analytics

MDPI

Article

Bankruptcy Prediction Using Machine Learning and Data Preprocessing Techniques

Kamil Samara [1,2,*] and Apurva Shinde [3]

[1] Computer Science Department, University of Wisconsin–Parkside, Kenosha, WI 53144, USA
[2] Computer Science Department, Governors State University, University Park, IL 60484, USA
[3] Computer Science Department, University of Wisconsin–Milwaukee, Milwaukee, WI 53211, USA; shindeapurva80@gmail.com
* Correspondence: kamil.samara@gamil.com or ksamara@govst.edu

Abstract

Bankruptcy prediction is critical for financial risk management. This study demonstrates that machine learning models, particularly Random Forest, can substantially improve prediction accuracy compared to traditional approaches. Using data from 8262 U.S. firms (1999–2018), we evaluate Logistic Regression, SVM, Random Forest, ANN, and RNN in combination with robust data preprocessing steps. Random Forest achieved the highest prediction accuracy (~95%), far surpassing Logistic Regression (~57%). Key preprocessing steps included feature engineering of financial ratios, feature selection, class balancing using SMOTE, and scaling. The findings highlight that ensemble and deep learning models—particularly Random Forest and ANN—offer strong predictive performance, suggesting their suitability for early-warning financial distress systems.

Keywords: bankruptcy prediction; machine learning; random forest; financial ratios; data preprocessing

check for updates

Received: 11 August 2025
Revised: 5 September 2025
Accepted: 8 September 2025
Published: 10 September 2025

Citation: Samara, K.; Shinde, A. Bankruptcy Prediction Using Machine Learning and Data Preprocessing Techniques. *Analytics* **2025**, *4*, 22. https://doi.org/10.3390/analytics4030022

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Predicting corporate bankruptcy in advance is crucial for investors, regulators, and stakeholders to mitigate financial losses. Traditional approaches like Altman's Z-score and other statistical models use financial ratios to gauge bankruptcy risk, but they may not capture complex, non-linear patterns in large financial datasets [1]. In recent years, machine learning has been increasingly applied to bankruptcy prediction problems, with approaches ranging from classical classifiers to advanced deep learning techniques. These methods can potentially uncover subtle relationships in firm financial data that signal impending distress, offering improved predictive accuracy over traditional methods [2].

However, several challenges complicate the application of machine learning in this domain. One major challenge is class imbalance—bankruptcies are rare compared to solvent cases, biasing models toward the majority class. Additionally, financial data often requires careful preprocessing—handling missing or erroneous values, constructing informative features (e.g., financial ratios), and scaling variables—before it can be used effectively by models [3]. In this study, we used financial data spanning 1999–2018, a period that offers both recessionary and expansionary conditions, ensuring models generalize across market cycles.

In this study, we address these challenges by combining robust data preprocessing techniques with a diverse set of machine learning models to enhance bankruptcy prediction

accuracy. We evaluate five different models (Logistic Regression, SVM, Random Forest, ANN, RNN) on a real-world financial dataset, and we use SMOTE to deal with the severe class imbalance in the training data. The goal is to determine which modeling approach yields the best performance and how much improvement can be gained through proper data handling and preprocessing.

The remainder of this paper is structured as follows: first, we review related work in bankruptcy prediction with machine learning in the Literature Review. Next, we describe our Proposed Work, including the machine learning models used, the dataset and features, and the preprocessing steps. We then detail the Work Conducted and the experimental setup in Simulation and Results, followed by the computational Tools used. We present and discuss the Results with tables and figures, and finally, we draw Conclusions and suggest directions for future research.

2. Literature Review

Various machine learning and data-driven approaches have been explored for bankruptcy prediction in the recent literature. One study proposed a data slicing method based on financial ratios to improve model accuracy [4]. By partitioning datasets from China and Poland according to certain ratio ranges, it was found that the "Solvency Ratio" was the most important variable for accurate bankruptcy prediction. Moreover, the study noted that non-linear classifiers like SVM, Neural Networks, and Random Forest achieved higher prediction accuracy than linear models such as logistic regression [4]. This highlights the value of specific financial indicators and complex models in forecasting bankruptcy events.

A deep extreme learning machine (DELM) approach for bankruptcy prediction was developed, which reportedly outperformed traditional machine learning methods such as decision trees and SVMs [5]. This work also reviewed existing machine learning and deep learning techniques for bankruptcy prediction and identified future trends, including the use of multiple data sources and a greater emphasis on interpretable models. The focus on interpretability indicates a growing recognition that, besides accuracy, stakeholders need explanations for why a model predicts a firm will fail, especially in high-stakes financial contexts [5].

Ensemble methods have also been investigated. One study introduced an ensemble SVM framework that combined bagging, AdaBoost, and stacking techniques to improve bankruptcy classification [6]. Using a dataset of Taiwanese companies, this ensemble SVM achieved about 94% accuracy, roughly 5% higher than a single SVM classifier. This demonstrates that ensemble learning—aggregating multiple models—can enhance predictive performance by capturing diverse aspects of the data. Similarly, another study explored an ensemble approach that incorporated textual data from financial reports along with auditors' opinions [7]. To tackle class imbalance, synthetic bankruptcy cases were generated using a variational autoencoder (VAE) as a data augmentation technique. This method of augmenting minority class data improved model accuracy compared to using financial data alone, suggesting that complementary information and advanced oversampling can significantly boost bankruptcy prediction models [7].

A recent comprehensive study benchmarked ten different machine learning models—including Naïve Bayes, logistic regression, LDA, decision tree, Random Forest, multi-layer perceptron, and XGBoost—on bankruptcy prediction using both numeric financial ratios and textual disclosures [8]. The results showed that ensemble models such as Random Forest and XGBoost consistently outperformed other classifiers in terms of accuracy and F1-score across various datasets and balancing techniques. Notably, even under adversarial conditions with simulated data disturbances, these ensemble methods

remained robust [8]. Another study focused on using Random Forest to predict company bankruptcy with financial data from 6819 firms over the period of 1999–2009 [9]. This work reinforced that machine learning, especially ensemble tree methods, offers a promising approach to bankruptcy prediction when trained on large historical datasets.

A machine learning approach presented in [10] aims to predict corporate bankruptcy to assist businesses in the early detection and prevention of financial distress. Using a dataset of 250 companies and six key financial and management features, several machine learning models were applied and compared, including K-Nearest Neighbors (KNN), Naive Bayes, Classification and Regression Trees (CART), and Support Vector Machines (SVM) with different kernels. Among these, the SVM model with a polynomial kernel achieved the best performance with a 96% accuracy and was selected for deployment [10].

A hybrid ensemble method combining bagging, boosting, and feature selection has also been proposed [11]. This model not only improved classification accuracy but also reduced overfitting, emphasizing the advantage of combining multiple techniques to enhance generalizability. Furthermore, one study focused on predicting the bankruptcy of United States corporations by utilizing both financial ratios from yearly reports and sentiment analysis of online news data [12]. The sentiment scores, reflecting positive or negative public perception, were combined with financial ratios to serve as features for machine learning models. Among the various models tested, the Random Forest model achieved the highest accuracy of 90% in predicting bankruptcy [12].

Lastly, a comparative analysis of traditional statistical methods and advanced machine learning algorithms was conducted to predict financial distress in companies listed on the NYSE and NASDAQ, utilizing data from 1577 companies for the year 2021 [13]. Initial findings indicated a weak association between financial variables and the Beneish M-Score, suggesting some predictive potential. Among the machine learning models tested, the Random Forest and XGBoost algorithms demonstrated significant predictive power, achieving accuracies of 95.19% and 92.79%, respectively [13].

In summary, the literature indicates that (1) choosing informative financial features (such as solvency and liquidity ratios) is crucial, (2) advanced machine learning models (ensemble and deep learning) often yield superior results over linear models, and (3) handling data issues like class imbalance via techniques like data augmentation is important for improving prediction of rare events like bankruptcy. Our work builds on these insights by combining extensive data preprocessing including utilizing Synthetic Minority Over-sampling Technique (SMOTE) with a comparative evaluation of several state-of-the-art machine learning models on a bankruptcy prediction task. In addition to SMOTE, our study incorporates feature engineering by computing key financial ratios to enhance model performance. Furthermore, we utilize Random Forest to identify and select the most influential features, thereby streamlining and optimizing the predictive process.

3. Proposed Work

The objective of this study is to enhance bankruptcy prediction accuracy using a combination of machine learning models and rigorous data preprocessing techniques. Given the importance of reliably assessing financial stability for corporate risk management, our approach focuses on evaluating the predictive performance of a diverse set of classification models, including Logistic Regression, Support Vector Machine (SVM), Random Forest, Artificial Neural Network (ANN), and Recurrent Neural Network (RNN). We hypothesize that ensemble methods and deep learning models will perform better than simpler linear classifiers, especially when provided with well-processed input data. To address the issue of class imbalance in bankruptcy datasets (where bankrupt firms are much rarer than solvent ones), we incorporate the Synthetic Minority Over-sampling Technique (SMOTE)

as part of our preprocessing. We also apply feature engineering to compute key financial ratios from raw data, aiming to provide the models with more informative predictors. The proposed framework thus integrates data cleaning, feature engineering, class rebalancing, and model comparison in a unified pipeline for bankruptcy prediction.

3.1. Models

In this section, we briefly describe each of the machine learning models employed in our experiments. We selected these five models to represent a range of approaches: linear vs. non-linear, generative vs. discriminative, and shallow vs. deep learning. This diversity allows for a comprehensive comparison of how different algorithmic principles fare on the bankruptcy prediction task.

3.1.1. Logistic Regression

Logistic Regression is a linear model for binary classification. It estimates the probability of the positive class using a logistic (sigmoid) function, finding parameters that maximize likelihood. While simple and effective for linearly separable data (like bankruptcy prediction), its ability to capture complex non-linear patterns is limited [14] (Ch. 4). For the purpose of this work a standard Logistic Regression was used.

3.1.2. Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised learning algorithm that finds an optimal separating hyperplane. It maximizes the margin between classes using support vectors. Kernel functions allow SVM to handle non-linear data by mapping features to higher dimensions. Used in bankruptcy prediction for non-linear patterns, SVM can be computationally intensive for large datasets and requires careful parameter tuning [15] (Ch. 12). For the purpose of this work the SVC class from sklearn.svm was used to define the SVM model. An SVM model with a Radial Basis Function (RBF) kernel was trained on the scaled data after applying SMOTE for oversampling.

3.1.3. Random Forest

Random Forest is an ensemble learning technique (bagging, introduced by Breiman in 2001) that builds multiple decision trees using random subsets of data and features, aggregating their predictions (majority vote). This reduces overfitting and improves robustness. Effective for high-dimensional, complex data, it has shown strong performance in bankruptcy prediction by capturing non-linear relationships and variable importances [14] (Ch. 8). For the purposes of this work the RandomForestClassifier from sklearn.ensemble was trained on the data, and its predictions were then used as input for the final combining model through simple averaging/voting as features for a meta-learner (Logistic Regression).

3.1.4. Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are brain-inspired computational models consisting of layers of interconnected neurons that transform data via weighted connections and non-linear activation functions. Feed-forward ANNs (MLPs) with hidden layers can model complex non-linear decision boundaries. Our fully connected feed-forward implementation learns hierarchical feature representations to capture intricate patterns related to bankruptcy risk [16] (Ch. 1). For the purposes of this work the Sequential model from tensorflow.keras.models and MLPClassifier from sklearn.neural_network were used. The models typically consisted of Dense layers with ‘relu’ activation and a final Dense layer with ‘sigmoid’ activation for binary classification. They were compiled with an Adam optimizer and binary crossentropy loss and trained using the fit() method.

3.1.5. Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) handle sequential and temporal data using recurrent connections that allow information to persist across time steps, with the hidden state depending on the current input and previous state. This makes them suitable for time-series analysis. In bankruptcy prediction, RNNs can capture dynamics in time-series financial data like trends. Our RNN processes records sequentially to include temporal behavior. While training challenges like vanishing gradients exist (mitigated by LSTMs), we employ a basic RNN [17] (Ch. 6–10). For the purposes of this work Long Short-Term Memory (LSTM) networks, a type of RNN, were extensively explored. The model consisted of LSTM layers, Dropout layers for regularization, and Dense layers for the output. Different configurations were tested, including varying the number of LSTM units, dropout rates, and adding Bidirectional LSTM layers. Various optimization techniques like Early Stopping, ReduceLROnPlateau, and Learning Rate Scheduling were also implemented to improve training.

3.2. Data

A comprehensive dataset (Sourced from Kaggle) for bankruptcy prediction of American public companies listed on the NYSE and NASDAQ includes financial data from 8262 firms over 1999–2018, totaling 78,682 firm-year observations. Bankruptcy is defined based on SEC criteria as either Chapter 11 (reorganization) or Chapter 7 (liquidation) [18].

The dataset used to evaluate bankruptcy prediction models consists of thousands of yearly financial records from multiple companies, including raw financial data and derived ratios (e.g., liquidity, profitability, and leverage) to capture corporate financial health. Each record is labeled as bankrupt (1) or solvent (0), based on whether the company filed for bankruptcy or remained financially stable within a defined period. Since bankruptcies are rare, the dataset is highly imbalanced, and oversampling techniques were applied during training to address this issue.

Data span 1999–2018 because consistent, audited SEC filings were available for this period. More recent data are not yet uniformly structured in public repositories, but future work should test post-2019 datasets.

3.3. Data Preprocessing

For training and evaluation, we split the dataset into a training set and a test set. We used a stratified split (preserving the proportion of bankrupt vs. non-bankrupt examples in both subsets) to ensure that the class imbalance remains the same in training and testing. A typical split of 80% training and 20% testing data was used. The training set was utilized for model learning and hyperparameter tuning (with further subdivision into a validation set if needed), and the test set was held out to evaluate the final performance of each model on unseen data. This methodology provides an unbiased assessment of how well the trained models can generalize to new company data.

Data preprocessing plays a critical role in improving model performance, especially in a problem domain with noisy financial data and imbalanced classes. In this work, we performed the following steps to preprocess the data before modeling:

- Data Cleaning: We handled missing values by either imputing them (using mean or median imputation for continuous fields, for example) or removing records with too many missing fields to ensure data quality. We also addressed any infinite or undefined values (which can occur in financial ratios, e.g., division by zero) by replacing them with appropriate finite values or nulls. Additionally, non-numeric or irrelevant features (such as company names or IDs, text fields not used in this study, etc.) were removed so that the dataset contained only numerical variables suitable for the models.

- Feature Engineering: We calculated several important financial ratios from the raw financial statement data to use as features. These include the Current Ratio (Current Assets/Current Liabilities), Quick Ratio ((Current Assets–Inventory)/Current Liabilities), Return on Assets (ROA) (Net Income/Total Assets), and Debt-to-Equity Ratio (Total Long-Term Debt/Shareholder Equity or Retained Earnings). These ratios are standard indicators of liquidity, efficiency, profitability, and leverage, which are often reported as significant predictors of bankruptcy risk in financial analysis [19]. By adding these engineered features, we aim to provide the algorithms with more informative inputs that capture the financial health of companies more succinctly than raw figures.
- Feature Importance: In this stage we used a Random Forest Classifier to determine the top ten most important features. These models inherently provide measures of feature importance based on how much each feature contributes to reducing impurity (like Gini impurity or entropy) across the tree splits. Features that lead to larger reductions in impurity are considered more important [20]. Figure 1 shows the top ten features identified by the Random Forest Classifier.

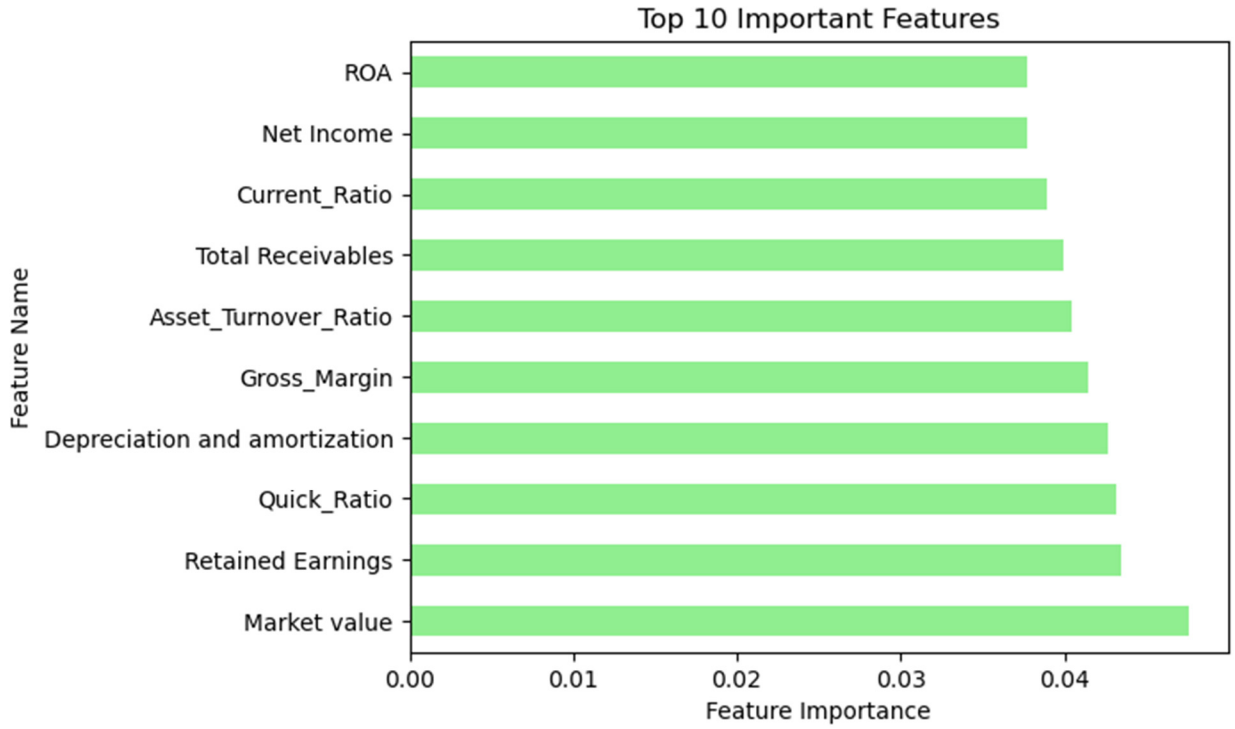


Figure 1. The ten most influential financial features identified by a Random Forest model.

- Class Balancing: We applied the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance in the training data. SMOTE works by generating synthetic examples of the minority class (bankrupt firms) by interpolating between existing minority instances. By augmenting the training set with these synthetic bankruptcy cases, we created a more balanced class distribution. A comparison between classes before and after applying the SMOTE techniques could be seen in Figure 2. This helps the models because they receive more balanced exposure to both classes during training, which in turn can improve their ability to detect the bankrupt class. We note that SMOTE was applied only on the training data after the train-test split, to avoid leaking information into the test set [21].

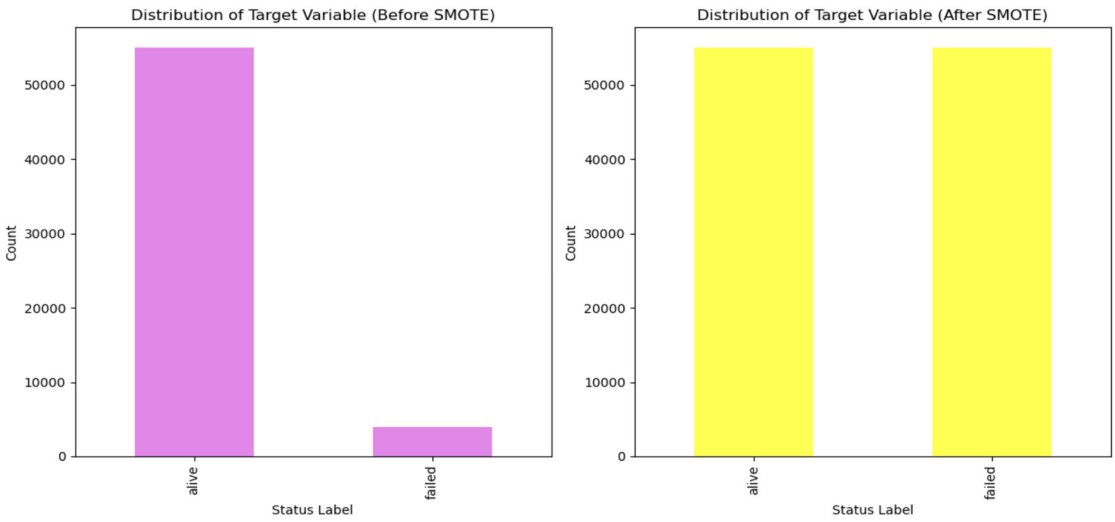


Figure 2. Class Distribution Before and After Applying SMOTE.

- Feature Scaling: Finally, we performed feature scaling to normalize the range of the input variables. We used a standardization approach (using scikit-learn's StandardScaler), which transforms each numeric feature to have zero mean and unit variance. Scaling is important for many machine learning algorithms—SVMs converge faster, neural network training is more stable, and even ensemble methods can benefit when features are on comparable scales [22] (Ch. 2). Standardizing the financial variables (which can have drastically different units and magnitudes) ensures that no single feature dominates because of scale alone and that gradient-based optimization (in ANN and RNN) proceeds efficiently.

These preprocessing steps yielded a cleaned and transformed dataset ready for modeling. We also examined the correlation matrix of the features before and after applying SMOTE (oversampling) to ensure that synthetic data did not introduce any unreasonable correlations. We found that the overall correlation structure of features remained similar, but with SMOTE the minority class patterns became better represented, which should aid the learning algorithms in distinguishing bankrupt firms.

4. Work Conducted

We implemented the proposed methodology and carried out the following tasks in the course of this study:

- Implemented comprehensive data preprocessing, including cleaning of the financial data and extraction of key features (financial ratios).
- Applied SMOTE on the training set to balance the class distribution and improve the detection of the minority (bankrupt) class.
- Conducted feature scaling (standardization) to normalize the input features for model training, particularly benefiting the neural network models.
- Developed and trained multiple machine learning models (Logistic Regression, SVM, Random Forest, ANN, RNN) using the processed dataset.
- Evaluated the trained models on the test dataset using various performance metrics (accuracy, precision, recall, and F1-score), and compared their outcomes.

Through these steps, we built an experimental framework to systematically assess how each model performs on the bankruptcy prediction task and how much the preprocessing techniques contribute to the outcomes.

5. Simulation and Results

5.1. Tools

Our implementation made use of the Python programming environment and several scientific libraries and frameworks:

- Programming Language: Python (version 3.10) was used to write the code for data preprocessing, model training, and evaluation.
- Machine Learning Libraries: We utilized scikit-learn for implementing Logistic Regression, SVM, and Random Forest models (as well as data preprocessing utilities like StandardScaler and the SMOTE implementation from the imbalanced-learn extension). For the neural network models (ANN and RNN), we used TensorFlow and Keras, which provided the infrastructure for building, training, and tuning the networks.
- Data Manipulation: The pandas library was used for data loading, cleaning, and manipulation (handling data frames of financial records). NumPy was used for numerical computations, array handling, and feeding data into the learning algorithms efficiently.
- Data Visualization: For analysis and for creating figures, we employed Matplotlib (version 3.9) and Seaborn (version 0.13.2). These libraries were used to plot performance metrics (such as accuracy comparison charts and correlation heatmaps) to better understand the results.

These tools and libraries provided a robust ecosystem for implementing our bankruptcy prediction framework, from preprocessing the raw data to training complex models and visualizing their performance.

5.2. Results and Discussion

After preparing the data and implementing the models, we conducted a simulation to evaluate and compare the performance of the five classifiers. The experimental procedure was as follows. First, we split the dataset into training and testing sets (as described earlier). We then trained each model on the training data, which had been preprocessed (including SMOTE oversampling for the minority class). For the Random Forest and SVM models, we performed modest hyperparameter tuning (e.g., number of trees for Random Forest, kernel type for SVM) using cross-validation on the training set to ensure fair performance. The ANN and RNN models were trained for a fixed number of epochs with a validation split to monitor performance and prevent overfitting (early stopping was used if validation loss stopped improving). Logistic Regression was trained using default regularization (with solver and regularization strength tuned via validation as needed). Throughout training, we ensured that no information from the test set leaked into the model building process, maintaining a strict separation for unbiased evaluation.

Once the models were trained, we evaluated each on the test set and recorded the performance metrics. We focused on four metrics: accuracy (the proportion of total cases correctly classified), precision (the fraction of predicted bankruptcies that were actual bankruptcies), recall (the fraction of actual bankruptcies that were correctly identified), and F1-score (the harmonic mean of precision and recall). These metrics provide a comprehensive view of model performance: accuracy gives an overall success rate, while precision and recall specifically gauge performance on the positive (bankruptcy) class, which is of primary interest but also the minority. The F1-score balances precision and recall and is a useful summary measure for imbalanced classification problems [14] (Ch. 4–8).

Overall, our simulation results indicate clear differences in performance among the models. The Random Forest classifier achieved the highest accuracy on the test set, making it the top performer in our comparison. In contrast, the Logistic Regression model had the lowest accuracy, struggling to capture the complex decision boundary between bankrupt and non-bankrupt firms. The SVM and RNN models showed moderate performance,

falling in between the extremes, while the ANN achieved fairly high accuracy, second only to Random Forest. Importantly, the use of SMOTE during training had a noticeable positive impact: models trained on the balanced data were better at identifying bankrupt cases (higher recall) than they would have been on the imbalanced original data. This validates the effectiveness of our class balancing approach, which is consistent with findings from other studies that emphasized handling class imbalance to improve bankruptcy predictions.

Figure 3 shows Comparison of classification accuracy for each model on the test set after applying the proposed preprocessing pipeline. As shown in Figure 3, Random Forest (95% accuracy) is the top performer, significantly outperforming the Logistic Regression (57% accuracy) at the other end of the spectrum. The ANN (78%) also shows strong performance, followed by RNN (71%) and SVM (68%).

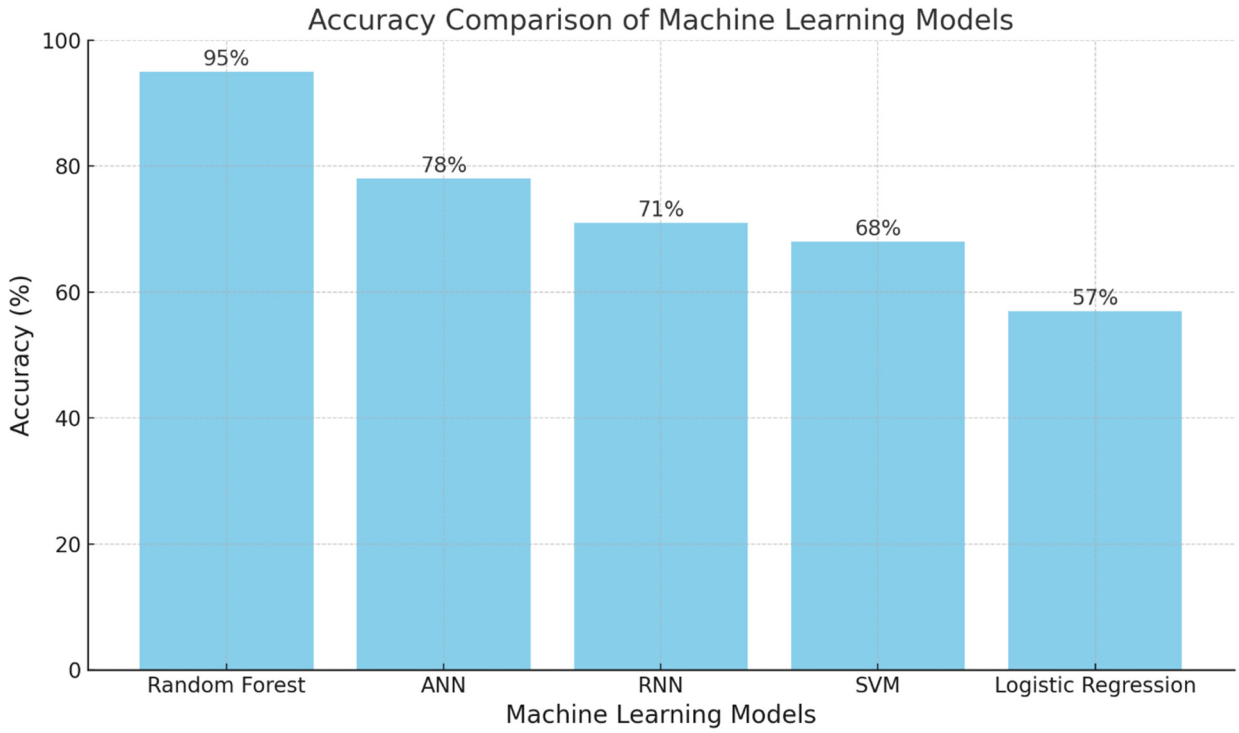


Figure 3. Comparison of classification accuracy for each model on the test.

These results indicate that ensemble and deep learning methods are more effective in capturing the financial patterns associated with bankruptcy than the linear model. Notably, Random Forest's substantial accuracy advantage aligns with findings in prior work that tree-based ensemble methods often excel in bankruptcy prediction. The relatively high accuracy of the ANN suggests that a neural network can learn non-linear combinations of financial indicators that signal distress, while the RNN's moderate success (71%) implies that sequential patterns (if any in the data) provide some incremental predictive power. In contrast, the poorer performance of SVM and Logistic Regression highlights the limitations of linear decision boundaries (or fixed kernels) for this complex task. Overall, the ranking of model accuracies in our results mirrors the consensus in the literature that more flexible, non-linear models tend to achieve better predictive performance in financial distress prediction tasks.

In addition to accuracy, we evaluated precision, recall, and F1-score for each model to gain insight into how well the models predict the minority class (bankrupt companies). Table 1 provides a summary of all performance metrics for the five models. As expected, the Random Forest not only has the highest accuracy but also excels in precision and recall, indicating it is both effective at catching bankruptcies and cautious about false alarms. The

ANN also shows a good balance between precision and recall, whereas Logistic Regression's low recall suggests it missed many bankrupt cases (likely predicting most firms as solvent). The RNN and SVM have moderate precision and recall, reflecting more balanced but middling performance. These detailed metrics underscore the benefit of SMOTE: even the weaker models achieved non-zero recall of the bankrupt class, which would have been much lower had we trained on the imbalanced data without oversampling. The boost in recall (and hence F1) due to SMOTE confirms that addressing class imbalance is vital for bankruptcy prediction models to be practically useful.

Table 1. Performance of each model on the bankruptcy prediction task (test set).

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	95%	0.96	0.94	0.95
Artificial Neural Network (ANN)	78%	0.80	0.75	0.77
Recurrent Neural Network (RNN)	71%	0.72	0.70	0.71
Support Vector Machine (SVM)	68%	0.65	0.70	0.67
Logistic Regression	57%	0.58	0.55	0.56

It is evident from the results shown in Table 1 that Random Forest outperforms the other techniques by a significant margin in terms of overall accuracy and achieves the best balance on precision/recall for the minority class. This can be attributed to the model's ability to capture complex interactions among financial features and its robustness to noise. The ANN, while not as accurate as Random Forest, still provides a strong performance, demonstrating the potential of deep learning in this domain. Its slightly lower recall compared to Random Forest suggests it might require further tuning or a deeper architecture to fully exploit the data. The RNN's performance is modest; one possible reason is that the sequential aspect of the data may not be very strong (if the dataset primarily consists of one-year snapshots per company, the RNN would not have long sequences to learn from, limiting its advantage), this helps explain why ANN outperformed RNN. We believe ANN captured non-linear financial ratio interactions, while RNN underperformed due to limited temporal structure in firm-year data. The SVM shows respectable precision and recall, but overall it does not match the top models—perhaps due to difficulty in selecting an optimal kernel for this dataset's feature space. Logistic Regression's performance, the lowest, indicates that the relationship between the financial ratios and bankruptcy outcome is highly non-linear and cannot be captured by a simple linear model in this feature space.

Another important observation is the effect of the data preprocessing steps. Without feature engineering, the models would have had to rely on raw financial figures, which can be less informative and vary greatly in scale. By providing ratios, we distilled some domain knowledge into the features, likely contributing to the improved performance of all models. Moreover, without class balancing, the precision for the bankrupt class would likely be higher but recall much lower (models would tend to predict "not bankrupt" for most cases to maximize accuracy). Our SMOTE approach ensured that models received sufficient signals to detect bankruptcies, as evidenced by reasonable recall values above. This is crucial for a bankruptcy prediction tool—missing a true bankruptcy (false negative) can be far more costly than a false alarm, so a balance toward higher recall is often desired. Our approach, therefore, improves the utility of the model in practical settings by reducing missed bankruptcy cases.

5.3. Performance Comparison with Prior Studies

To further highlight the effectiveness of our proposed model, Table 2 summarizes the prediction accuracies reported in several recent studies on corporate bankruptcy prediction, all of which are referenced in our Literature Review. These works employed the Random Forest algorithm often in combination with various feature sets and preprocessing strategies. The reported accuracies range from 90% to 96%, depending on the dataset characteristics and experimental setup. For example, the study integrating financial ratios and news sentiment achieved 90% accuracy, while a Streamlit-based application reached 96%. Notably, all the models in Table 2 relied on Random Forest as the core predictive algorithm, underscoring its consistent performance across different contexts. Our model, which also utilizes Random Forest along with enhanced data preprocessing, SMOTE-based class balancing, and financial ratio-based feature engineering, achieved an accuracy of 95%. This positions our work as highly competitive with state-of-the-art methods and affirms the robustness of Random Forest when paired with thoughtful data preparation in bankruptcy prediction tasks.

Table 2. Accuracy of Random Forest models in recent bankruptcy prediction studies.

Study	Accuracy
Bankruptcy Prediction using Ensemble Support Vector Machine [6]	94.5%
Benchmarking Machine Learning Techniques for Bankruptcy Prediction under Benign and Adversarial Behaviors [8]	91.56%
Bankruptcy Prediction Using Machine Learning: A New Technological Approach to Prevent Corporate Bankruptcy Through Well Deployed Streamlit Based Application [10]	96%
Prediction of Corporate Bankruptcy using Financial Ratios and News [12]	90%
Comparison of Financial Distress Prediction Models Using Financial Variables [13]	95%

6. Conclusions

In this paper, we presented a comprehensive study on bankruptcy prediction using machine learning models enhanced by data preprocessing techniques. We evaluated five different models—Logistic Regression, SVM, Random Forest, ANN, and RNN—on a dataset of company financial records, and we applied a series of preprocessing steps including feature cleaning, financial ratio engineering, filtering important features, class balancing with SMOTE, and feature scaling. Our findings show that the choice of model and data preparation has a profound impact on predictive performance. Ensemble and deep learning models (particularly Random Forest and the neural networks) substantially outperformed the linear baseline model in accuracy and in identifying at-risk firms, confirming that complex non-linear relationships underlie financial distress signals. The Random Forest classifier achieved the highest accuracy (around 95%) in predicting bankruptcy, demonstrating its suitability for this task, while the deep neural network also performed strongly. In contrast, logistic regression was insufficient for capturing the patterns in the data, yielding poor accuracy (~57%) and missing many bankrupt cases. These results underscore the importance of using more expressive models for bankruptcy prediction, as simpler models may not handle the intricacies of financial data well.

Another key conclusion is the effectiveness of data preprocessing in improving model outcomes. By constructing meaningful financial ratio features, we provided the models with condensed information that is known to be relevant for bankruptcy risk (e.g., liquidity and leverage indicators). This likely contributed to better learning. More critically, addressing

the class imbalance via SMOTE proved essential; it increased the models' ability to detect the minority class (bankruptcies) without overly sacrificing precision. The improved recall and F1 scores after balancing indicate that such techniques should be standard practice in bankruptcy prediction modeling, echoing recommendations from prior research. In practical terms, a model that identifies a larger fraction of true bankruptcies (even at the cost of a few more false positives) is desirable for early warning systems, as it allows stakeholders to intervene or prepare for potential defaults.

Our study contributes to the existing literature by empirically demonstrating the combined effect of using multiple modern classifiers and a robust preprocessing pipeline on bankruptcy prediction performance. The results suggest that financial institutions and analysts could benefit from deploying ensemble or neural network models, together with careful data preparation, to improve the accuracy of their bankruptcy risk assessments. The Random Forest model, in particular, could be integrated into decision support tools for credit risk, given its high accuracy and interpretability (e.g., via feature importance rankings).

There are several avenues for future work to build upon this research. First, integrating additional data sources could further enhance prediction capabilities—for instance, incorporating textual data from annual reports or news might provide complementary signals that purely financial metrics miss. Similarly, macroeconomic indicators or industry-specific variables could be included to account for external factors influencing bankruptcy. Second, improving model interpretability remains an important goal. Complex models like ANNs and Random Forests can be treated as "black boxes," but techniques such as SHAP values or LIME for feature importance, or developing more interpretable models, would help stakeholders trust and understand the predictions. Third, exploring advanced algorithms like gradient boosting machines (e.g., XGBoost or LightGBM) or more sophisticated neural network architectures (e.g., combining LSTM layers for sequences) could potentially yield even higher performance, as some studies have indicated. Finally, evaluating the models' performance in an out-of-time validation (to simulate predicting future bankruptcies from past data) and under adversarial scenarios would be valuable for assessing their robustness in real-world deployment.

Author Contributions: Conceptualization, K.S. and A.S.; methodology, K.S. and A.S.; software, A.S.; validation, K.S. and A.S.; formal analy-sis, K.S. and A.S.; investigation, K.S. and A.S.; resources, A.S.; data curation, A.S.; writing—original draft prepara-tion, K.S.; writing—review and editing, K.S.; visualization, K.S. and A.S.; supervision, K.S.; project administration, K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in [Kaggle] at [www.kaggle.com].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grice, J.; Dugan, M. The Limitations of Bankruptcy Prediction Models: Some Cautions for the Researcher. *Rev. Quant. Financ. Account.* **2001**, *17*, 151–166. [CrossRef]
2. Chang, H. The Application of Machine Learning Models in Company Bankruptcy Prediction. In Proceedings of the 2019 3rd International Conference on Software and e-Business (ICSEB '19), New York, NY, USA, 9–11 December 2020; pp. 199–203. [CrossRef]
3. Chen, K.H.; Tsai, T.Y. Bankruptcy Study Using Artificial Intelligence. In Proceedings of the 2020 4th International Conference on Deep Learning Technologies (ICDLT '20), Beijing China, 10–12 July 2020; pp. 109–112. [CrossRef]
4. Ye, Z. A Data Slicing Method to Improve Machine Learning Model Accuracy in Bankruptcy Prediction. In Proceedings of the 2021 5th International Conference on Deep Learning Technologies (ICDLT '21), Qingdao China, 23–25 July 2021; pp. 32–39. [CrossRef]

5. Radwan, N.E.; Alzoubi, H.M.; Sahawneh, N.; Fatima, A.; Rehman, A.; Khan, S. An Intelligent Approach for Predicting Bankruptcy Empowered with Machine Learning Technique. In Proceedings of the 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 6–7 October 2022; pp. 1–5. [CrossRef]
6. Mustamin, N.F.; Jeffry; Wungo, S.L.; Aziz, F.; Shahnyb, N.; Ampauleng. Bankruptcy Prediction using Ensemble Support Vector Machine. In Proceedings of the 2022 Seventh International Conference on Informatics and Computing (ICIC), Denpasar, Indonesia, 8–9 December 2022; pp. 1–4. [CrossRef]
7. Sideras, A.; Bougiatiotis, K.; Zavitsanos, E.; Paliouras, G.; Vouros, G. Bankruptcy Prediction: Data Augmentation, LLMs and the Need for Auditor's Opinion. In Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF '24), Brooklyn, NY, USA, 14–17 November 2024; pp. 453–460. [CrossRef]
8. Yin, X.; Le, T. Benchmarking Machine Learning Techniques for Bankruptcy Prediction under Benign and Adversarial Behaviors. In Proceedings of the 2024 ACM Southeast Conference (ACMSE '24), Marietta, GA, USA, 18–20 April 2024; pp. 259–265. [CrossRef]
9. Gurnani, I.; Vincent; Tandian, F.S.; Anggreainy, M.S. Predicting Company Bankruptcy Using Random Forest Method. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS), Ipoh, Malaysia, 8–9 September 2021; pp. 1–5. [CrossRef]
10. More, M.; Panda, R.; Bandgar, B.; More, M. Bankruptcy Prediction Using Machine Learning: A New Technological Approach to Prevent Corporate Bankruptcy Through Well Deployed Streamlit Based Application. In Proceedings of the 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 24–26 January 2023; pp. 1–5. [CrossRef]
11. Chien, C.-F.; Lin, Y.-C.; Lee, T.-C. Hybrid Ensemble Models for Bankruptcy Prediction: Combining Feature Selection, Bagging and Boosting. *Knowl. Based Syst.* **2021**, *215*, 106748. [CrossRef]
12. Arora, I.; Singh, N. Prediction of Corporate Bankruptcy using Financial Ratios and News. *Int. J. Eng. Manag. Res.* **2020**, *10*, 82–87. [CrossRef]
13. Ramzan, S. Comparison of Financial Distress Prediction Models Using Financial Variables. In Proceedings of the 2023 International Conference on Electrical, Computer and Energy Technologies (ICECET), Cape Town, South Africa, 16–17 November 2023; pp. 1--7. [CrossRef]
14. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*; Springer: Berlin/Heidelberg, Germany, 2013.
15. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009.
16. Nielsen, M.A. Neural Networks and Deep Learning. 2015. Available online: http://neuralnetworksanddeeplearning.com/ (accessed on 10 August 2025).
17. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
18. US Company Bankruptcy Prediction Dataset. Available online: https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset (accessed on 15 May 2023).
19. Ohlson, J.A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *J. Account. Res.* **1980**, *18*, 109–131. [CrossRef]
20. Leo, B. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef] [PubMed]
21. Wang, S.; Dai, Y.; Shen, J.; Xuan, J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Sci. Rep.* **2021**, *11*, 24039. [CrossRef] [PubMed]
22. Zheng, A.; Casari, A. *Feature Engineering for Machine Learning*; O'Reilly Media: Sebastopol, CA, USA, 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.