Company Bankruptcy Prediction Model

*

Aryan Gupta
School Of Computer Engineering
Kalinga Institute Of Industrial Technology
Bhubhaneswar, India
2228017@kiit.ac.in

Dattatreyo Basu
School Of Computer Engineering
Kalinga Institute Of Industrial Technology
Bhubhaneswar, India
2228022@kiit.ac.in

Parijat Roy
School Of Computer Engineering
Kalinga Institute Of Industrial Technology
Bhubhaneswar, India
2228125@kiit.ac.in

Kanish Srivastava
School Of Computer Engineering
Kalinga Institute Of Industrial Technology
Bhubhaneswar, India
2228173@kiit.ac.in

***Abstract*—Predicting corporate bankruptcy is crucial for risk management and financial stability. This project leverages machine learning to develop a predictive model for identifying companies at risk of bankruptcy. Using financial indicators, market trends, and historical data, we employ algorithms such as Logistic Regression, Random Forest, and XGBoost to enhance prediction accuracy. The model aims to assist investors, regulators, and businesses in making informed decisions, reducing financial risks, and improving early intervention strategies.**

***Index Terms*—machine learning, bankruptcy predicition, XGBoost, Random Forest**

I. Introduction

The prediction of corporate bankruptcy has been a critical area of study in financial research and practice for decades. Accurately forecasting financial distress allows investors, regulators, and policymakers to take preemptive measures to mitigate economic risks. Traditional bankruptcy prediction models, such as the **Altman Z-Score** (Altman, 1968) and **Ohlson O-Score** (Ohlson, 1980), have been widely utilized for their simplicity and interpretability. These statistical approaches rely on financial ratios and linear discriminant analysis, offering reasonable predictive accuracy in stable economic conditions. However, their effectiveness diminishes in dynamic and complex financial environments, especially in **emerging economies** such as **India**, where business cycles, regulatory frameworks, and corporate governance structures differ significantly from those in developed nations.

With **rising corporate debt**, increased economic **uncertainty**, and **regulatory changes** such as the **Insolvency and Bankruptcy Code (IBC) of 2016**, the need for robust and adaptable bankruptcy prediction models has never been more pressing. Early identification of financially distressed firms not only aids **investors** in risk assessment but also assists **lenders** in making informed credit decisions. Furthermore, regulators and policymakers can leverage such predictive models to design interventions that prevent **systemic financial crises** and stabilize the economy.

Identify applicable funding agency here. If none, delete this.

While global research has explored various machine learning (ML) techniques for bankruptcy prediction, studies specifically tailored to **Indian companies** remain scarce. Given India's **unique financial landscape**, which includes **highly leveraged firms**, **a growing startup ecosystem**, and **sector-specific risks**, applying advanced ML techniques can enhance predictive accuracy beyond traditional models.

This study aims to bridge this research gap by leveraging **machine learning algorithms** to predict corporate bankruptcy in India. We develop and evaluate multiple ML models, comparing their predictive performance using financial statement data. The results provide insights into the effectiveness of data-driven approaches in enhancing bankruptcy prediction, ultimately contributing to improved **financial risk management** and **decision-making processes**.

II. Literature Review

A. Traditional Approaches In Bankruptcy Prediction

Bankruptcy prediction models have evolved significantly over the past few decades, transitioning from traditional statistical techniques to more advanced machine learning methodologies. **Early models**, such as the **Altman Z-Score** (1968), were among the first quantitative approaches used for financial distress prediction. The Z-Score model applied **Multivariate Discriminant Analysis (MDA)** to financial ratios such as liquidity, profitability, and leverage to classify firms as bankrupt or non-bankrupt. While this approach demonstrated **reasonable accuracy** in stable economic conditions, it relied on **linear assumptions**, making it less effective in capturing the complexities of modern financial environments.

Recognizing the need for more flexible models, **Ohlson's O-Score** (1980) introduced **logistic regression**, incorporating **non-linear relationships** into bankruptcy prediction. The O-Score model expanded the feature set by including firm-

Altman Z-score	Meaning of the cut-off points
Z > 2.67	Non-distress Zones
1.81 <Z < 2.67	Grey Zones
Z < 1.81	Distress Zones

Source: Adapted from "Business Bankruptcy Prediction Models" by Anjum, 2012, p. 216

Fig. 1. Model Of Altman Z-Score

specific characteristics, such as size, financial structure, and market variables. However, it remained **constrained by statistical assumptions**, such as independence among predictors and a fixed functional form, limiting its applicability to dynamic and evolving financial systems.

Over time, researchers explored additional statistical models to enhance predictive accuracy. **Logit and Probit Models** gained popularity due to their ability to estimate bankruptcy probabilities based on a firm's financial attributes. These models improved upon MDA by addressing some of its statistical limitations, particularly in handling binary classification problems. However, their reliance on predefined relationships between variables still posed challenges in adapting to **real-world financial complexities**, where bankruptcy is influenced by a multitude of **macro- and microeconomic factors** beyond just financial ratios.

With rapid advancements in **information technology** and the proliferation of **big data**, a wealth of information on both **financial and non-financial indicators** is now accessible. Traditional models primarily utilized **static financial ratios** over a specific period, often overlooking key external factors such as **macroeconomic conditions, industry trends, corporate governance, and market sentiment**. Moreover, financial distress is rarely caused by a **single factor**—it is a cumulative process influenced by multiple interdependent variables, including **management decisions, regulatory changes, and external shocks**. As a result, while traditional approaches laid the foundation for bankruptcy prediction, their limitations in capturing the **complex, multi-dimensional nature of financial distress** necessitated the shift towards **machine learning and data-driven models**.

B. Modern Machine Learning Approaches

The advent of **machine learning (ML)** has revolutionized bankruptcy prediction by introducing more sophisticated and data-driven approaches. Unlike traditional statistical models, ML techniques can capture **non-linear relationships, high-dimensional interactions, and hidden patterns** within financial and non-financial data. Various ML algorithms, including **Logistic Regression, Support Vector Machines (SVM), Random Forests, XGBoost, and Neural Networks**, have been widely explored for their predictive capabilities.

1, Supervised Learning Models

Among ML techniques, **Logistic Regression (LR)** continues to serve as a baseline model due to its interpretability and probabilistic output. However, its linear decision boundary limits its ability to handle complex relationships in financial distress data. **Support Vector Machines (SVM)** improve upon LR by finding an optimal hyperplane for classification, making it particularly effective in **high-dimensional feature spaces**. However, its computational cost can be high, especially for large datasets.

2. Ensemble Methods:

Recent studies have demonstrated that **ensemble learning techniques**, such as **Random Forests and XGBoost**, significantly enhance bankruptcy prediction accuracy. **Random Forests**, an aggregation of multiple decision trees, help **reduce overfitting and improve model stability** by averaging multiple weak classifiers. **XGBoost**, an optimized gradient boosting algorithm, further refines predictions by sequentially learning from errors, making it a popular choice for financial classification problems. These models **outperform traditional methods** in terms of adaptability, feature importance ranking, and handling missing data.

3. Deep Learning & Neural Networks:

With the growth of **deep learning**, **Artificial Neural Networks (ANNs) and Recurrent Neural Networks (RNNs)** have gained traction in bankruptcy prediction. These models can learn **complex, non-linear dependencies** in financial data, making them particularly effective in capturing **time-series trends** and patterns that influence corporate failure. Although deep learning models achieve high predictive accuracy, they often suffer from **lack of interpretability**—a crucial factor in financial decision-making.

III. Methodology

A. Data Sourcing And Preprocessing

The data for this model has been sourced from a verified dataset of Taiwanese Economic Journal. The dataset encompasses companies across various sectors, including manufacturing, services, and technology, to ensure robustness and generalizability. The data includes balance sheets, profit and loss accounts, cash flow statements, and additional financial disclosures from annual reports.

The collected data undergoes extensive preprocessing to ensure quality and consistency. Missing data is addressed through imputation techniques, while outliers are detected using the Interquartile Range (IQR) method and handled through capping or transformation. To enhance model performance, financial ratios and key metrics are calculated, including liquidity ratios, solvency ratios, profitability ratios, and efficiency metrics. Data normalization and scaling are performed to ensure uniformity across features.

B. Imbalance In Dataset

On exploration of the data, it was found that the dataset was highly imbalanced with a majority of data points belonging to the Non-Bankrupt category.

To address this, we employed the **Synthetic Minority Oversampling Technique (SMOTE)**, a widely used resampling method that generates synthetic samples for the minority class rather than merely duplicating existing instances.

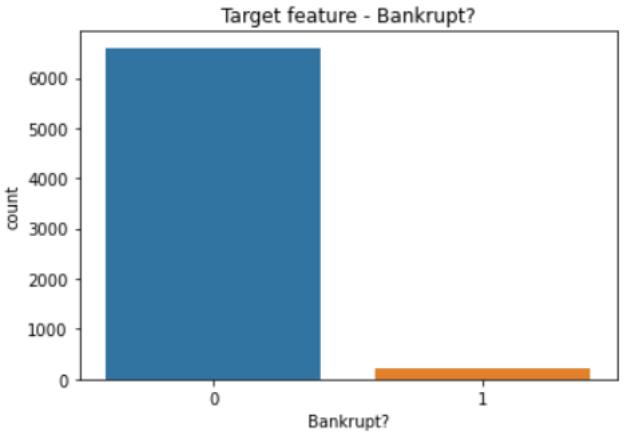

Fig. 2. Count Plot portraying the class imbalance in the dataset

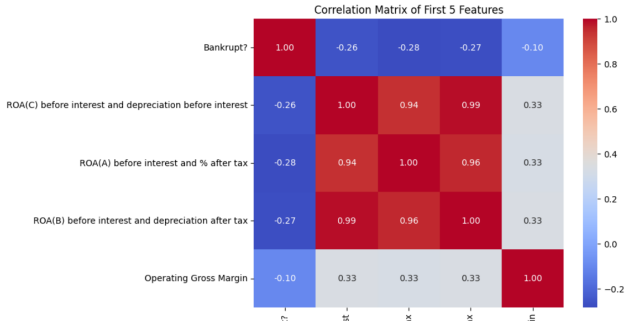

Fig. 3. A section of the Correlation Matrix

SMOTE creates synthetic data points by interpolating between existing minority class samples within their feature space. It selects a random sample from the minority class and computes new instances along the line connecting it to its nearest neighbors. This process enhances model generalizability by ensuring that the classifier learns meaningful patterns rather than memorizing a few examples.

By integrating SMOTE into our preprocessing pipeline, we balanced the dataset before training our **Random Forest classifier**, reducing bias toward non-bankrupt companies. The effectiveness of this approach was validated through improved classification performance across multiple evaluation metrics.

C. Feature Selection

Feature selection plays a crucial role in improving the performance of bankruptcy prediction models by identifying the most relevant financial and non-financial indicators that contribute to corporate distress. Effective feature selection helps reduce **overfitting, computational complexity, and noise**, ensuring that the model generalizes well to unseen data.

*a) **Key Financial Indicators for Bankruptcy Prediction**:* Several financial ratios and parameters have been widely used in bankruptcy prediction models:

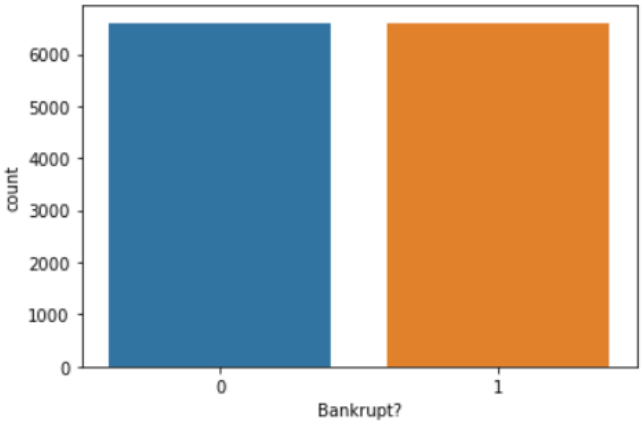

Fig. 4. Count Plot After applying SMOTE

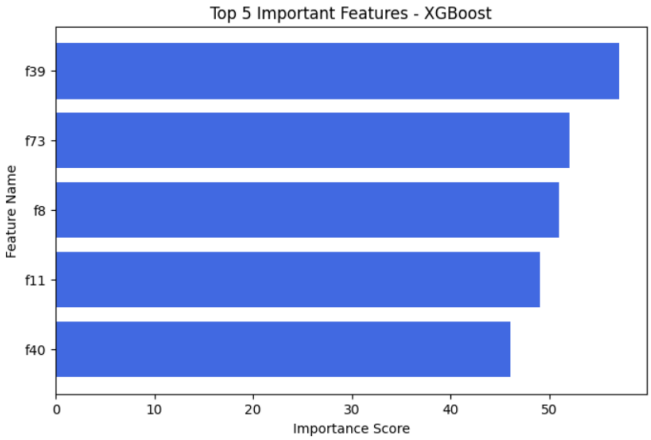

Fig. 5. Top 5 most important features in XGBoost

1) **Liquidity Ratios** – Indicate a firm's ability to meet short-term obligations:
   - *Current Ratio (Current Assets / Current Liabilities)*
   - *Quick Ratio (Liquid Assets / Current Liabilities)*
2) **Profitability Ratios** – Measure a company's ability to generate profit:
   - *Return on Assets (Net Income / Total Assets)*
   - *Return on Equity (Net Income / Shareholder's Equity)*
   - *Net Profit Margin (Net Profit / Revenue)*
3) **Leverage Ratios** – Assess the extent of financial risk and indebtedness:
   - *Debt-to-Equity Ratio (Total Debt / Shareholder's Equity)*
   - *Interest Coverage Ratio (EBIT / Interest Expense)*
4) **Efficiency Ratios** – Indicate operational effectiveness and asset utilization:
   - *Asset Turnover Ratio (Revenue / Total Assets)*

- *Receivables Turnover Ratio (Net Credit Sales / Average Accounts Receivable)*

5) **Market-Based Indicators** – Reflect investor confidence and market perception:
   - *Market Capitalization*
   - *Earnings Per Share (EPS)*

D. Model Development

Based on our literature review, prior research has consistently demonstrated that **XGBoost** outperforms traditionally used models such as **Logistic Regression** and **Altman Z-Score** in accurately predicting corporate bankruptcy. XGBoost's ability to handle complex, non-linear relationships and its robustness against overfitting make it particularly well-suited for financial distress prediction. Additionally, decision trees and logistic regression remain valuable benchmark models, offering interpretability and insights into key financial indicators that contribute to bankruptcy prediction.

In our study, we have employed three distinct classification techniques:

1) **XGBoost** – a gradient boosting algorithm known for its high predictive accuracy, feature importance evaluation, and ability to handle imbalanced datasets effectively.
2) **Decision Tree Classifier** – a rule-based algorithm that provides clear decision boundaries, enabling interpretability in understanding the factors leading to bankruptcy.
3) **Logistic Regression** – a widely used statistical method that serves as a baseline model for binary classification problems, offering explainability through coefficient analysis.

Given the inherent class imbalance in bankruptcy datasets—where bankrupt companies are significantly fewer than non-bankrupt ones—we have applied the **Synthetic Minority Over-sampling Technique (SMOTE)** to balance the dataset. As detailed in the previous section, **SMOTE generates synthetic samples** for the minority class, ensuring that our models are trained on a more balanced distribution. This technique helps mitigate bias toward the majority class, allowing for improved generalization and more reliable bankruptcy predictions.

By employing these three models and leveraging SMOTE for data preprocessing, our research aims to evaluate and compare their effectiveness in predicting corporate bankruptcy while identifying the most influential financial indicators driving bankruptcy risk.

IV. Results of Model Training

The training results clearly indicate that **XGBoost outperformed Decision Tree Classifier and Logistic Regression** across all performance metrics. As highlighted in the literature, XGBoost's ensemble-based boosting approach helps reduce overfitting, improves generalization, and enhances predictive power by sequentially refining weak learners.

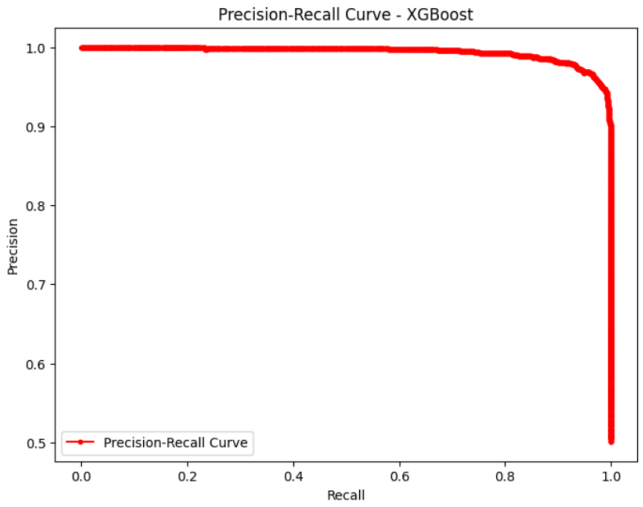

Fig. 6. Enter Caption

*a) **Performance Analysis of the Models**:*

- **XGBoost**: This model delivered the **best results**, achieving an **F1-score of 0.97, a Precision of 0.97, and an Accuracy of 0.97**. Its superior performance can be attributed to its ability to capture complex patterns in financial data, handle missing values efficiently, and optimize performance using regularization techniques. The gradient boosting mechanism allows XGBoost to learn from previous misclassifications, leading to a more refined decision boundary. Additionally, feature importance analysis in XGBoost provided insights into which financial variables contributed the most to bankruptcy prediction.
- **Decision Tree Classifier**: While Decision Trees are known for their interpretability, they **tend to overfit**, especially when working with imbalanced datasets. Even after applying **SMOTE for data balancing**, the Decision Tree model achieved an **F1-score of 0.91, Precision of 0.91, and Accuracy of 0.91**. The model struggled to generalize well, likely due to its tendency to create deep trees that memorize the training data rather than identifying underlying patterns.
- **Logistic Regression**: As a **linear model**, Logistic Regression has inherent limitations when applied to complex financial datasets where relationships between variables may be **non-linear**. While it remains a widely used baseline model in financial risk prediction, its inability to capture intricate dependencies led to **slightly lower predictive performance** (**F1-score: 0.91, Precision: 0.90, Accuracy: 0.91**). The slightly lower precision score indicates that it generated more **false positives**, potentially misclassifying financially stable firms as bankrupt.

XGBoost achieved an **F1-score of 0.97, a Precision Score of 0.97, and an Accuracy of 0.97**, significantly outperforming the other two models. The Decision Tree Classifier and

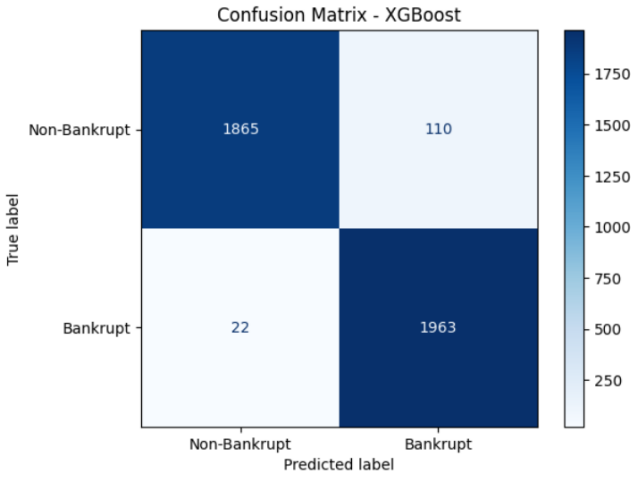

Fig. 7. Enter Caption

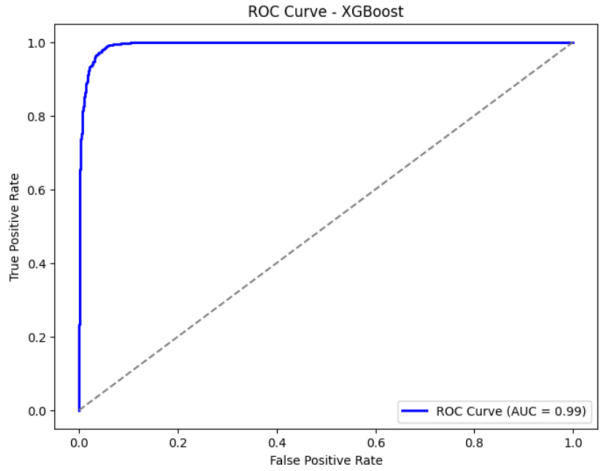

Fig. 8. Enter Caption

Logistic Regression models, while still effective, exhibited comparatively lower performance due to their inherent limitations. Decision Trees, though interpretable, tend to overfit on training data, reducing generalization. Logistic Regression, being a linear model, struggled to capture the complex decision boundaries present in bankruptcy data. To ensure a fair evaluation, all models were trained on the **balanced dataset** obtained using **SMOTE**, as described in the previous section. The performance metrics for each model are summarized below:

*b) **Impact of Class Imbalance and SMOTE**:* Class imbalance is a major challenge in bankruptcy prediction since the number of bankrupt firms is significantly lower than non-bankrupt ones. Without proper handling, models tend to be biased towards the majority class, failing to correctly predict financially distressed firms. To counter this, we applied **Synthetic Minority Over-sampling Technique (SMOTE)**, which artificially generates new instances of the minority class. This balancing process **improved model generalization and prevented bias** toward non-bankrupt firms.

- Without SMOTE, models exhibited **lower recall scores**, meaning that many actual bankrupt firms were misclassified as non-bankrupt.
- With SMOTE, models were better equipped to **identify financially distressed companies**, improving their **ability to assist investors, policymakers, and regulators in risk assessment**.

Model	F-1 Score	Precision	Accuracy
XGBoost	0.97	0.97	0.97
Decision Tree Classifier	0.91	0.91	0.91
Logistic Regression	0.91	0.90	0.91

TABLE I
PERFORMANCE METRICS OF THE THREE MODELS

V. AREAS OF FUTURE RESEARCH

As financial environments become increasingly complex, the next generation of **machine learning (ML) techniques** aims to improve bankruptcy prediction through enhanced **accuracy, interpretability, and adaptability**. Below are some promising ML advancements that could shape the future of bankruptcy prediction models.

*a) 1. **Graph Neural Networks (GNNs) for Bankruptcy Prediction**:*

- Traditional ML models treat companies as isolated entities, but in reality, firms operate within interconnected financial ecosystems.
- **GNNs** can model **supply chain relationships, interbank transactions, and investor networks**, capturing the ripple effects of financial distress.
- This approach is particularly useful for predicting systemic risks in financial markets.

*b) 2. **Self-Supervised and Semi-Supervised Learning**:*

- Bankruptcy data is often **imbalanced**—firms that survive far outnumber those that fail.
- **Self-supervised learning** leverages **unlabeled data** to learn feature representations before fine-tuning on smaller labeled datasets.
- **Semi-supervised techniques** use both labeled and unlabeled data to improve model generalization in low-data environments.

*c) 3. **Transformer-Based Models for Financial Time Series**:*

- Deep learning models like **LSTMs and GRUs** have been widely used for time-series forecasting but often struggle with long-term dependencies.
- Transformer-based architectures, such as **Temporal Fusion Transformers (TFTs)** and **FinanceBERT**, can improve predictive performance by capturing sequential trends in financial data more effectively.

VI. CONCLUSION

This study identifies **XGBoost** as the most effective model for corporate bankruptcy prediction, significantly outperforming traditional models such as **Decision Tree Classifier and**

Logistic Regression. The findings reinforce the growing role of **machine learning** in financial risk assessment, particularly in the context of **Indian companies**, where early detection of financial distress is crucial for investors, creditors, and policymakers.

Our research highlights the importance of **data-driven feature selection** and the ability of advanced machine learning models to capture complex financial patterns. By addressing class imbalance through **SMOTE**, we ensure that the model generalizes well to real-world bankruptcy scenarios, improving its reliability in predicting distressed firms. These insights can help financial institutions, regulatory bodies, and businesses make informed decisions to mitigate risks and enhance financial stability.

While our study demonstrates the superiority of **XGBoost**, several areas remain open for further exploration. **Future research directions include:**

1) **Integrating macroeconomic indicators** such as interest rates, inflation, GDP growth, and exchange rate fluctuations to examine their influence on bankruptcy prediction.
2) **Exploring deep learning architectures**, including recurrent and convolutional neural networks, to model sequential and non-linear financial behaviors.
3) **Assessing the role of corporate governance factors**, such as board structure, ownership patterns, and managerial decisions, in influencing financial distress.
4) **Developing real-time bankruptcy monitoring systems** that utilize streaming financial data to provide dynamic risk assessments.
5) **Investigating sector-specific bankruptcy trends**, tailoring predictive models to industries with unique financial characteristics (e.g., manufacturing, technology, or banking).

By incorporating these enhancements, future research can further refine bankruptcy prediction models, enabling **proactive financial risk management** and contributing to a more **resilient corporate landscape**. A robust bankruptcy prediction framework has the potential to support not only firms and investors but also regulatory bodies in fostering a more stable economic environment.

*a) 6. **Hybrid Models Combining ML and Economic Theories**:*

- Combining **economic theories** (e.g., Merton's Structural Model) with **deep learning models** could improve predictive insights.
- **Hybrid approaches** integrate ML's pattern recognition with economic intuition, making models more interpretable for financial analysts.

REFERENCES

[1] Altman, E.I. , "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy.", The Journal of Finance, 23(4), 589–609.

[2] Ohlson, J.A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research, 18(1), 109–131.

[3] Apostolos Dasilas, Anna Rigani, Machine learning techniques in bankruptcy prediction: A systematic literature review, Expert Systems with Applications, Volume 255, Part C, 2024, 124761, ISSN 0957-4174

[4] Shetty, Shekar & Musa, Mohamed & Brédart, Xavier. (2022). Bankruptcy Prediction Using Machine Learning Techniques. Journal of Risk and Financial Management. 15. 35. 10.3390/jrfm15010035.

[5] Meena, Manish and Pandey, Ashish and Garg, Dr. Ajay Kumar, Machine Learning Models Comparison for Bankruptcy Predication for Indian Companies a Study Based on India's Insolvency and Bankruptcy Code (Ibc '2016). Available at SSRN: https://ssrn.com/abstract=4690797https://ssrn.com/abstract=4690797 , http://dx.doi.org/10.2139/ssrn.4690797http://dx.doi.org/10.2139/ssrn.4690797

[6] CMIE Prowess Database. (2024). Financial Statement Data of Indian Companies.