**Request Proposal**

**Title**

Foreseeing Bankruptcy: Sensitivity-Weighted Interdependent Feature Translation for Early Warning of Financial Distress

**Course Name**

Masters In Machine Learning & Artificial Intelligence

**Presented by**

Asheesh Jakhmola

**Date**

September 2025

**Abstract**

Bankruptcy prediction is vital for ensuring business continuity and maintaining financial stability across enterprises. In recent years, machine learning-based bankruptcy prediction models combined with explainable AI techniques such as SHAP have demonstrated strong performance, providing both predictive accuracy and interpretability. However, these approaches also inherit limitations. SHAP assumes feature independence, which rarely holds true in real-world financial data where features are highly correlated. Moreover, SHAP can be computationally expensive and do not talk about how much change in input can change the output. These limitations restrict its practical utility. These shortcomings reduce accuracy and hinder the ability to generate reliable explanations, particularly in scenarios that demand transparency for critical financial decision-making. To address these gaps, this research proposes an enhanced interpretability framework that explicitly incorporates feature correlations into the explanation process, aiming to generate more realistic and trustworthy outcomes. The study further investigates techniques to model feature dependencies within bankruptcy prediction tasks and introduces methods to capture the sensitivity of model outputs to input changes. In addition, the research emphasizes improving computational efficiency without compromising explanation fidelity. By systematically analyzing the trade-offs between accuracy, speed, and interpretability, this work seeks to deliver a more robust and transparent framework for enterprise bankruptcy prediction

**Contents**

[**LIST OF FIGURES** 4](#_Toc207649327)

[**LIST OF ABBREVIATIONS** 4](#_Toc207649328)

[**1.** **Background** 5](#_Toc207649329)

[**2. Literature Review** 6](#_Toc207649330)

[**3.** **Research Questions** 15](#_Toc207649331)

[**4.** **Aim & Objective** 15](#_Toc207649332)

[**4.1 Aim** 15](#_Toc207649333)

[**4.2 Objectives:** 15](#_Toc207649334)

[**5.** **Significance of the Study** 16](#_Toc207649335)

[**6.** **Scope of Study** 16](#_Toc207649336)

[**7.** **Research Methodology** 16](#_Toc207649337)

[**7.1 Data Collection and Preparation** 17](#_Toc207649338)

[**7.2 Exploratory Data Analysis (EDA)** 18](#_Toc207649339)

[**7.3 Feature Engineering** 20](#_Toc207649340)

[**7.4 Handling Class Imbalance** 20](#_Toc207649341)

[**7.5 Model Development** 20](#_Toc207649342)

[**7.6 Model Evaluation** 21](#_Toc207649343)

[**7.7 Explainability Assessment** 22](#_Toc207649344)

[**8.** **Requirement Resources** 24](#_Toc207649345)

[**9.** **Research Plan** 25](#_Toc207649346)

[**References** 27](#_Toc207649347)

# **LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Propose model chart for FWCE | 08 |
| Figure 2. Framework of LIME | 11 |
| Figure 3. Different XAI Categorization | 13 |
| Figure 4. Research Methodology Flow | 16 |
| Figure 5. Research Plan | 22 |

# **LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| AI | Artificial Intelligence |
| ML | Machine Learning |
| XAI | Explainable Artificial Intelligence |
| SHAP | SHapley Additive exPlanations |
| LIME | Local Interpretable Model-agnostic Explanations |
| SWIFT | Sensitivity-Weighted Interdependent Feature Translation |
| SMOTE | Synthetic Minority Oversampling Technique |
| MIP | Modified Informative Position |
| FWCE | Feature-Weighted Counterfactual Explanation |
| EU | European Union |
| CNN | Convolutional Neural Network |
| LSTM | Long Short-Term Memory |
| GPU | Graphics Processing Unit |
| AI | Artificial Intelligence |
| ML | Machine Learning |
| XAI | Explainable Artificial Intelligence |
| SHAP | SHapley Additive exPlanations |
| LIME | Local Interpretable Model-agnostic Explanations |
| AUC | Area Under the Curve |

# **Background**

Bankruptcy forecasting is an essential for ensuring business continuity and financial stability, especially in high-stakes domains like corporate finance. With the rapid growth of data, machine learning has become central to prediction tasks, offering higher accuracy than traditional statistical models. However, as models become more complex, interpretability has emerged as a critical challenge, particularly in regulated financial sectors where transparency and trust are vital.

SHAP (SHapley Additive Explanations), Explainable Artificial Intelligence (XAI) method is widely used to interpret model outputs and support decision-making. While SHAP is effective in providing feature-level explanations, it faces limitations, including the assumption of feature independence, high computational cost and inability to capture how input changes affect predictions These shortcomings restrict its utility in real-world scenarios where features are often correlated and dataset has good amount of data.

To address these gaps, this study proposes an improved interpretability framework that incorporates feature dependencies, accounts for sensitivity of model outputs, and enhances computational efficiency. The goal is to provide more accurate, transparent, and reliable bankruptcy predictions while balancing the trade-offs between accuracy, speed, and interpretability.

# **2. Literature Review**

[12] Machine learning models outperform traditional statistical methods such as logistic regression in predicting corporate bankruptcy particularly on imbalanced datasets. Random Forest achieved the strongest performance in both accuracy and F1-score while XGBoost and LightGBM also performed well after tuning and resampling. Since bankruptcy data are heavily imbalanced. Methods like SMOTE improve prediction reliability.

SHAP analysis showed that financial indicators such as working capital, retained earnings, total assets and EBIT are key drivers. By combining models with SHAP complex algorithms become interpretable and meets regulatory demands. These models can predict bankruptcy a year in advance, supporting timely intervention and financial stability.

[4] Black-box models such as deep learning perform well in predicting financial risk but their opacity limits trust and acceptance by regulators and institutions. Explainable AI (XAI) is essential for compliance. XAI has methods like LIME and SHAP that generate post-hoc explanations for any model and these methods are broadly applicable to risk-related tasks.

Risk management benefits from both global and local explanations. Global insights reveal which features drive predictions overall while local explanations clarify individual outcomes such as loan denials. Among available methods SHAP stands out for its stability and consistency. That makes SHAP most robust approach in finance. AI also supports feature selection by identifying key drivers which helps reducing dimensionality while maintaining accuracy.

[3] Deep learning models often surpass human experts in accuracy yet their decision-making process remains opaque. The paper emphasizes the accuracy-opacity trade-off and identifies three forms of opacity.

* Internal opacity: Internal opacity is due to the complexity of neural networks.
* Link opacity: link opacity reflects the weak connection between model features and real-world meaning.
* Structure opacity: Structure opacity comes from the incomprehensibility of the underlying phenomena itself

Methods such as SHAP and counterfactual explanations increase transparency by identifying key features or showing minimal changes needed for different outcomes. XAI can reduce link opacity when strong empirical knowledge exists but it cannot overcome structure opacity. Opacity also varies with the user’s expertise and may diminish over time as science advances.

Even when models reveal correlations these do not always constitute explanations. In high-stakes domains like healthcare, opacity challenges trust, responsibility and liability, requiring caution, empirical validation and ethical oversight.

[17] Common explainability tools like SHAP and LIME assume features are independent which rarely holds in real datasets. When correlations exist these methods misestimate importance and producing mathematically unfaithful attributions. Author proposes a new proxy method called Modified Informative Position (MIP) that modifies existing XAI ranking methods by considering feature dependency. Experiments show MIP outperforms SHAP, particularly in highly correlated settings. Kernel SHAP’s reliance on marginal perturbations and LIME’s linear surrogates make them unreliable with complex feature interactions. MIP uses generative models or empirical distributions to give more accurate explanations. MIP requires more computation for this. This extra accuracy is much more important and helpful in areas such as healthcare and finance

[5] Most counterfactual explanation methods treat all features equally which often produces unrealistic or less interpretable results. This paper introduces Feature-Weighted Counterfactual Explanation (FWCE) which is a method that incorporates feature importance weights into counterfactual generation.

By prioritizing influential features FWCE creates counterfactuals that are both realistic and feasible. It draws on locally extracted weights from techniques such as SHAP to guide which features are perturbed and it is also ensuring changes align with human intuition. This approach minimizes both the distance between original and counterfactual inputs and the weighted impact of feature changes, resulting in plausible and minimally distorted outcomes.

Propose model chart for FWCE:

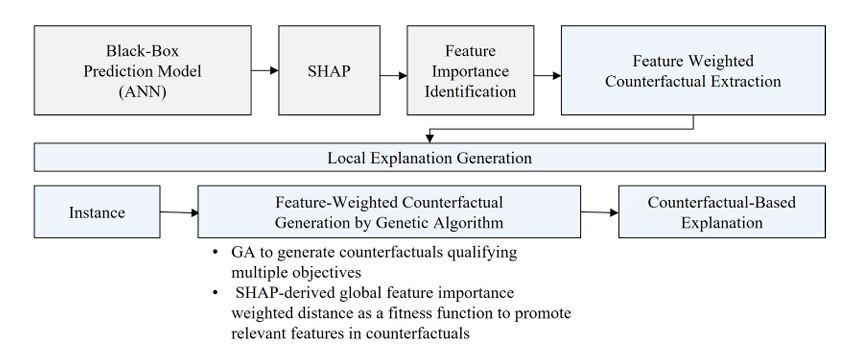


Figure: 1

FWCE is model-agnostic. It is effective across neural networks, decision trees and ensembles. It produced sparser and interpretable counterfactuals.

[2] SHAP provides accurate feature attributions but iscomputationally expensive particularly on large datasets and complex models. To address this author, propose ReSHAP, a framework that approximates SHAP values using a small but representative subset of the training data.

The formula to find the size of the sample population is

n = N / (1 + Ne^2) Where:

n is the required sample size

N is the population size.

e is the margin of error (expressed as a decimal).

ReSHAP employs K-means clustering to select diverse and dense instances. This ensures interpretability is preserved while reducing computational cost. Empirical results show up to 90% reduction in runtime with minimal accuracy loss in explanations. The method is model-agnostic. ReSHAP is working across linear, tree-based and neural network models.

Importantly ReSHAP only affects the explanation layer and do not touch predictive performance. Its adjustable clustering parameters allow users to balance speed and fidelity. This makes SHAP more practical in resource-constrained environments.

[13] The growing use of complex AI deep neural network models in sensitive areas like healthcare and finance has created strong demand for explanation methods that build trust and transparency. The paper distinguishes between ante-hoc models which are inherently interpretable and post-hoc techniques which explain black-box models. Both play complementary roles.

XAI methods are classified as

* Model-specific (e.g. Grad-CAM for CNNs)
* Model-agnostic (e.g. SHAP, LIME).
* SHAP is for global explanation and valued for theoretical consistency.
* LIME is for local explanation, explaining any specific instance.
* Counterfactual approach shows minimal changes needed to flip decisions.

Tools such as saliency maps, Grad-CAM and LRP are effective for CNNs but vulnerable to input perturbations. Visual aids like heatmaps, dependency plots and trees improve accessibility for non-experts.

A key challenge remains balancing fidelity with interpretability. While simpler models are easier to explain, they sacrifice accuracy. The paper highlights risk of manipulation, the absence of universal evaluation metrics and calls for XAI standards, benchmarks and legal frameworks.

[7] The XAI program emerged from concerns that advances in AI particularly deep learning which produces highly accurate but black box systems. This reduces trust and understanding. Its goal was to design models that not only predict accurately but also provide human understandable explanations.

The program focused on three areas:

* Building explainable models
* Studying psychological aspects of explanations
* Testing methods through real-world problems in autonomy and analytics

Research teams explored interpretable models, deep explanations and proxy-based model induction.

Over 12,700 users participated in experiments showing that explanations improved trust and decision-making, especially for complex or error-prone tasks. A scoring system (ESS) was introduced to evaluate consistency and utility. The work emphasized interdisciplinary collaboration, user-centred evaluation and the value of explanations in supporting human-AI alignment.

[6] Structured documentation is essential for accountability, interpretability and oversight in AI. The CLeAR framework provides such structure through four components:

* Contextualization
* Legal and ethical information
* Assessments of performance
* Recommendations for use

CLeAR is designed for both technical and non-technical audiences. CLeAR is fostering shared understanding among developers, regulators and the public. Developers are encouraged to document laws, norms and ethical standards, alongside fairness, bias, safety and performance evaluations.

The framework is proposed to use across the AI lifecycle, from design to post-deployment monitoring and across domains such as healthcare, education and criminal justice. By clarifying uses, limits and risks CLeAR improves trust, enables oversight and prevents inappropriate repurposing of AI systems.

[9] The paper points out that in ML and DL there is a gap in making models both responsible and easy to understand. Since the measures of explainability are still unclear. A systematic review using PRISMA guidelines and sources like Scopus, Google Scholar, and Arxiv found 6122 XAI studies, but only 884 used evaluation metrics. Even though XAI is widely used, there is still no clear agreement on what counts as an “explanation.”

Two taxonomies are proposed:

* Methodology-Based
* Application-Based

The review shows that many explanations are judged subjectively and that can cause bias. The authors ask for clear, validated metrics to build trust, allow fair comparison and support safe use of XAI in areas like healthcare and security.

[15] AI has grown quickly but the complexity of neural networks makes it hard to understand how models make decisions. This creates challenges for trust and accountability. The idea of trustworthy AI has therefore gained importance. It is requiring techniques to meet societal and ethical standards.

This paper surveys technologies that improve explainability, tracing emerging trends and methods. It highlights the link between explainability and meta-reasoning which aligns with the goals of XAI. Transparency is a key ethical requirement of XAI. That is also aligned to EU AI high-level expert group’s principles of autonomy, harm prevention, fairness and explicability.

The study reviews approach such as post-modelling analysis, reward driven explainability and integration of methods (like Chain-of-Thoughts and Tree-of-Thoughts in large language models). Rule-based models including decision trees are noted as more interpretable than complex black-box models like neural networks or SVMs.

The authors conclude that explainability is essential to improve transparency, foster trust and enable responsible decision-making in AI applications.

[10] Local Interpretable Model-agnostic Explanations (LIME) is among the most widely used methods in explainable AI. This gives local insights into how complex models make individual decisions. Its flexibility allows it to work across diverse models including neural networks, CNNs, LSTMs, transformers, decision trees and random forests.

Despite its popularity LIME faces key challenges such as instability, inefficiency and difficulty in handling certain data types. To address this, many studies have proposed modifications for a systematic review of LIME and its variants. This paper presents a taxonomy that categorises extensions along two dimensions: technical modifications within LIME and the specific problems they address.

LIME’s four-step process

* Feature generation
* Sample generation
* Feature attribution
* Explanation representation

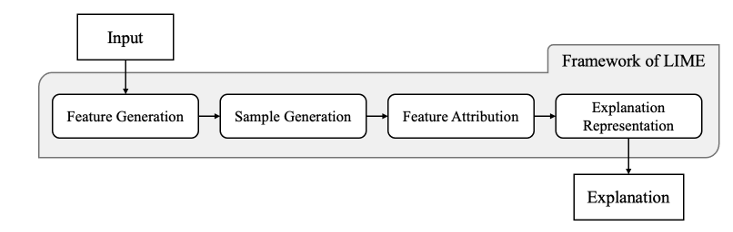


Figure:2

Evaluation of LIME is inconsistent as many studies only compare it with the standard version and ignore wider benchmarks.

Trustworthiness in AI underscores the importance of improving evaluation.

Future work should focus more on the user (explainee), incorporating contextual knowledge and aligning explanations with user expectations to strengthen both utility and trust.

[14] The paper asserts that interpretability and explainability should be understood as a multi-faceted concept. It introduces 12 conceptual properties (Co-12) which should be evaluated to comprehensively assess explanation quality. These properties divided into 3 categories.

Content:

* Correctness: Describe how faithful the explanation is with respect to the black box.
* Completeness: Describe how black box behaviour is described in the explanation.
* Consistency: Describe how deterministic and implementation-invariant the explanation method is.
* Continuity: Describe how stable the explanation is for slight variations in the input.
* Contrastivity: Describe how explanation differentiates the explained instance from others.
* Covariate complexity: Describe how complexity of the features and their interactions used in the explanation.

Presentation:

* Compactness: Describe size of the explanation.
* Composition: Describe presentation format and organization of the explanation.
* Confidence: Describe presence and accuracy of probability information in the explanation.

User:

* Context: Describe how relevant the explanation is to the user and their needs.
* Coherence: Describe how plausible and understandable the explanation is to a human.
* Controllability: Describe how extent to which a user can control, correct, or interact with an explanation.

The study provides a practical and publicly available interactive website listing 312 papers that introduce explainable AI methods.  
The authors call that relying solely on anecdotal evidence for evaluation is insufficient and can be misleading. Authors are emphasising the need for more robust and objective validation methods.  
The paper highlights the significance of objective evaluation methods that do not require human studies particularly because the sheer volume of different explanations that can be generated far exceeds the capacity for human assessment.

[1] The study highlights the need for advanced Default Prediction Models (DPMs), especially given the risks posed by large-scale financial defaults such as the 2008 crisis. Traditional models do not have such capability. Traditional model lack transparency, rely on static assumptions and unable to capture non-linear financial patterns.

Recent research points to a shift toward hybrid and intelligent models that combine statistical and AI methods. Deep learning approaches, including LSTM, CNN, LightGBM and XGBoost, show superior performance, often achieving 80 to 95% accuracy in predicting financial distress. These models also benefit from diverse data inputs, such as macroeconomic indicators, credit bureau records and textual information, with careful feature selection proving critical.

Transparency remains central. Explainable AI (XAI) techniques like SHAP and LIME are increasingly adopted to make predictions interpretable, aligning with regulatory frameworks (like EU AI Act). Even though challenges remain same, including inconsistent evaluation practices, difficulties in assessing textual models and reliance on historical data.

The paper empathises the need for domain-specific benchmarks, cross-sector collaboration and future research into underexplored XAI goals for ensuring financial prediction models remain both accurate and accountable.

[16] The review highlights a significant gap in evaluating Explainable AI (XAI) results. Most studies rely on expert judgment or anecdotal evidence rather than robust quantitative measures. This pointing to need for standardised evaluation frameworks that improve reliability, interpretability and stability.

Healthcare dominates XAI research like applications in cancer diagnosis, COVID-19 and medical imaging. XAI is also used in finance, law, cybersecurity, agriculture, transportation and education. It shows its growing role in sensitive and regulated areas where clarity and compliance are important.

Following is the overview of different XAI approaches and evaluation methods. These categories were used to classify the XAI application papers reviewed in this study

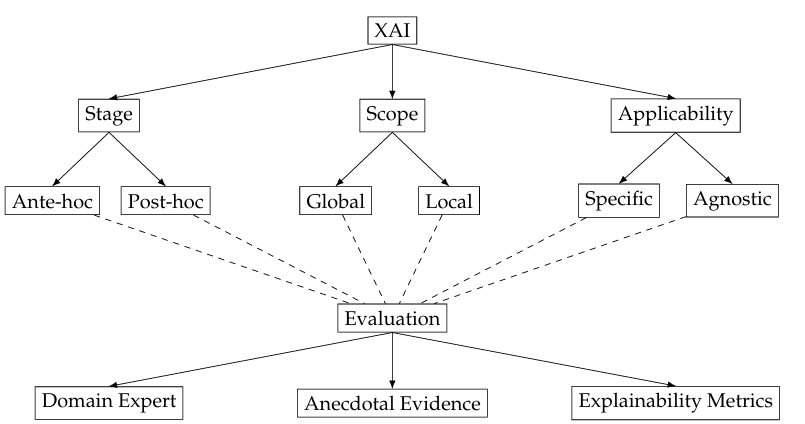


Figure:3

Methodologically, most studies apply local explanations (53%), while 29% combine local and global approaches. SHAP and LIME are the leading techniques, with SHAP preferred for its mathematical guarantees and stability. Gradient-based method, such as Grad-CAM and Integrated Gradients are commonly used for complex data like images.

Neural networks remain the most frequent model type (59%), followed by tree-based models (37%). Feature importance is the dominant form of explanation, often paired with visualisations, saliency maps or counterfactuals. Despite this 58% of studies lacked quantitative evaluation.

# **Research Questions**

* 1. In what ways can feature correlation be included into interpretability methods to generate more realistic explanations for enterprise bankruptcy prediction?

Real-world datasets frequently contain correlated features, but SHAP assumes independence. This assumption often distorts the attribution of importance.

* 1. How can interpretability approaches capture the sensitivity of model outputs to changes in feature values for enterprise bankruptcy prediction?

Beyond ranking feature importance, decision-makers need to understand how variations in input values influence predictions. This raises the question of how to quantify the directional impact and magnitude of input changes.

* 1. What kind of trade-offs appears among accuracy, computational efficiency and interpretability when designing an improved algorithm?
  2. How can computational efficiency in feature attribution methods be improved without sacrificing accuracy?

# **Aim & Objective**

**4.1 Aim:** To improve machine learning interpretability by including feature correlations and generating more realistic and trustworthy explanations for bankruptcy model predictions.

## **4.2 Objectives:**

1. To analyze the limitations of current methods that ignore feature dependencies.
2. To find technique that integrates these correlations into the explanation generation process.
3. To validate the new method by demonstrating that its explanations are more plausible and faithful to the model's logic compared to existing approaches.

# **Significance of the Study**

1. **Addressing Limitations**: The study examines the weaknesses of current explainability methods that ignore feature dependencies, which often lead to incomplete or biased interpretations.
2. I**ntroducing Correlation-Aware Explanations:** It proposes a new technique that incorporates feature correlations into the explanation generation process, thereby improving interpretability.
3. **Enhancing Plausibility and Faithfulness**: The method ensures that explanations align more closely with the model’s internal logic, making them more realistic and trustworthy.
4. **Validation and Practical Utility**: By comparing with existing approaches, the study highlights its improved effectiveness, particularly for regulated financial decision-making.

# **Scope of Study**

In this working on a novel Explainable AI (XAI) framework integrating three recent advances: efficient data reduction, correlation-aware feature ranking and feature-weighted counterfactual generation. The scope includes selecting reduced representative datasets using Slovin’s formula, identifying feature dependencies through correlation-sensitive ranking and generating minimal yet realistic counterfactuals via a multi-objective optimization. The project focuses on financial risk prediction, specifically bankruptcy classification, but its methodology is model-agnostic and applicable to broader domains requiring trustworthy, interpretable machine learning.

# **Research Methodology**

This research methodology follows the systematic steps to study, develop, implement and evaluate an improved explainable artificial intelligence (XAI) algorithm for enterprise bankruptcy prediction. This methodology integrates principles from machine learning, financial risk modelling and interpretability research.

The process is organized into Seven phases:

1. Data collection and preparation
2. Exploratory analysis
3. Feature Engineering
4. Handling class imbalance
5. Model development
6. Model evaluation
7. Explainability assessment

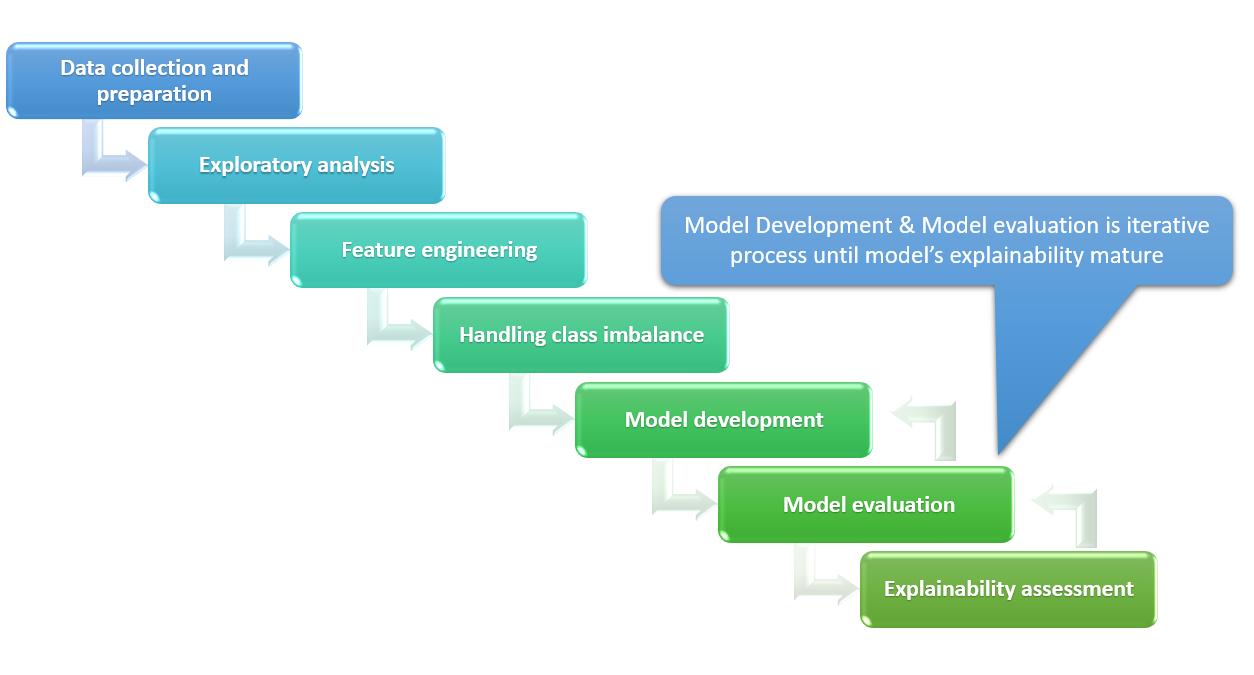


Figure: 4

## **7.1 Data Collection and Preparation**

* + **Data Source**: This study uses the publicly available Company Bankruptcy Prediction dataset hosted on Kaggle. This is originally compiled from the Taiwan Economic Journal. The dataset contains financial information on 6,819 companies. This data labeled as either bankrupt or non-bankrupt across 96 financial features. These features cover multiple dimensions of firm performance, including liquidity ratios, profitability indicators, leverage ratios, and efficiency measures.
  + **Data Loading**: The dataset will be loaded into Python using the Pandas library. Initial checks will be performed to inspect structure, feature types and missing values. This step ensures readiness for downstream preprocessing and modelling.
* **Data Cleaning**
* Missing values: find the missing values and fix the missing values using strategies such as mean, median, or k-nearest neighbour.
* Outliers: Outlier detection will rely on statistical thresholds (z-scores) and visualization (boxplots).
* Normalization and Scaling: Since features represent financial ratios with differing magnitudes, StandardScaler or MinMaxScaler will be used to normalize them.
* Encoding: If categorical attributes exist, they will be encoded using one-hot or label encoding.

## **7.2 Exploratory Data Analysis (EDA)**

Exploratory Data Analysis will be carried out to understand data distribution, feature behavior, and interrelationships before model development.

* **Class Distribution**
  + Assess imbalance between bankrupt and non-bankrupt firms.
  + Apply visualization techniques (bar charts, pie charts) to highlight imbalance.
  + Quantify imbalance ratios to guide the choice of resampling methods (SMOTE, undersampling).
* **Statistical Summaries**
  + Compute mean, variance, skewness, and kurtosis for all financial ratios and indicators.
  + Detect outliers and anomalies (e.g., extreme debt-to-equity ratios).
  + Identify non-normal distributions requiring transformation (e.g., log-scaling).
* **Correlation Analysis**
  + Use Pearson correlation for linear dependencies and Spearman correlation for rank-based relationships.
  + Identify multicollinearity clusters of financial ratios.
  + Perform variance inflation factor (VIF) analysis to further confirm redundant features.
* **Visualization Techniques**
  + **Histograms & Boxplots:** To inspect skewness and detect outliers**.**
  + **Scatter Plots:** To visualize non-linear relationships between key ratios**.**
  + **Heatmaps:** To identify correlated groups of features**.**
  + **Pairplots:** For deeper insight into interactions among selected features**.**
* **Missing Value Analysis**
  + Detect missing values and patterns (random vs. systematic).
  + Compare imputation strategies (mean/median imputation, regression imputation).
* **Outlier Detection**
* Apply z-score and IQR methods to flag extreme values.
* Consider domain-driven thresholds (e.g., debt ratio > 100%).
* Decide whether to cap, transform, or retain them for model robustness.
* **Feature Distributions**
* Analyze distribution of key financial features across bankrupt vs. non-bankrupt groups.
* Check discriminatory power of features (e.g., liquidity ratios may show stark differences).
* **Data Quality Checks**
  + Identify duplicate records.
  + Ensure consistent units across financial ratios.
  + Standardize data formatting (e.g., percentage vs. ratio form).

## **7.3 Feature Engineering**

* **Feature Selection: Feature** redundancy will be addressed through correlation-based selection, variance thresholds, and recursive feature elimination to retain only the most informative predictors.
* **Correlation Grouping:** For the interpretability component, correlated features will be grouped to ensure that the algorithm considers their joint contribution, avoiding misleading attributions that can occur in SHAP.
* **Feature Transformation:** Non-linear relationships will be captured using transformations such as logarithmic, polynomial, or interaction terms. This step enhances the model’s ability to recognize complex financial patterns.
* **Dimensionality Reduction:** Techniques like Principal Component Analysis (PCA) and Autoencoders will be explored to reduce dimensionality while retaining maximum variance, thereby improving efficiency without losing interpretability.
* **Domain-Specific Feature Construction:** Financial ratios such as Altman Z-score, Debt-to-Equity ratio, Quick Ratio, and Liquidity indicators will be engineered to embed domain knowledge directly into the model.

## **7.4 Handling Class Imbalance**

The dataset is expected to show imbalance, with significantly fewer bankrupt firms compared to solvent firms. To address this:

* **Oversampling**: Synthetic Minority Oversampling Technique (SMOTE) will be applied to generate synthetic examples of the minority class.
* **Undersampling**: Majority class undersampling may be used for comparison to assess effects on performance.
* **Class Weights**: Machine learning models will be trained with adjusted class weights to penalize misclassification of the minority class more heavily.

## **7.5 Model Development**

* **Baseline Models**
* Build Decision Trees using current features for transparent feature importance.
* Build Random Forests for robustness and ensemble-based performance.
* Save baseline models for comparison.
* **Advanced Models**
  + Implement gradient boosting methods: XGBoost, LightGBM, and CatBoost.
  + Train with current features to create advanced benchmarks.
  + Save models for future evaluation.
* **Training and Hyperparameter Optimization**
  + Use 80/20 split of training and testing data with stratification to maintain class ratios.
  + Apply 5-fold cross-validation for robust performance evaluation.
  + Perform hyperparameter tuning using GridSearchCV and Bayesian optimization.
* **Feature Gap Analysis with SHAP**
  + Apply SHAP on baseline and advanced models to identify feature gaps.
  + Correct redundant or misleading features.
  + Retrain and save improved SHAP-driven models.
* **Feature Gap Analysis with SWIFT**
  + Apply SWIFT to capture feature interdependencies
  + Introduce sensitivity-weighted feature transformations.
  + Retrain and save improved SWIFT-driven models.
* **Final Output**
* Maintain all model versions (baseline, advanced, SHAP-based, SWIFT-based).
* Document improvements for transparency and reproducibility.

## **7.6 Model Evaluation**

* **Performance Metrics**
  + Evaluate models using accuracy, precision, recall, F1-score, and AUC-ROC to capture classification effectiveness.
  + For bankruptcy prediction, prioritize recall and F1-score to minimize false negatives (missing bankrupt firms).
* **Baseline vs. Advanced Models**
  + Compare Decision Trees and Random Forests with gradient boosting methods (XGBoost, LightGBM, CatBoost).
  + Establish whether advanced models significantly outperform baseline models.
* **SHAP-based Models**
  + Assess models improved through SHAP-driven feature selection.
  + Check whether removal of redundant features and better interpretability enhance predictive power.
  + Compare SHAP models against baseline/advanced in terms of both performance metrics and explainability scores.
* **SWIFT-based Models**
  + Evaluate SWIFT-enhanced models that consider feature interdependencies.
  + Validate improvements in accuracy, recall, and interpretability over SHAP-only models.
  + Highlight cases where sensitivity-weighted feature grouping provides clearer insights.
* **Cross-Validation and Stability**
  + Use 5-fold cross-validation results to ensure robustness across different data splits.
  + Track standard deviation of performance metrics to assess model stability.
* **Model Interpretability**
  + Compare explanation consistency across SHAP and SWIFT methods.
  + Ensure selected features align with domain knowledge in corporate bankruptcy prediction.
* **Final Model Selection**
  + Select the best-performing model balancing accuracy, recall, and interpretability.
  + Document the trade-offs between complexity, transparency, and predictive power.

## **7.7 Explainability Assessment**

* **Feature Correlation Integration**
  + Group highly correlated financial ratios to provide joint explanations instead of isolated contributions.
  + Overcomes SHAP’s independence assumption, producing more realistic and trustworthy interpretations.
  + Validates whether explanations match domain knowledge (e.g., debt-equity and leverage ratios working together).
* **Sensitivity Analysis**
  + Apply perturbation-based methods to check how small changes in input values affect predictions.
  + Identify high-sensitivity features where slight variations strongly influence bankruptcy predictions.
  + Compare sensitivity results with SHAP-based importance to ensure robustness and stability.
* **Computational Efficiency**
  + Measure runtime improvements of the proposed approach compared to SHAP.
  + Use approximation and optimization techniques (sampling, gradient-based) to reduce explanation cost.
  + Maintain a balance between speed and explanation fidelity.
* **Trade-off Analysis**
  + Assess trade-offs among correlation integration, sensitivity depth, and computational efficiency.
  + Document scenarios where speed optimization may slightly reduce granularity of explanations.
  + Provide guidelines for enterprises to select the right balance based on their needs (accuracy vs. speed).
* **Benchmarking Against Existing Methods**
  + SHAP: Benchmark improvements in interpretability accuracy, runtime, and stability.
  + Counterfactuals: Use as a complementary method to evaluate minimal input changes needed for different outcomes (e.g., reducing liabilities to avoid bankruptcy).
  + Ensure combined approaches (correlation-aware SHAP/SWIFT) provide richer and actionable insights.
* **Explainability Evaluation Metrics**
  + **Fidelity**: Do explanations accurately reflect model reasoning?
  + **Stability**: Are explanations consistent across data subsets and perturbations?
  + **Comprehensibility**: Are outputs clear and usable for financial decision-makers?
  + **Efficiency**: Is explanation generation feasible in enterprise-scale deployments?

# **Requirement Resources**

* **Data Resources**
* Bankruptcy Prediction Datasets: Kaggle “Company Bankruptcy Prediction”.
* Preprocessing libraries for handling imbalance (SMOTE, undersampling).
* **Software & Libraries**
* Programming Language: Python (≥ 3.10).
* ML Libraries: scikit-learn, XGBoost, TensorFlow/PyTorch (for ANN/SVM models).
* XAI Libraries: SHAP, LIME (for baseline), Alibi (for counterfactuals).
* Optimization: DEAP (Genetic Algorithm), Nevergrad, or PyGAD.
* Data Handling: pandas, numpy, scipy.
* Visualization: matplotlib, seaborn, plotly (for SHAP plots, counterfactual comparisons).
* **Hardware Requirements**
* Mid-to-high performance CPU (i.e. Intel i5/AMD Ryzen 5 or higher).
* At least 16 GB RAM for handling SHAP + GA iterations.
* Storage: 5-10 GB (to keep datasets, checkpoints and results).

# **Research Plan**

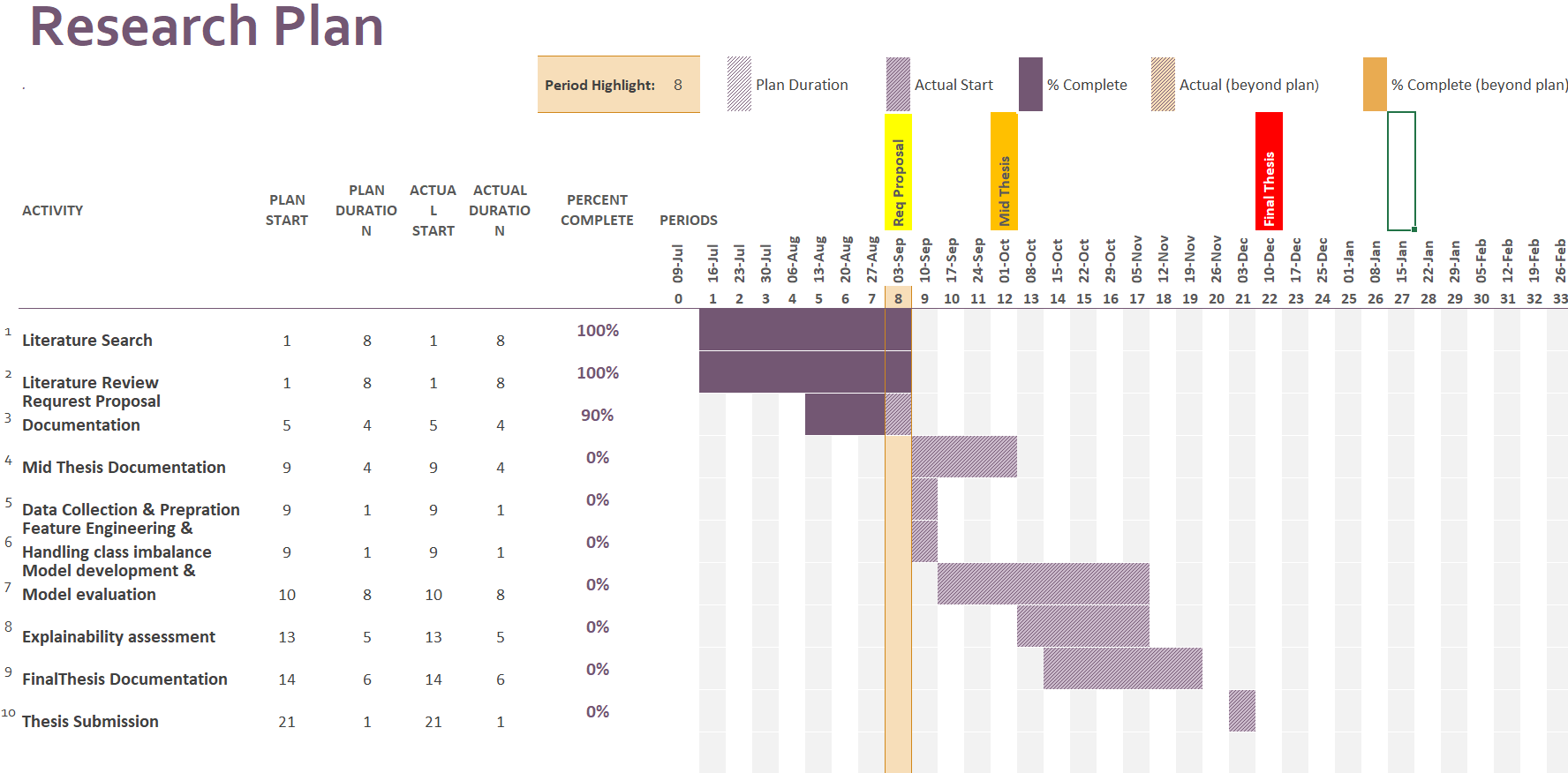


Figure:5

**Research Plan in detail**

* **Week 1-8: Literature Search**  
  Conduct a search of academic journals, books, and databases to identify existing work. Gather relevant studies that form the foundation of the research topic.
* **Week 1-8: Literature Review**  
  Evaluate collected literature to identify research gaps, trends, and challenges. Summarize insights that justify the need for the proposed study.
* **Week 5-8: Request Proposal Documentation**  
  Prepare and draft the research proposal, outlining objectives, methodology, and expected outcomes. Submit for review and approval from academic supervisors.
* **Week 9-12: Mid Thesis Documentation**  
  Develop mid-stage thesis documentation, detailing progress, methodology, and preliminary results. Submit interim findings for feedback and guidance.
* **Week 9: Data Collection & Preparation**  
  Gather relevant datasets and perform necessary preprocessing steps. Ensure data quality by handling missing values and inconsistencies.
* **Week 9: Feature Engineering & Handling Class Imbalance**  
  Transform raw data into meaningful features for modelling. Apply balancing techniques like SMOTE to address class imbalance.
* **Week 10-18: Model Development & Model Evaluation**  
  Build machine learning models tailored for bankruptcy prediction. Evaluate performance using appropriate metrics to validate accuracy and robustness.
* **Week 13-19: Explainability Assessment**  
  Incorporate XAI methods to interpret model predictions. Assess explanation quality to ensure transparency and trustworthiness.
* **Week 14-19: Final Thesis Documentation**  
  Integrate all sections into a complete thesis draft. Refine structure, formatting, and references for submission readiness.
* **Week 21: Thesis Submission**  
  Submit the finalized thesis to the academic committee. Ensure compliance with formatting and submission guidelines.

**Risk and Contingency Plan**

1. **Computational Power**: A major risk is that high computational demand may slow down processing or even halt large-scale model training. To handle this, cloud-based platforms and parallel computing techniques will be used. Additionally, lighter model versions will be explored to reduce resource usage.
2. **Data Challenge**: Data quality, missing values, or imbalance can limit the model’s performance. To address this, multiple reliable sources of data will be collected. Preprocessing methods such as cleaning, normalization, and data augmentation will be applied to ensure consistency. For imbalance, re-sampling and advanced balancing techniques will be used.
3. **Model Complexity**: Overly complex models may reduce interpretability and increase training time. The contingency plan is to use simpler baseline models for comparison, perform model pruning, and apply dimensionality reduction to control complexity while maintaining accuracy.
4. **Complex Algorithms**: Advanced algorithms may create challenges in implementation and explainability. To manage this, modular coding practices will be followed, multiple explanation methods (like visual graphs, statistic formula )will be tested, and documentation will be maintained to ensure clarity and reproducibility.

# **References**

1. Alvi, J., Arif, I. & Nizam, K., 2024. *Advancing financial resilience: A systematic review of default prediction models and future directions in credit risk management*.
2. Bachmann, S., 2023. *Efficient XAI: A Low-Cost Data Reduction Approach to SHAP Interpretability*.
3. Bjerring, J.C., Mainz, J. & Munch, L., 2025. *Deep learning models and the limits of explainable artificial intelligence*.
4. Bowden, J., Cummins, M., Dao, D. & Jain, K., 2024. *Explainable AI For Financial Risk Management*.
5. Cho, S.H. & Shin, K., 2023. *Feature-Weighted Counterfactual-Based Explanation for Bankruptcy Prediction*.
6. Chmielinski, K., Newman, S., Kranzinger, C.N., Hind, M., Wortman Vaughan, J., Mitchell, M., Stoyanovich, J., McMillan-Major, A., McReynolds, E., Esfahany, K., Gray, M.L., Chang, A. & Hudson, M., 2024. *The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners & Context for Policymakers*.
7. Gunning, D., Vorm, E., Wang, J.Y. & Turek, M., 2021. *DARPA’s explainable AI (XAI) program: A retrospective*.
8. Karunagaran, S., Lucic, A. & Custis, C., 2022. *XAI Toolsheet: Towards A Documentation Framework for XAI Tools*.
9. Kadir, M.A., Mosavi, A. & Sonntag, D., 2023. *Evaluation Metrics for XAI: A Review, Taxonomy and Practical Applications*.
10. Knab, P., Marton, S., Schlegel, U. & Bartelt, C., 2025. *Which LIME should I trust? Concepts, Challenges and Solutions*.
11. Long, B., Liu, E., Qiu, R. & Duan, Y., n.d. *Explainable AI - the Latest Advancements and New Trends*.
12. Maktum, T., Pulgam, N., Chandgadkar, V., Pathak, P. & Solanki, A., 2025. *A Machine Learning Based Framework For Bankruptcy Prediction In Corporate Finances Using Explainable AI Techniques*.
13. Mathew, D.E., Ebem, D.U., Ikegwu, A.C., Ukeoma, P.E. & Dibiaezue, N.F., 2025. *Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human*.
14. Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schloetterer, J., van Keulen, M. & Seifert, C., 2023. *From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI*.
15. Nguyen, H.H., Viviani, J-L. & Jabeur, S.B., 2023. *Bankruptcy prediction using machine learning and Shapley additive explanations*.
16. Saarela, M. and Podgorelec, V., 2024. Recent applications of explainable AI (XAI): A systematic literature review. *Journal/Publisher Name*
17. Salih, A.M., Galazzo, I.B., Raisi-Estabragh, Z., Petersen, S.E., Menegaz, G. & Radeva, P., 2024. *Characterizing the Contribution of Dependent Features in XAI Methods*.