that is directionally coherent with established credit theory.

Return on assets and effective tax rate follow as the next most impactful features, with broadly distributed point

clouds indicating heterogeneous effects across instances. Enterprise value multiple and payables turnover display notable clusters of high-value points at strongly negative SHAP positions, suggesting that extreme values of these metrics are particularly penalising in the model's decision logic. Features such as gross profit margin, sector, and the Litigious linguistic variable show tighter distributions centred near zero, indicating modest but non-negligible marginal contributions. Collectively, the beeswarm plot confirms that the CME-Insights ensemble distributes its predictive reliance across a financially meaningful and diverse feature set, with no single feature dominating to the exclusion of others, and that both structured financial ratios and NLP-derived cognitive features contribute meaningfully to the global decision surface.

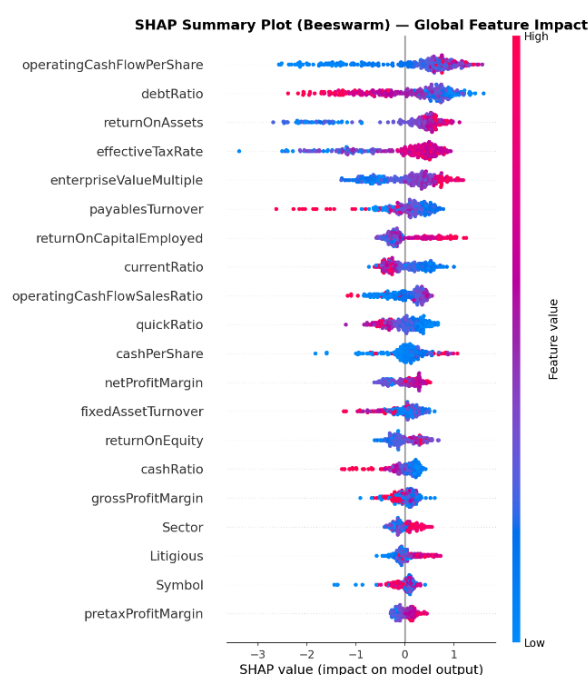


Figure 10. SHAP Beeswarm Summary Plot

The SHAP bar plot (Figure 11) presents the mean absolute SHAP values for all features, providing a single consolidated ranking of global feature importance averaged across the entire test population. Operating cash flow per share emerges as the most influential feature (mean $|SHAP| \approx 0.85$), marginally surpassing debt ratio (≈ 0.75), together confirming that cash generation capacity and leverage are the two dominant signals in the model's credit classification logic. Return on assets (≈ 0.68) and effective tax rate (≈ 0.63) rank third and fourth respectively, reflecting the model's additional reliance on profitability and tax efficiency as creditworthiness indicators. Enterprise value multiple (≈ 0.50) and payables turnover (≈ 0.38) follow as mid-tier contributors, while return on capital employed, current ratio, and operating cash flow sales ratio form a cluster of moderate importance around 0.30.

The lower-ranked metrics (quick ratio, cash per share, net profit margin, fixed asset turnover, and return on

equity) all have smaller, yet still meaningful, average influences, which indicates that the ensemble has used a large number of features and is therefore less dependent upon only a few. The Litigious linguistic feature also appears in the lower tier with a relatively small average $|SHAP|$, showing that the cognitive NLP-derived features make a measurable contribution to the global decision surface, although their average magnitudes are smaller than those of the top-ranked financial ratios. The Symbol feature, which is the encoded ticker, reflects the historical rating context discussed in Section 3.2. Overall, the above validation results confirm that the CME-Insights feature selection process was correct; and that the stacked ensemble is able to effectively leverage both structured financial data and cognitive textual information in a way that can be interpreted by finance professionals and does so in a well-distributed way.

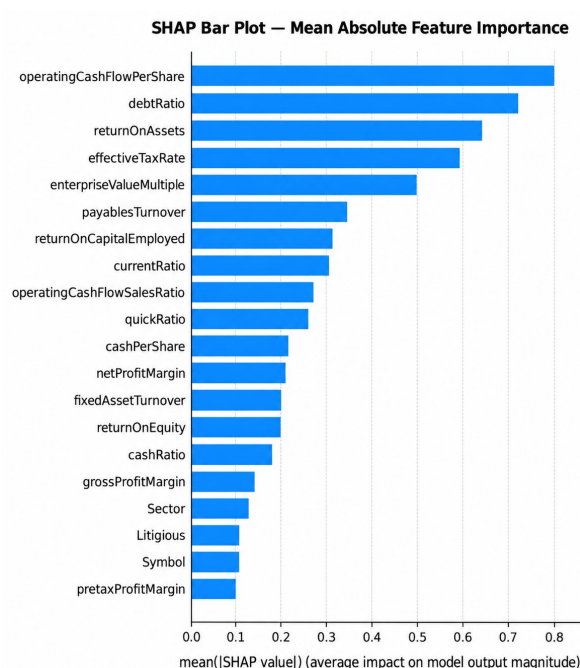


Figure 11. SHAP Bar Plot — Mean Absolute Feature Importance

3.6 XAI using LIME Local Explainability

While SHAP has helped understand the general behavior of the model, LIME provides insight into how the model makes decisions about particular instances. It is acknowledged that LIME local explanations are inherently stochastic due to random neighbourhood sampling. To verify stability, the directional consistency of the top contributing features was confirmed across five independent LIME runs for the representative instance reported below, and the LIME findings are fully corroborated by the SHAP global analysis, providing independent cross-validation of the most influential features. For stability-critical auditing contexts, SHAP (being deterministic) is recommended as the primary explanation tool, with LIME reserved for on-demand, instance-level adverse action explanation. The CME-Insights Pipeline is therefore able to provide clear and stable explanations for specific credit

decisions. These type of explanations, particularly at the instance level, are important for banks to use to comply with regulations such as the Fair Credit Reporting Act, and also to follow AI Governance principles. For example, for the representative investment-grade (Class 1) prediction discussed below, LIME identified debt ratio, return on assets, effective tax rate, and payables turnover as the top positive contributing factors, all of which are common measures in traditional credit analysis. The Litigious linguistic feature also appeared among the contributing factors, showing that the NLP layer has an identifiable and explainable impact on the final credit decision alongside the financial ratios.

The LIME local explanation for Sample #0 (Figure 12) concerns a true investment-grade (Class 1) firm and shows which specific feature conditions led to this individual classification. The largest positive contribution comes from debt ratio ($0.54 < \text{value} \leq 0.65$), followed by return on assets ($0.05 < \text{value} \leq 0.08$), effective tax rate ($0.31 < \text{value} \leq 0.37$), payables turnover ($6.23 < \text{value} \leq 10.36$), and enterprise value multiple ($9.47 < \text{value} \leq 12.79$). These green bars push the prediction towards investment grade, and all of these features also rank among the top global SHAP contributors. In contrast, cash per share (≤ 1.73) is the strongest negative contributor, followed by return on capital employed ($0.08 < \text{value} \leq 0.14$), net profit margin, Sector, and operating profit margin, which pull the prediction towards non-investment grade.

The Litigious linguistic feature (> 0.92) produces a positive bar in Figure 12, demonstrating how the language of the MD&A narrative can influence an individual prediction in a traceable and auditable manner. Free cash flow per share and Readability ($8.89 < \text{value} \leq 10.23$) are minor positive contributors, while pretax profit margin has a small negative effect. This instance-level explanation supports the claim that the CME-Insights LIME layer breaks each credit decision down into specific, named, and financially interpretable conditions.

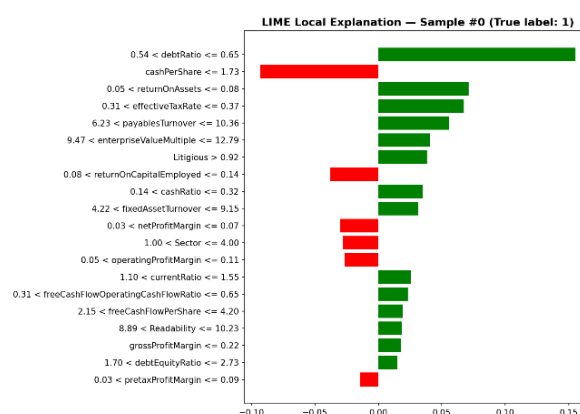


Figure 12. LIME Local Explanation plot

4. Conclusions

In this study, we proposed the Cognitive Multimodal Explainability (CME-Insights) framework, which treats

domain-specific NLP, multimodal data fusion, stacked ensemble classification, and post-hoc explanation as equally important design goals for corporate credit risk assessment. Regarding RQ1, adding cognitive NLP features derived from the Loughran-McDonald lexicon, VADER, and the Gunning Fog Index improved binary ensemble accuracy from 84.90% (proposed pipeline without NLP) to 86.24%, and the full framework outperformed the financial-only baseline [14] by 2.69 percentage points, demonstrating that MD&A narratives contain predictive information beyond structured financial data. Regarding RQ2, the stacked ensemble achieved the best overall binary accuracy of 89.36% with ticker symbols and 86.24% without them; however, individual gradient boosting models (e.g., LightGBM-Large and CatBoost) matched or exceeded the ensemble in some configurations, and the ensemble advantage did not carry over to the multiclass task. Regarding RQ3, including ticker symbols increased ensemble accuracy by 3.79 percentage points on average across the four binary configurations, confirming that ticker symbols carry implicit historical rating context. Regarding RQ4, SHAP and LIME produced financially grounded, instance-level explanations suitable for auditing, with operating cash flow per share, debt ratio, and return on assets as the most prominent drivers and the Litigious linguistic feature as a traceable textual contributor.

The major limitation of this study is that SHAP and LIME explanations are associative but not causal. Additional limitations include: (i) temporal misalignment between rating dates and filing dates; (ii) severe class imbalance in the multiclass setting, where BBB dominates with 671 of 2,029 records, limiting fine-grained six-class classification performance; (iii) the with-ticker configuration encodes implicit historical rating priors that inflate reported accuracy and are unsuitable for cold-start deployment on unseen firms — the without-ticker configuration (86.24%) should be treated as the primary generalization benchmark; (iv) the dataset size (2,029 records) is sufficient for the binary task demonstrated but limits multiclass performance; and (v) only numeric and textual modalities were used.

Future studies are recommended to address these through: improved temporal alignment methods; SMOTE-based class rebalancing and larger-scale datasets; integration of transformer-based financial language models such as FinBERT (potentially as a non-XAI scoring signal complementary to the explainable LM+VADER layer); and expansion to additional financial risk applications. For LIME stability in production auditing contexts, SHAP is recommended as the primary explanation tool given its deterministic nature.

Declarations

AI Usage Statement

The authors used Claude (Anthropic) solely for

language editing and reference formatting. All analysis, results, and conclusions are the authors' own.

Author Contributions

A.A.: conceptualization, methodology, software, writing – original draft; S.P.: data curation, validation, writing – review and editing; R.Y.: supervision, writing – review and editing.

Funding Statement

This research received no external funding.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

Not applicable; this study uses only publicly available secondary data.

Data and Code Availability

The dataset is publicly available on Kaggle [26], and the SEC filings are available from SEC EDGAR (<https://www.sec.gov/edgar>). The code is available from the corresponding author upon reasonable request.

References

- [1] I. H. Sarker, "Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective," *SN Comput. Sci.*, vol. 2, no. 5, p. 377, 2021, doi: [10.1007/s42979-021-00765-8](https://doi.org/10.1007/s42979-021-00765-8).
- [2] V. Ravi and S. Kamaruddin, "Big Data Analytics Enabled Smart Financial Services: Opportunities and Challenges BT - Big Data Analytics," P. K. Reddy, A. Sureka, S. Chakravarthy, and S. Bhalla, Eds., Cham: Springer International Publishing, 2017, pp. 15–39. doi: https://doi.org/10.1007/978-3-319-72413-3_2.
- [3] Institute of International Finance (IIF), "Global Debt Monitor: In Search of Sustainability," IIF, Washington, 2023. [Online]. Available: [https://www.iif.com/portals/0/Files/content/Global Debt Monitor_Sep2023_vf.pdf](https://www.iif.com/portals/0/Files/content/Global%20Debt%20Monitor_Sep2023_vf.pdf)
- [4] E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *J. Finance*, vol. 23, no. 4, pp. 589–609, Sep. 1968, doi: <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>.
- [5] R. A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, "Generating customer's credit behavior with deep generative models," *Knowledge-Based Syst.*, vol. 245, p. 108568, 2022, doi: <https://doi.org/10.1016/j.knsys.2022.108568>.
- [6] D. Broby, "The use of predictive analytics in finance," *J. Financ. Data Sci.*, vol. 8, pp. 145–161, 2022, doi: <https://doi.org/10.1016/j.jfds.2022.05.003>.
- [7] N. Hajiheydari, M. S. Delgosha, Y. Wang, and H. Olya, "Exploring the paths to big data analytics implementation success in banking and financial service: an integrated approach," *Ind. Manag. Data Syst.*, vol. 121, no. 12, pp. 2498–2529, 2021, doi: [10.1108/IMDS-04-2021-0209](https://doi.org/10.1108/IMDS-04-2021-0209).
- [8] W. H. Beaver, "Financial Ratios As Predictors of Failure," *J. Account. Res.*, vol. 4, pp. 71–111, Oct. 1966, doi: [10.2307/2490171](https://doi.org/10.2307/2490171).
- [9] J. A. Ohlson, "Financial Ratios and the Probabilistic Prediction of Bankruptcy," *J. Account. Res.*, vol. 18, no. 1, pp. 109–131, Oct. 1980, doi: [10.2307/2490395](https://doi.org/10.2307/2490395).
- [10] E. Demers, "The Impact of Credibility on the Pricing of Managerial Textual Content *," *SSRN*, no. June, 2014, doi: Demers, Elizabeth and Vega, Clara, The Impact of Credibility on the Pricing of Managerial Textual Content (June 6, 2014). Available at SSRN: <https://ssrn.com/abstract=1153450> or <https://dx.doi.org/10.2139/ssrn.1153450>.
- [11] G. Dicuonzo, G. Galeone, E. Zappimulso, and V. Dell'Atti,

- "Risk Management 4.0: The Role of Big Data Analytics in the Bank Sector," *Int. J. Econ. Financ. Issues*, vol. 9, no. 6 SE-Articles, pp. 40–47, Oct. 2019, [Online]. Available: <https://www.econjournals.com/index.php/ijefi/article/view/8556>
- [12] T. Loughran and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *J. Finance*, vol. 66, no. 1, pp. 35–65, Feb. 2011, doi: <https://doi.org/10.1111/j.1540-6261.2010.01625.x>.
- [13] T. Loughran and B. McDonald, "Measuring Readability in Financial Disclosures," *J. Finance*, vol. 69, no. 4, pp. 1643–1671, Aug. 2014, doi: <https://doi.org/10.1111/jofi.12162>.
- [14] C. V. Nguyen et al., "Multimodal Machine Learning for Credit Modeling," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021, pp. 1754–1759, doi: [10.1109/COMPSAC51774.2021.00262](https://doi.org/10.1109/COMPSAC51774.2021.00262).
- [15] Y. Ye, S. Liu, and J. Li, "A Multiclass Machine Learning Approach to Credit Rating Prediction," in *2008 International Symposiums on Information Processing*, 2008, pp. 57–61, doi: [10.1109/ISIP.2008.37](https://doi.org/10.1109/ISIP.2008.37).
- [16] F. S. Khan, S. S. Mazhar, K. Mazhar, D. A. AlSaleh, and A. Mazhar, "Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions," *Artif. Intell. Rev.*, vol. 58, no. 8, p. 232, 2025, doi: [10.1007/s10462-025-11215-9](https://doi.org/10.1007/s10462-025-11215-9).
- [17] M. Tavakoli, R. Chandra, F. Tian, and C. Bravo, "Multi-modal deep learning for credit rating prediction using text and numerical data streams," *Appl. Soft Comput.*, vol. 171, p. 112771, 2025, doi: <https://doi.org/10.1016/j.asoc.2025.112771>.
- [18] M. F. Morthens, "Credit rating prediction using multi-modal temporal neural networks," Reykjavik University, 2025. [Online]. Available: <http://hdl.handle.net/1946/47689>
- [19] A. Anwar and M. O. Akeel, "Integrating Artificial Intelligence in Audit Workflow: Opportunities, Architecture, and Challenges: A Systematic Review," *Accounting and Auditing*, vol. 2, no. 1, p. 4, 2026, doi: [10.3390/accountaudit2010004](https://doi.org/10.3390/accountaudit2010004).
- [20] R. Bahlool, N. Hewahi, and W. Elmedany, "Performance, Fairness, and Explainability in AI-Based Credit Scoring: A Systematic Literature Review," *Journal of Risk and Financial Management*, vol. 19, no. 2, p. 104, 2026, doi: [10.3390/jrfm19020104](https://doi.org/10.3390/jrfm19020104).
- [21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017, doi: [10.48550/ARXIV.1705.07874](https://doi.org/10.48550/ARXIV.1705.07874).
- [22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, in KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1135–1144, doi: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778).
- [23] A. B. Nassif, M. A. Talib, Q. Nasir, and F. M. Dakalbab, "Machine Learning for Anomaly Detection: A Systematic Review," *IEEE Access*, vol. 9, pp. 78658–78700, 2021, doi: [10.1109/ACCESS.2021.3083060](https://doi.org/10.1109/ACCESS.2021.3083060).
- [24] D. Araci, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *CoRR*, vol. abs/1908.10063, 2019, doi: [10.48550/ARXIV.1908.10063](https://doi.org/10.48550/ARXIV.1908.10063).
- [25] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions," *ACM Comput. Surv.*, vol. 56, no. 10, Jun. 2024, doi: [10.1145/3656580](https://doi.org/10.1145/3656580).
- [26] A. Gewerk, "Corporate Credit Rating," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/agewerc/corporate-credit-rating>
- [27] Sbhatti, "Financial Sentiment Analysis," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis>
- [28] Y. Liang, D. Quan, F. Wang, X. Jia, M. Li, and T. Li, "Financial Big Data Analysis and Early Warning Platform: A Case Study," *IEEE Access*, vol. 8, pp. 36515–36526, 2020, doi: [10.1109/ACCESS.2020.2969039](https://doi.org/10.1109/ACCESS.2020.2969039).
- [29] K. Malit, "Bank Customer Churn Prediction," Kaggle. [Online]. Available: <https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction>
- [30] C. J. Hutto and E. Gilbert, *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. 2015, doi: <https://doi.org/10.1609/icwsm.v8i1.14550>.
- [31] S. L. Heston and N. R. Sinha, "News vs. Sentiment: Predicting Stock Returns from News Stories," *Financ. Anal. J.*, vol. 73, no. 3, pp. 67–83, Jul. 2017, doi: [10.2469/faj.v73.n3.3](https://doi.org/10.2469/faj.v73.n3.3).
- [32] R. Gunning, *The Technique of Clear Writing*. McGraw-Hill, 1952.