Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7, 238-259
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i7.8563

Financial statement analysis in the ERA of big data: A comparative study of machine learning techniques and traditional methodologies

Liuyu Sun[1*]
[1]A business college, Shanghai Industrial and Commercial Polytechnic, Shanghai, 201806, China; 5021652598@163.com (L.S.).

Abstract: In the modern world of big data, traditional financial statement analysis procedures are being challenged and supplemented with the use of machine learning (ML) techniques. This research aims to provide a comparative analysis of ML techniques and traditional financial statement analysis methodologies in predicting stock market reactions to earnings announcements. The traditional approach is based on financial measurements, particularly profitability ratios (PR). These ratios assist in forecasting stock movements based on past performance. In contrast, the ML approach processes vast amounts of financial and market data using advanced algorithms such as Dynamic Regularization Support Vector Machine with Random Forest (DRSVM-RF). The research uses a dataset of historical financial statements and market data, big data refers to the vast, intricate datasets that require innovative tools for storage, processing, and analysis together with data normalization using min-max normalization and feature extraction using Principal Component Analysis (PCA), to find important parameters like free cash flow and corporate characteristics. The findings show that the DRSVM-RF model produces the most accurate forecasts and the most abnormal returns. This research indicates that ML techniques outperform traditional methods in predicting accuracy (95.62%), providing a more dynamic and scalable approach to financial analysis, although traditional methods are still beneficial for stable predictions and risk mitigation.

Keywords: *Dynamic regularization support vector machine with random forest (DRSVM-RF), Financial analysis, Market data, Profitability ratios (PR), Stock movements.*

1. Introduction

Financial statement analysis involves examining and analyzing a firm's financial reports to better understand its financial health, performance, and sustainability [1]. The three primary financial reports that provide accurate information on a company's assets, liabilities, revenues, costs, and cash flows are the balance sheet, cash flow statement, and income statement [2]. This process transforms raw financial data into meaningful insights regarding operational efficiency, profitability, liquidity, and long-term solvency [3].

Financial statement analysis comprises several crucial steps. Initially, relevant financial statements are collected and organized, often covering multiple periods to facilitate temporal comparisons [4]. Several analytical techniques are then employed, such as ratio analysis, which involves calculating financial ratios to evaluate profitability, liquidity, efficiency, and leverage [5]. Trend analysis identifies patterns over time, while vertical and horizontal analyses compare financial statement elements within and across years. Finally, comparative analysis benchmarks a company's performance against competitors or industry standards, offering a broader interpretative framework [6].

Financial statement analysis is vital across industries and for various stakeholders. Investors utilize it to decide whether to invest in a company by assessing profitability and risk levels [7]. Creditors and lenders evaluate financial statements to determine a firm's capacity to service debt and meet financial obligations. Management employs financial analysis to improve operational efficiency, establish budgets,

History: Received: 9 April 2025; Revised: 6 June 2025; Accepted: 10 June 2025; Published: 3 July 2025
* Correspondence: 5021652598@163.com

and make strategic decisions [8]. Regulators and auditors scrutinize financial statements to ensure transparency and compliance with accounting standards. Additionally, potential business partners or acquirers perform financial analysis as part of due diligence during mergers, acquisitions, or strategic partnerships [9].

The benefits of financial statement analysis are extensive. It enhances understanding of a firm's financial condition, supporting better decision-making for stakeholders. It aids in risk management and strategic planning by revealing financial strengths and weaknesses [10]. Comparing historical and current financial data with industry benchmarks helps identify trends and performance metrics [11]. Financial analysis also facilitates forecasting future financial conditions and cash flows, enabling better planning and growth management. Moreover, it promotes accountability, boosts investor confidence, and contributes to overall financial stability and organizational prosperity [12].

Traditional financial statement analysis methods face challenges, including handling noisy, incomplete, and high-dimensional data, as well as accurately categorizing financial health or detecting fraud due to complex, nonlinear relationships. The DRSVM-RF method addresses these limitations by combining the robust feature selection and ensemble learning capabilities of Random Forest (RF) with the dynamic regularization and high generalization power of Support Vector Machines (SVM). This hybrid model effectively manages data irregularities, reduces overfitting, and uncovers intricate patterns, resulting in more accurate, consistent, and explainable financial analysis outcomes.

2. Related Works

Machine learning and extensive financial data were used in Chen, et al. [13] to forecast profit changes one year in advance. Annual size-adjusted gains to hedge portfolios range from 5.02% to 9.74%, indicating that the models had strong predictive potential. It exceeded traditional models and professional analysts' estimates, which were due to the ignored nonlinear predictor interactions and machine learning usage of more precise financial data.

It examined how business factors affected the integrity of financial statements for firms trading on the stock market as evaluated in [14]. From 2014 to 2020, it analyzed data from 2225 observations using regression models and ML methods. Profitability, business size, and Board of Directors diversity all correlated positively with financial statement excellence, but policy on dividends, ownership by the state, and company listing time all correlated negatively.

A stability prediction model based on conflicting factors that used sigmoid deep learning was proposed in [15]. The model detected possible stabilizations of affecting elements for attractive stock exchanges. Stability was projected using both sigmoid and non-sigmoid levels until the optimal level was reached. The model matched real and expected stock market fluctuations, with the sigmoid function adjusted to satisfy new ranges. The non-sigmoid layer was left unaltered to increase forecast precision.

Natural language processing was employed in Liu, et al. [16] to anticipate stock performance based on a database of analyst reports. The sentiment of the reports was graded as good, neutral, or negative. The data revealed that strong positive sentiment was related to increased excess returns and instantaneous volatility, whereas strong negative sentiment increases volatility and trade volume but lowers future excess returns. The effect was stronger for positive mood reports. It added to the empirical data on sentiment analysis and how the stock market responded to the data.

As explained in Mabelane, et al. [17] 13 financial ratios from 2012 to 2018 financial statements were utilized to develop an acceptable model for assessing and forecasting municipal audit results. Three ML techniques were examined: logistic regression models, decision trees, and artificial neural networks (ANNs). The findings demonstrated that ANN was an extremely effective system for forecasting audit opinions.

A strong, deep learning-based time series technique for estimating future stock market prices was proposed in [18]. The system used a Conventional neural network-long short-term memory (CNN-LSTM) combination to predict the closing prices of Tesla and Apple. In terms of stock market price forecasting, the CNN-LSTM approach beat both the single deep learning LSTM and traditional

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

techniques. Artificial intelligence models have grown increasingly important as market capitalization has increased, and precise predictions are required.

To identify financial mistakes in the general ledger (GL) record, employing both supervised and unsupervised ML methodologies was investigated in [19]. Seven supervised ML techniques and two unsupervised techniques were used by the analysts. Models were trained on a real dataset of GL and removed the entry size variation by data vectorization. The findings presented a great possibility for the identification of abnormalities and efficient sampling of data.

The Hybrid Financial Risk Predictor (HFRP) model, created with CNN-LSTM, enhanced financial risk prediction by merging techniques from the field of finance [20]. The model achieved great accuracy and stability, which considerably reduced training and testing losses. It also considerably reduced credit, liquidity, market, and operational risks, increasing financial stability and functioning as a dependable model for current financial markets. The technique promoted smart decision-making and risk assessment.

Credit card usage has expanded due to technological improvements and e-commerce, resulting in higher financial transaction volume. Detecting fraudulent activity was critical in Hashemi, et al. [21] and it investigated class weight-tuning hyperparameters for controlling both fraudulent and valid transactions. Bayesian optimization was used to improve hyperparameters while maintaining imbalanced data. Deep learning fine-tuning was employed to adjust the hyperparameters.

An ML strategy for detecting and forecasting financial fraud in publicly traded organizations was suggested in [22]. The analysis gathered 18,060 transactions and 363 financial indicators, discarding 9 unrelated ones and retaining 13 important ones. Five single classification and three combination models were created, with the best single model obtaining better accuracy. The ensemble model outperformed, suggesting its efficacy in identifying fraudulent behavior.

Data mining and ML technologies to enhance trade finance firms' ability to deal with excessive financialization were employed in [23]. A risk assessment model was developed using genetic algorithms, neural networks, and PCA techniques. The model's performance was evaluated using particular scenarios, demonstrating that, while most businesses do not face increased financial risk, assets were increasingly financialized. The model's accuracy rate exceeds 91%, with 80% achieved even with small sample sizes.

2.1. Problem Statement

Traditional financial forecasting models have significant limitations. Machine learning in Chen, et al. [13] successfully predicts profits but does not account for the intricate nonlinear relationships among financial factors. It also does not have provisions for modeling temporal dependencies, which are essential for discovering changing financial trends. The model in Wang [15] is based on sigmoid deep learning but is incapable of learning long-term dependencies in sequential data. It constrains its capacity to cope with dynamic and volatile stock market activities. Fraud detection models in Zhao and Bai [22] are highly accurate but mostly static, ignoring the time-evolving aspects of fraudulent behavior. The proposed DRSVM-RF addresses these challenges by combining dynamic regularization, robust feature selection, and temporal learning for enhanced predictive performance.

3. Methodology

Financial analysis involves scrutinizing the financial health and performance of a firm by examining the financial documents and stock market records. The dataset integrates past financial reports and stock market data to guarantee robust model calibration in different market scenarios. Data preprocessing involved Min-Max normalization to normalize variables evenly, and this led to improved algorithms and faster convergence. Feature extraction using PCA is performed to acquire lower dimensions and identify important predictors like free cash flow and firm-specific variables. The proposed DRSVM-RF method combines nonlinear classification and ensemble learning, which improves

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

forecasting performance and detects market anomalies following earnings announcements. Figure 1 shows the overall suggested flow for financial statement analysis.

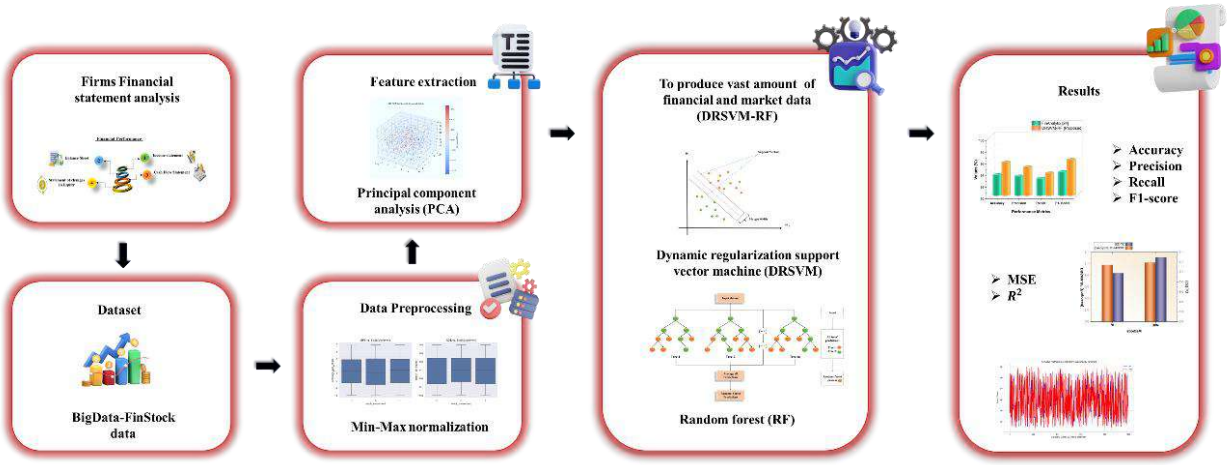

Figure 1.
Overall proposed flow for financial statement analysis.

3.1. Dataset

The BigData-FinStock dataset is designed to allow for the evaluation of traditional financial statement measures and contemporary methodologies for forecasting stock market movements after earnings releases. It integrates comprehensive data points such as financial ratios (e.g., Return on Assets, Debt to Equity, Earnings per Share (EPS)), market indicators (e.g., stock prices around announcements, volatility), and firm-specific characteristics. It also includes derived measures such as anomalous return and a categorical stock movement label, making it ideal for financial modeling, predictive analytics, and big data-driven investing strategies.
Source: https://www.kaggle.com/datasets/ziya07/bigdata-finstock/data

3.2. Data Exploration

The correlations between return on assets (ROA), net profit margin (NPM), and return on equity (ROE) as shown by stock movement (-1: decrease, 0: neutral, 1: rise) are shown in **Figure 2**. The KDE plots along the diagonal illustrate that most businesses have favorable profitability indicators. However, the overlap across multiple stock movement categories shows that these measurements alone lack a significant impact on short-term stock success. The scatter plots show modest correlations between the variables, indicating minimal dependency. Overall, while ROA, ROE, and net profit margin are important indications of financial health, they cannot be sufficient to forecast market reactions in financial statement analysis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

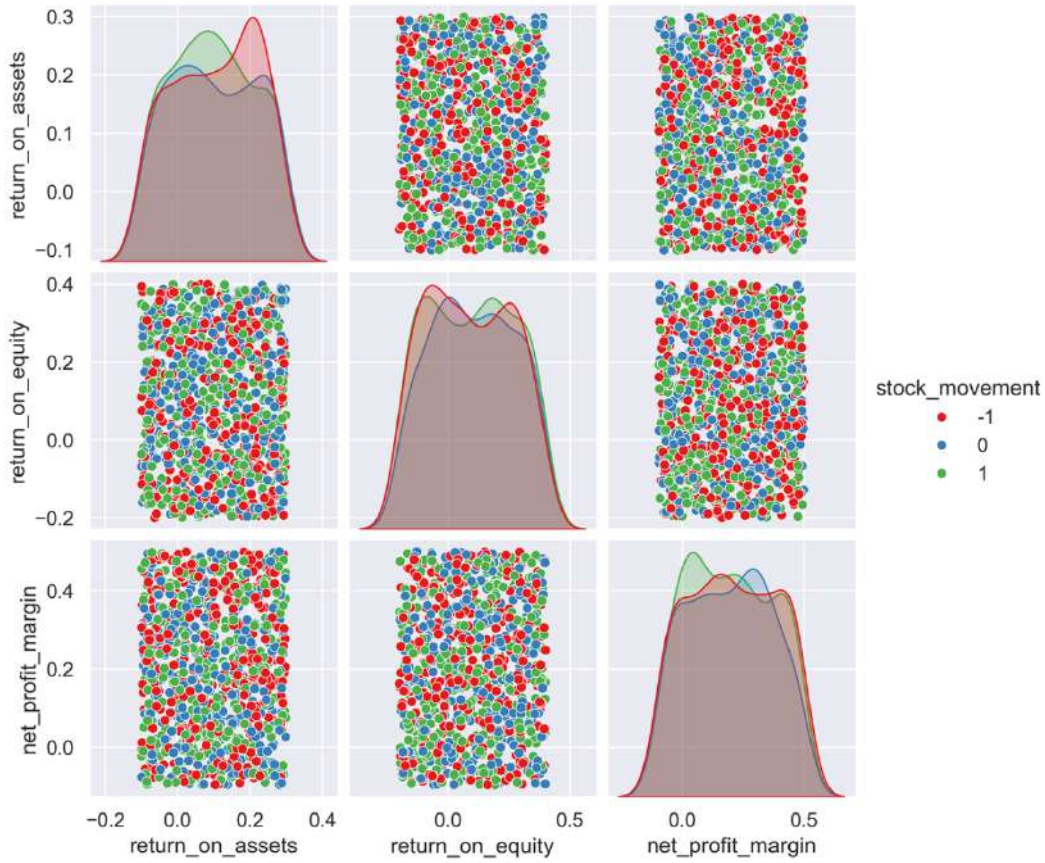

Figure 2.
Data exploration for the BigData-FinStock dataset.

3.3. Data Preprocessing Using Min-Max Normalization

Min-Max normalization adjusts financial statement data to a specific range, often [0, 1], improving comparability across financial measures. This method ensures that elements such as revenue, assets, and liabilities contribute appropriately to the analysis, hence enhancing model accuracy and efficiency as shown in Figure 3. Min-max normalization involves applying linear modifications to the original data to provide a balanced comparison of values before and after the operation. This approach could involve the following equation (1).

$$W_{new} = \frac{W - \min(W)}{\max(w) - \min(w)} \quad (1)$$

Where W_{new} indicates the new value obtained from the normalized findings, W equals the old value, $\max(W)$ is the highest value, and $\min(W)$ represents the minimum value.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

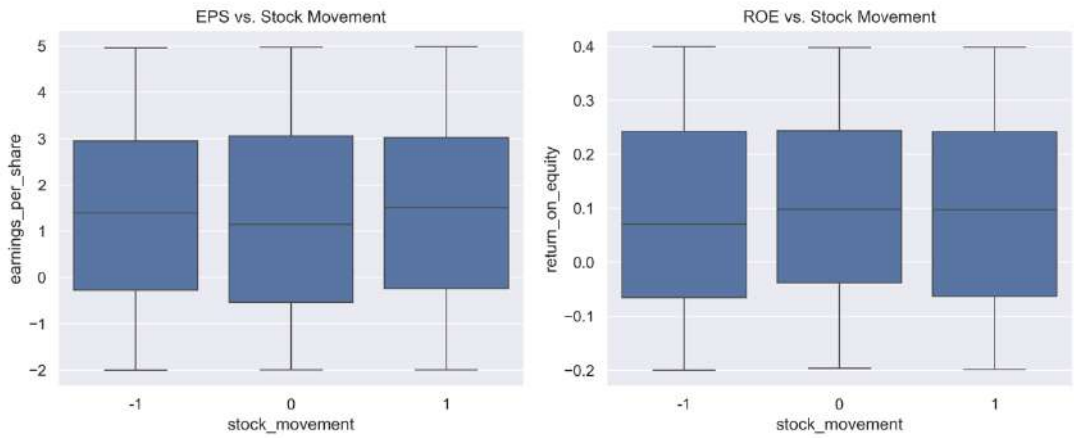

Figure 3.
Preprocessed features in the BigData-FinStock data using Min-Max normalization.

3.4. Feature Extraction Using PCA

PCA simplifies financial statement analysis by reducing data dimensionality and identifying significant aspects such as profitability, liquidity, and solvency ratios. It reveals hidden patterns, reduces redundancy, and improves decision-making by concentrating on the most informative elements. PCA is an information-driven modeling approach that reduces correlated variables to a smaller set of uncorrelated variables while preserving most of the original data.

Consider the dataset W, where each column represents a sequence of n-dimensional inputs. Consider that each function in the set of values has an average value of zero $(F(W) = 0)$. A typical original data matrix has m samples and n variables, as shown in equation (2):

$$W = [w_1, w_2, \ldots\ldots, w_m]^S = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix} \quad (2)$$

PCA transforms meteorological variables and performance factors into a new event space while preserving as much dataset as possible from the original data.

The largest variance directions in the input data are identified and projected to a new subspace with equal or lower dimensions than the original space.

An orthonormal transformation Y can move W to a new space S, as shown below in equation (3):

$$S = YW \quad (3)$$

The S-matrix of scores consists of orthonormal vectors derived from W-matrix components, which represent how samples are connected. The covariance matrix of S is in equation (4):

$$D_S = YD_WY^S \quad (4)$$

D_W represents the covariance matrix of W.

The loading matrix Y can be calculated using the eigenvalue equation as follows in equation (5):

$$(D_S - \lambda J)f_j = 0 \quad (5)$$

The covariance matrix holds pairwise covariances between input variables. The extracted features from the data are shown in Figure 4.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

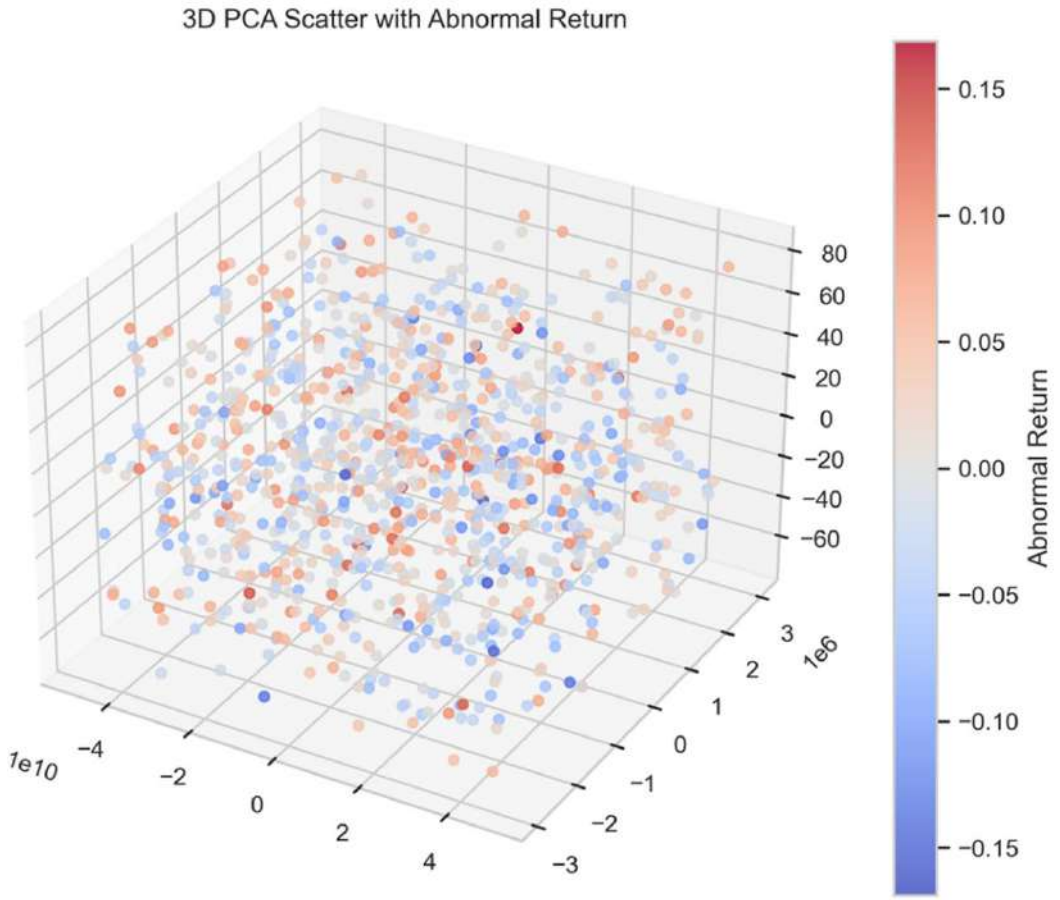

Figure 4.
Relevant features are extracted from the BigData-FinStock data using PCA.

3.5. Traditional Approach

The traditional method of using profitability ratios in financial statement analysis involves assessing a company's capability to earn relative to sales, assets, or equity. Important ratios such as gross profit margin, ROA, net profit margin, and ROE indicate operational efficiency, cost control, and overall profitability. By comparing these ratios with past periods or against industry standards, analysts can evaluate the firm's financial well-being, trends in performance, and management's effectiveness in making profits. This method is extensively utilized because of its ease of use and direct relationship with profitability results.

3.5.1. PR

This ratio evaluates managerial performance based on a firm's earnings from sales and investments. There are two sorts of probabilities: one based on sales and one based on investment. These two profitability ratios measure a company's operational performance. The Gross Profit Margin (GPM) is determined as follows in equation (6):

$$GPM = \frac{Gross\ profit}{Net\ sales} \tag{6}$$

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

3.5.2. NPM

This ratio measures a company's performance over time and provides insight into its management capabilities. This ratio represents the net profit from each sale in Rupiah. This ratio compares net income to net sales and is calculated as follows in equation (7):

$$NPM = \frac{Net\ income}{Net\ sales} \quad (7)$$

3.5.3. Time Interest Earned (TIE)

This ratio compares the amount of interest owed in the current year to operational profit. This ratio indicates that the firm can generate operational profit by lowering selling and operating costs compared to total revenue. The ratio can be computed as follows in equation (8):

$$TIE = \frac{interest\ payment}{Operational\ profit} \quad (8)$$

The main problem with traditional profitability ratio analysis is that it relies on fixed financial data, which fails to represent market shifts and irregular patterns. The proposed DRSVM-RF improves on the previous techniques by using advanced machine learning methods that represent complex links in finance, yield better accuracy, and responsively change to new data to ensure solid analysis of financial statements.

3.6. DRSVM-RF

The DRSVM-RF approach links SVM and RF together to give more strength to financial statement analysis. With dynamic regularization, the SVM module can manage its penalties as it deals with financial data of varying complexities. Such dynamic adjustments make it possible to include more data without overfitting since financial records usually contain noise, outliers, and unusual patterns. Meanwhile, feature selection and initial classification are carried out using the RF method, which is well known for its ensemble learning capabilities. By assessing the significance of each financial indicator (for example, liquidity ratios, profitability margins, and solvency measurements), RF determines the most relevant characteristics, reducing dimensionality and improving interpretability. The revised feature set is subsequently fed into the dynamically regularized SVM, which performs specific classification tasks including recognizing financial distress, measuring credit risk, and forecasting company performance. This two-stage design greatly enhances classification accuracy, robustness, and the dependability of decision-making. DRSVM-RF, therefore, provides a strong analytical tool for finance stakeholders, allowing for more precise, data-driven insights from complicated financial statements in real-world company environments.

Algorithm 1: DRSVM-RF

Input:
- Dataset D = {X, y}
- Initial regularization parameter C_init
- Regularization update rule ΔC
- Number of RF trees T
- SVM kernel function K (e.g., RBF, linear)
- Stopping criteria ε

Output:
- Trained DRSVM-RF model
- Classification predictions

1. Preprocess dataset D:
   a. Normalize or standardize X
   b. Split D into training and testing sets
2. Train Random Forest (RF) on training data:

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

a. $RF\_model \leftarrow Train\_RF(X\_train, y\_train, T)$
b. Feature_Importances ← Extract_Importances(RF_model)
3. Select important features based on RF:
a. Selected_Features ← Top-k features using Feature_Importances
b. $X\_train\_sel \leftarrow X\_train[:, Selected\_Features]$
c. $X\_test\_sel \leftarrow X\_test[:, Selected\_Features]$
4. Initialize:
a. C ← C_init
b. previous_accuracy ← 0
5. Repeat until convergence:
a. Train SVM_model ← Train_SVM(X_train_sel, y_train, C, Kernel=K)
b. Predict y_pred ← SVM_model.predict(X_test_sel)
c. Compute accuracy ← Accuracy(y_test, y_pred)
d. If |accuracy - previous_accuracy| < ε:
Break
Else:
i. Update regularization parameter:
C ← Update_C(C, accuracy, ΔC)
ii. previous_accuracy ← accuracy
6. Combine RF and SVM results (optional ensemble):
a. RF_pred ← RF_model.predict(X_test_sel)
b. Final_pred ← Majority_Vote(y_pred, RF_pred)
7. Return Final_pred and trained models (SVM_model, RF_model)

3.6.1. RF

RF is a sophisticated ensemble learning algorithm used in financial statement analysis to improve forecasting accuracy. It identifies critical financial indicators, finds abnormalities, and allows for more informed decision-making in credit scoring, fraud detection, and risk assessment by aggregating several decision trees. RF combines various classifiers that are generated during the learning process. The first training data are used to create several classification and regression trees (CART) based on bootstrap selection. The best place to split is decided by comparing data from a set of input variables chosen at random. Initially, the training data, at the root node, are divided into child nodes by CARTs. RF operates by letting each tree decide its vote for input w nd the classifier determines its output by taking the most common answer. RF can handle many variables and uses a large set of trees in its model. RF features that are very important include:

- It provides an excellent strategy for predicting missing data.
- The weighted random forest (WRF) algorithm balances faults in unequal data.
- Estimates the relevance of classifying variables.

3.7. Splitting Criterion

RF, CART, selects the split with the lowest impurity at each node by using the Gini measure. Gini impurity measures the distribution of class labels within a node. The Gini impurity is a number between $0\ and\ 1$, where 0 means that every item in a node belongs to a similar category. For the parameter $W = \{w_1, w_2, \ldots\ldots, w_i\}$ at node s, the Gini impurity is found using i for the number of children at s, M for the number of samples, m_{dj} for the number of samples with a value w_j of class d and b_j for the number of samples with a value w_j at node s.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

$$J\left(s_{w_j}\right) = 1 - \sum_{d=0}^{d}\left(\frac{m_{dj}}{b_j}\right)^2 \quad (9)$$

To get the Gini index of splitting, first take the weighted average of the Gini values for the variable W.

$$Gini(s, W) = \sum_{j=1}^{i}\frac{b_j}{M}J\left(s_{w_j}\right) \quad (10)$$

The attributes with the lowest Gini impurity values will indicate the right way to split the dataset. Every tree in RF takes a distinct group of n variables to produce its splitting rules.

3.8. Variable Importance

The changing significance output is an important component of RF. Variable significance assesses the strength of the association between a variable and its grouping results. These four measurements are used by RF to see how important each variable is: the score for the class 0, the score for class 1, the change in accuracy, and the Gini Index. Out-of-bag (OOB) data are sent down the tree, and the parameter's importance is figured out by measuring the prediction accuracy. When the values for a parameter j are swapped within the OOB samples, the parameter's accuracy is once again evaluated. All calculations are made on a tree-by-tree basis during RF building. The importance of the variable i is then estimated by averaging the accuracy loss of these combinations across all trees. A substantial decline in prediction accuracy suggests a close connection between the response and variable I. A set of parameters can be rated by RF according to their relative importance. The RF structure is demonstrated in Figure 5.

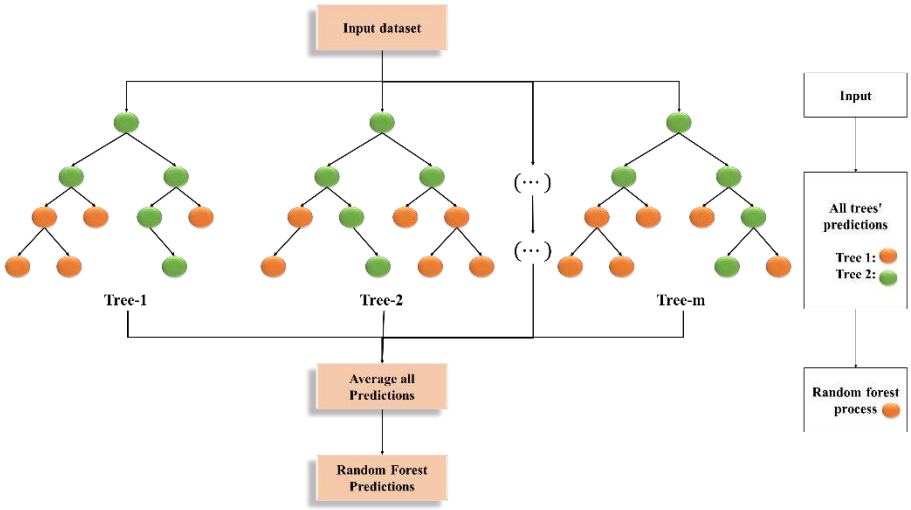

Figure 5.
Random forest structure.

Let β_s represent the OOB sampling for the tree s, $s \in \{1, \ldots, mtree\}$, and $z_j{}'^s$. In a tree s, j stands for the class predicted, for instance, j before permutations, while $z_j{}'^s$ means the class that is expected after the permutations. In tree s, the variable significance VI for the parameter i could be calculated as in equation (11)

$$VI_i^s = \frac{\sum_{j=1}^{M}\beta_s J(\gamma_j = \gamma_s'^s)}{|\beta_s|} - \frac{\sum_{j=1}^{M}\beta_s J(\gamma_j = \gamma_{j,\alpha}'^s)}{|\beta_s|} \quad (11)$$

The raw significance value of variable i is calculated as an average over all trees in the RF (equation 12).

$$VI_i = \frac{\sum_{s=1}^{mtree} VI_n^s}{mtree} \quad (12)$$

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

It relies on the Mean Decrease Gini (MDG) significance measure that comes from the Gini splitting criteria. The MDG is used to measure the change in ΔJ that coincides with the splitting reaction (equation 9). Reducing the group impurities (equation 13) from the weighted mean Gini of the node (equation 10) results in the quantity J (equation 14) for a two-class model.

$$J(s) = 1 - \sum_{d=0}^{d} \left(\frac{m_j}{M}\right)^2 \quad (13)$$

$$\Delta J(s) = J(s) - Gini(s, W) \quad (14)$$

Gini Importance (GI) is calculated by recording the reduction in Gini impurity across all nodes (s) in all trees $(mtree)$ in RF for all variables (equation 15).

$$GI = \sum_{mtree} \sum_{s} \Delta J(s) \quad (15)$$

3.9. DRSVM

DRSVM improves financial statement analysis by automatically modifying regularization environments based on data complexity. This method enhances classification accuracy by successfully detecting financial abnormalities, fraud, and performance trends in dynamic and unbalanced financial information. The empirical risk function gives each insensitive mistake between predicted and actual values the same weight D. The compromise of risk measurement and the standard term is determined by the regularization constant D. The empirical risk exceeds the established term as D rises in importance. As an illustration, the empirical risk factor is provided as follows in equation (16):

$$F_{SVMs} = D \sum_{j=1}^{m} (\zeta_j + \zeta_j^*) \quad (16)$$

DRSVM use a weighted function as the regularization constant, rather than a constant number (Equation 17-18).

$$F_{d-ASVMs} = \sum_{j=1}^{m} D_j (\zeta_j + \zeta_j^*) \quad (17)$$

$$D_j = w(j)D \quad (18)$$

Where $w(j)$ is the weighted function fulfilling $w(j) > w(j-1), j = 2, \ldots, m$. The ascending regularization constant D_j gives more weight to the latest data used for training points, as weights tend to shift from distant to nearby locations. The descriptions are as follows.

Regarding the linear weighting function in equation (19),

$$w(j) = \frac{j}{m(m+1)/2} \quad (19)$$

That is in equation (20),

$$D_j = D\frac{j}{m(m+1)/2} \quad (20)$$

Regarding the exponential weighting function in equation (21),

$$w(j) = \frac{1}{1+\exp\left(b-\frac{2bj}{m}\right)} \quad (21)$$

That is in equation (22),

$$D_j = D\frac{1}{1+\exp\left(b-\frac{2bj}{m}\right)} \quad (22)$$

The exponential weighting function is modeled after the discounted least squares (DLS), with a parameter controlling the ascending rate. Figure 6 demonstrates the weighting function's behavior, which could be stated as follows.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

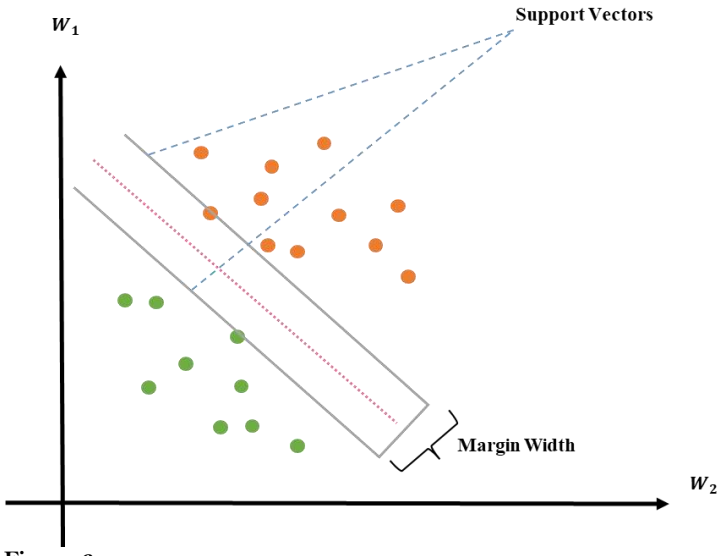

Figure 6.
DRSVM structure.

When $b \to 0$ then $\lim_{b\to 0} D_j = \frac{1}{2}D$. In this scenario, all training data points have identical weights, resulting in $F_{D-ASVMs} = \frac{1}{2}F_{SVMs}$ (equation 23).

When $b \to \infty$, then

$$\lim_{b\to 0} D_j = \begin{cases} 0, & j < \frac{m}{2} \\ D, & j \geq \frac{m}{2} \end{cases} \tag{23}$$

The weights for the initial half of the training information are zero, while those for the subsequent half are set to one (equation 24).

$$F_{D-ASVMs} = \begin{cases} 0, & j < \frac{m}{2} \\ F_{SVMs}, & j \geq \frac{m}{2} \end{cases} \tag{24}$$

As the number grows from $b \in [0, \infty]$, the initial half of the data has smaller weights, while the second half has heavier weights later in training.

Thus, the standardized threat function is computed as follows in equation (25):

$$minimize \quad Q_{D-ASVMs}\left(x, \zeta^{(*)}\right) = \frac{1}{2}||w||^2 + \sum_{j=1}^{m} D_j\left(\zeta_j + \zeta_j^*\right)$$

Subject to,

$$c_j - w\phi\left(x_j\right) - a_j \leq \varepsilon + \zeta_j,$$
$$w\phi\left(x_j\right) + a_j - c_j \leq \varepsilon + \zeta_j^*$$
$$\zeta^{(*)} \geq 0 \tag{25}$$

The dual function remains in its initial state (equation 26).

$$Q\left(b_j, b_j^*\right) = \sum_{j=1}^{m} c_j\left(b_j - b_j^*\right) - \varepsilon \sum_{j=1}^{m}\left(b_j - b_j^*\right)$$
$$-\frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{m}(b_j - b_j^*)(b_j - b_j^*)L(x_j, x_i) \tag{26}$$

However, the limitations have altered as follows in equation (27):

$$\sum_{j=1}^{m}\left(b_j - b_j^*\right) = 0$$
$$0 \geq b_j \leq D_j, \quad j = 1,2, \ldots\ldots, m$$
$$0 \geq b^*{}_j \leq D_j, \quad j = 1,2, \ldots\ldots, m \tag{27}$$

The essential hyperparameters for SVM, Random Forest, and hybrid models are covered in table 1, including regularization, kernel choice, tree count and depth, feature selection, and voting technique, which allow for dynamic customization for best classification performance and balanced model complexities.

Table 1.
Hyperparameters for DRSVM-RF.

Hyperparameter	Symbol	Example Value	Range	Component	Description
Initial Regularization	C_init	1.0	(0.01 – 1000)	SVM / Dynamic Reg.	Starting value for SVM regularization parameter C.
Regularization Step	ΔC	0.1	(0.001 – 1.0)	Dynamic Reg.	Step size for adjusting the regularization parameter dynamically.
Kernel Function	K	RBF	['linear', 'rbf', 'poly', 'sigmoid']	SVM	Specifies the kernel type used by SVM.
Convergence Threshold	ε	0.01	(0.001 – 0.1)	Dynamic Reg.	Minimum change in accuracy to stop updating C.
Max Iterations (opt.)	max_iter	10	(1 – 100)	Dynamic Reg.	Optional limit on the number of update loops.
# of Trees	T	100	(10 – 1000)	Random Forest	Number of decision trees in the Random Forest.
Max Tree Depth	max_depth	None	(5 – 100 or None)	Random Forest	Maximum depth of trees. None means full growth.
Min Samples per Split	min_samples_split	2	(2 – 10)	Random Forest	Minimum number of samples to split an internal node.
Feature Selection Count	k	10	(5 – 30 or % of total)	Feature Selection	Number of top features selected based on importance from RF.
Voting Strategy	vote_mode	Majority Vote	['majority', 'weighted']	Ensemble (optional)	Strategy to combine SVM and RF predictions.

4. Results and Discussion

In the big data era, financial statement analysis is aided by modern machine learning algorithms, which are developed and implemented using Python. Existing approaches, such as FinAnalytix [24] combine ML's pattern recognition capabilities with financial statement analysis to identify patterns and risks, whereas Decision Trees (DT) [25] provide explicit, rule-based insights. The proposed DRSVM-RF improves on these by combining SVM precision with the RF ensemble strength. Its dynamic regularization responds to data complexity, improving prediction accuracy and resilience in real-time decision-making.

4.1. Experimental Results

The relationship between numerous financial and market variables and anomalous return, which is an important measure of stock performance relative to expectations, is depicted in Figure 7. Among all characteristics, the stock movement exhibits the largest positive correlation (0.88) with anomalous return, demonstrating that significant price swings following announcements are strongly linked to abnormal returns. Traditional financial statement measures, such as net profit margin, gross margin, and free cash flow, exhibit poor or negative correlations, implying that these indicators have limited predictive potential for unexpected short-term returns.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

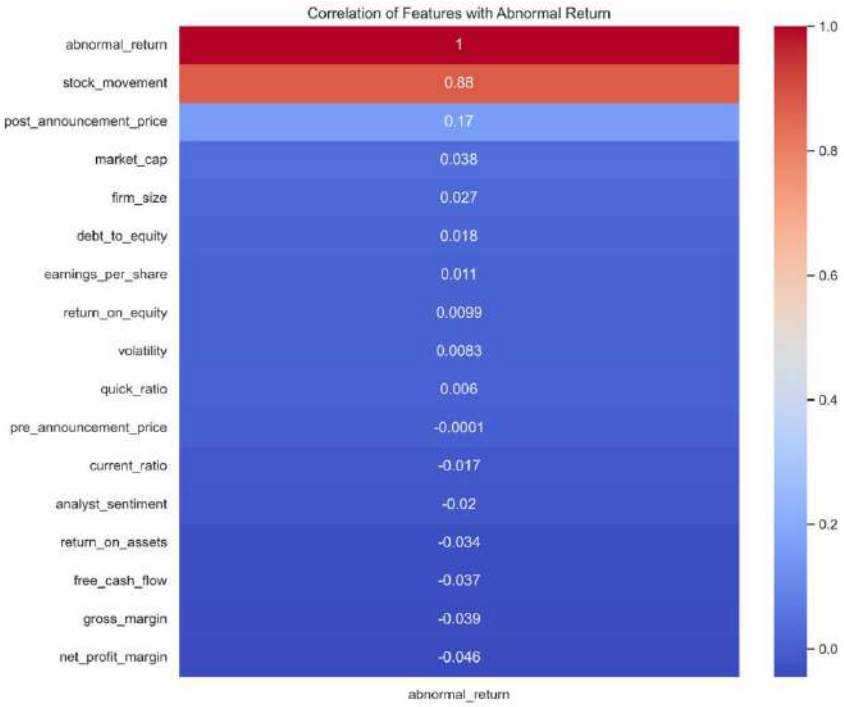

Figure 7.
Correlation of features with abnormal return of financial statement analysis.

The link between free cash flow (FCF) and EPS, with bubble size reflecting market capitalization and color signifying stock movement (positive red, negative blue, and neutral grey), was shown in Figure 8. The distribution shows that there is no clear linear link between FCF and EPS since enterprises are distributed throughout the axes. Larger market-cap businesses (larger bubbles) appear at all levels of FCF and EPS, demonstrating that size does not influence performance indicators. The mix of red and blue across the plot demonstrates that stock movement is not always linked with FCF or EPS, indicating that investor reactions to financial indicators can vary greatly and are likely influenced by larger market or sentiment-driven elements in financial statement evaluation.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

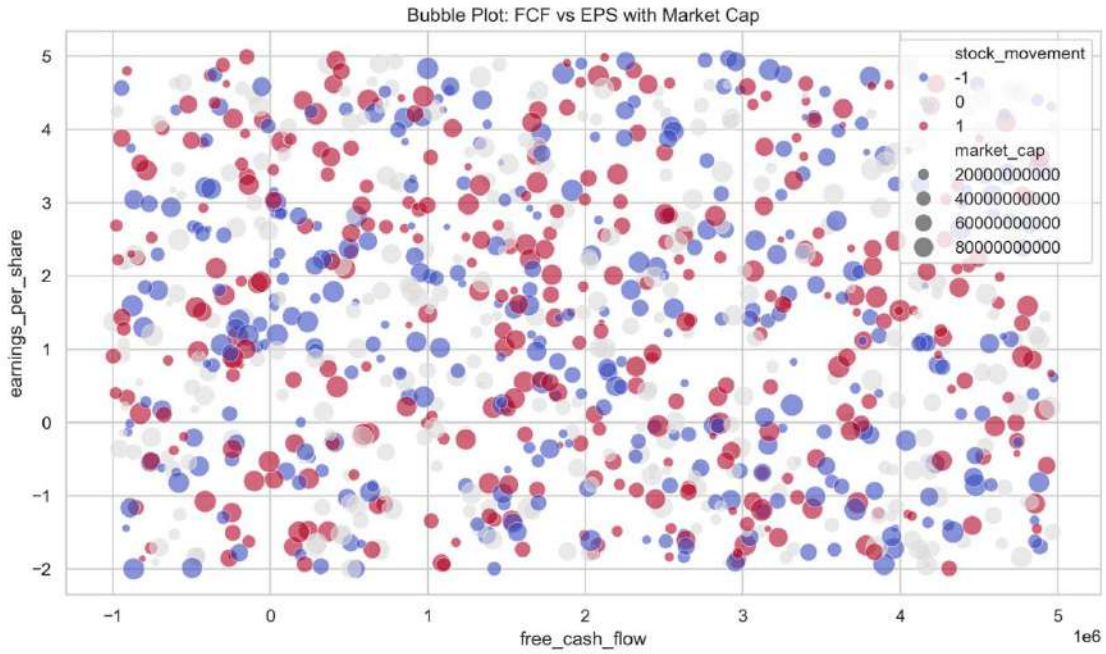

Figure 8.
FCF vs. EPS with the market cap of financial data.

The link between ROA and ROE, colored by stock movement (negative: red, neutral: green, and positive: pink), was depicted in Figure 9. The data points are evenly distributed, with no apparent grouping or strong association between ROA and ROE. Marginal KDE plots reveal generally symmetric distributions, with most values ranging from 0 to 0.2, indicating modest levels of profitability. Stock movement colors are uniformly distributed, indicating that there is no consistent relationship between stock price fluctuations and ROA or ROE variations. This suggests that profitability ratios alone could not be reliable predictors of short-term market reactions in financial statement analysis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

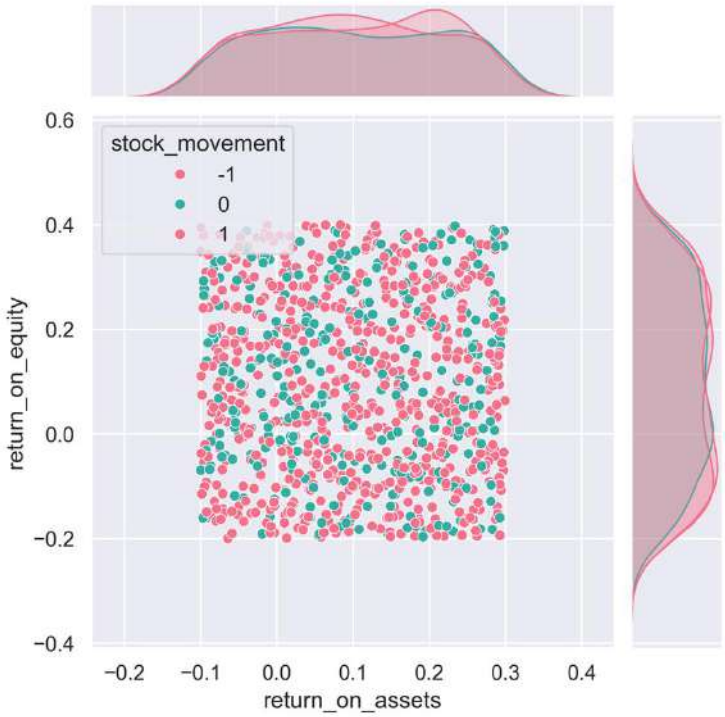

Figure 9.
ROE vs ROA performance of stock movement in financial data.

The simulated stock values for 1,000 firms before and after a financial statement, organized by analyst opinion, are shown in Figure 10. Each vertical line indicates a corporation, with blue indicating the stock price before the announcement and red representing the stock price after the announcement. The data appear to be highly erratic, with no apparent pattern throughout the sentiment spectrum, indicating that, while analyst sentiment could influence expectations, actual stock price movements could differ significantly. This emphasizes the intricate link between market sentiment and investor behavior, as well as the need to examine both qualitative mood and quantitative price data in financial statement analysis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

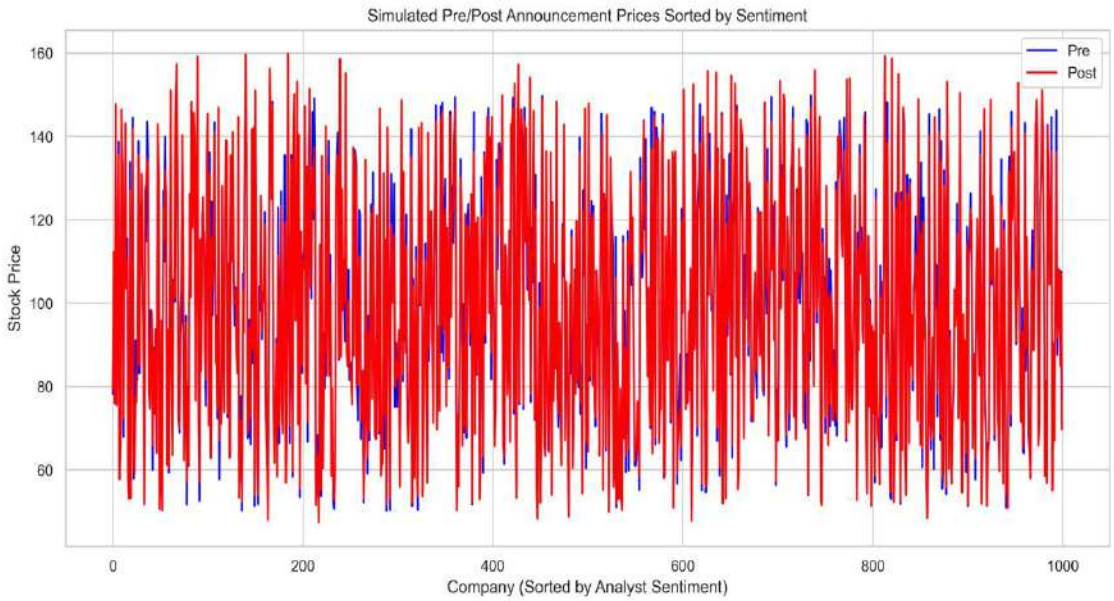

Figure 10.
Simulated pre/post-announcement prices sorted by sentiment.

4.2. Comparison With Existing Techniques

Financial statement analysis assesses a firm's financial well-being through the process of its income statement, balance sheet, and cash flow. It assists in measuring profitability, liquidity, solvency, and operating efficiency to make informed decisions. Fin Analytix may be restricted by certain high costs of implementation, reliance on precise input data, limited user customization for distinctive business models, and a high learning curve for fresh users. Also, it could not fully address qualitative factors or industry-distinctive distinctions, which could impact the depth and precision of financial analysis [24]. Decision trees tend to overfit financial data, particularly with compound statements. It may not be able to handle fine relationships and interdependencies. Instability arises from sensitivity to small changes in the data. Trees also lack transparency regarding the importance of features and could fail to generalize, thereby decreasing reliability for subtle financial statement analysis and forecasting [25]. The DRSVM-RF model improves financial statement analysis through the combination of strong classification and feature selection techniques. DRSVM corrects generalization with dynamic regularization and minimizes overfitting, while Random Forest enhances interpretability and operates efficiently with high-dimensional data. Together, they provide precise, consistent predictions, successfully capturing important financial indicators and trends in large and complex datasets.

4.3. Performance Metrics

Financial statement analysis is evaluated using performance metrics such as precision, accuracy, F1 score, recall, R^2, and mean square error (MSE).

Accuracy: The percentage of accurately anticipated financial outcomes relative to all projections is known as accuracy. Although it can be deceptive in unbalanced situations, such as the identification of uncommon fraud, it is helpful in balanced financial datasets (equation 28).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (28)$$

Precision: Precision in financial statement analysis is the proportion of firms that were accurately recognized as fraudulent or insolvent out of all those anticipated. It is critical for reducing false positives and ensuring that warned organizations present a financial danger (equation 29).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

$$Precision = \frac{TP}{TP+FP} \quad (29)$$

Recall: Recall measures the model's capacity to properly identify all relevant positive occurrences in financial data, all bankrupt enterprises, or those engaged in fraud. It is essential in situations when ignoring a risk is more costly than a false alert (equation 30).

$$Recall = \frac{TP}{TP+FN} \quad (30)$$

F1 score: F1-score strikes a balance between precision and recall, resulting in a single score that accounts for both false positives and false negatives. Financial statement analysis is extremely crucial when dealing with class imbalances, such as detecting unusual financial fraud (equation 31).

$$F1\ score = 2 \times \frac{precision \times recall}{precision+recall} \quad (31)$$

MSE: MSE calculates the average squared difference between actual and expected numerical financial values (such as sales and profit). A lower MSE recommends better predictive accuracy in financial forecasting and valuation models, as it penalizes significant errors more severely (equation 32).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (32)$$

R^2: The R^2 value reflects how well the regression model explains the variation in the financial variable (e.g., net income). A higher R^2 indicates that the model accurately reflects financial variability, which is crucial for valuation and forecasting purposes (equation 33).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (33)$$

Where TN= True positive, FP= False positive, TN= True negative, FN= False negative, y_i= actual financial value, $\hat{y}_i$= predicted value, $\bar{y}_i$= mean of the actual, n= number of observations

The FinAnalytix model achieves high effectiveness as demonstrated in Figure 11 and Table 2, with an accuracy of 93.48%, precision of 93.25%, recall of 92.85%, and an F1 score of 93.99%, indicating a strong ability to correctly identify relevant financial patterns while minimizing FP and FN. The traditional Profitability Ratios (PR) approach, trained with the BigData-FinStock dataset, utilizes key financial ratios to predict stock price movements following earnings announcements based on historical financial performance and market conditions. The PR model is less reliable, with an accuracy of 90.23%, a precision of 88.65%, a recall of 89.31%, and an F1 score of 87.43%, indicating that it frequently prefers precision above recall. Conclusively, these results point out that FinAnalytix provides unbiased and truthful forecasts for the analysis of financial statements, an important factor for an accurate financial analysis.

Table 2.
Values for the performance metrics using traditional and proposed methods.

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Fin Analytix [24]	93.48	93.25	92.85	93.99
PR	90.23	88.65	89.31	87.43
DRSVM-RF [Proposed]	95.62	94.82	93.77	96.1

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

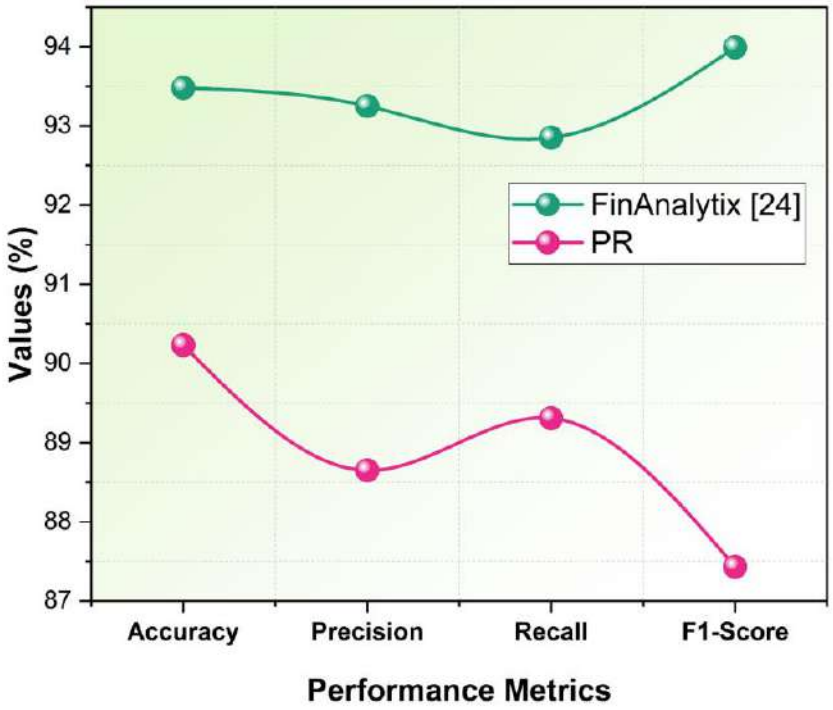

Figure 11.
Comparative analysis for traditional PR and FinAnalytix in financial statement analysis.

The proposed DRSVM-RF model outperforms FinAnalytix in financial statement analysis, as shown in Figure 12 and Table 2. It achieves an accuracy of 95.62%, which is higher than FinAnalytix's 93.48%, suggesting improved overall classification performance. Precision is also greater than 94.82% and 93.25%, implying that DRSVM-RF detects relevant positive instances more correctly. Furthermore, recall improves from 92.85% to 93.77%, indicating better identification of TP. The F1 score, which measures precision and recall, improves from 93.99% to 96.1%, indicating that the model is more effective and reliable. These findings illustrate DRSVM-RF's outstanding capacity to analyze financial information appropriately.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

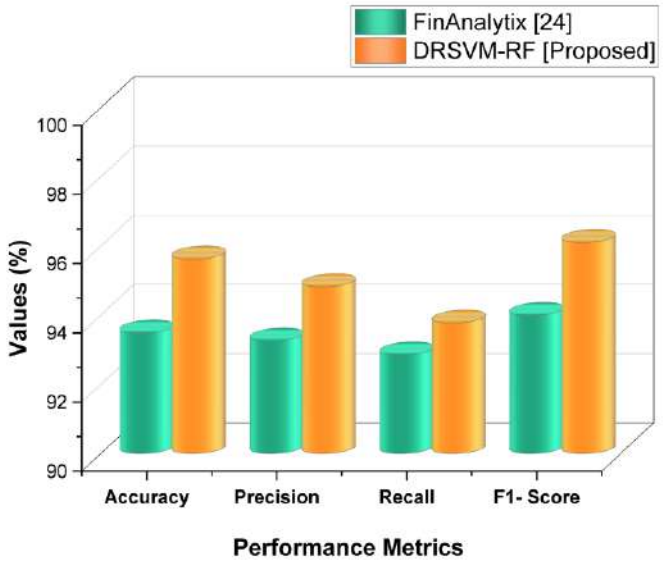

Figure 12.
Comparative performance for FinAnalytix and proposed in the financial statement.

Table 3.
Values for the performance in analyzing financial data using DT and DRSVM-RF.

Methods	MSE	R^2
DT [25]	1.293	0.977
DRSVM-RF [Proposed]	1.02	0.98

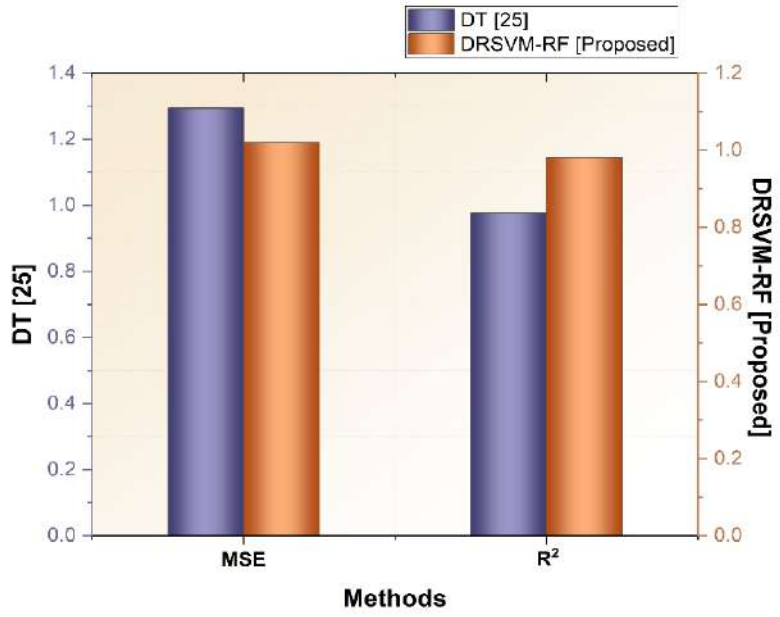

Figure 13.
Comparative performance for DT and the suggested method in financial statement analysis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

In financial statement analysis, reliable prediction models are critical for making accurate choices. Table 3 and Figure 13 compare the performance of two models: DT and the proposed method. The DRSVM-RF model outperforms the DT model, with a reduced MAE of 1.02 and 1.293 for DT, suggesting more accurate predictions. Furthermore, the greater R^2 of 0.98 for DRSVM-RF, compared to 0.977 for DT, indicates a more robust association between anticipated and actual financial values. This enhancement indicates the proposed model's increased capacity to capture complex financial patterns, making it more appropriate for accurate financial statement analysis.

5. Conclusion

In the modern big data era, traditional financial statement analysis methodologies are being challenged and complemented using ML approaches. It employs datasets of historical financial statements and market data. Big data refers to vast, intricate datasets that necessitate innovative tools for storage, processing, and analysis, as well as data normalization using min-max normalization and feature extraction using PCA, to identify key factors such as free cash flow and firm characteristics. It has been used to compare ML approaches to traditional financial statement analysis methodologies in anticipating stock market reactions to earnings releases. The conventional method centers on financial metrics, namely PR. In contrast, the ML technique employs complex algorithms like DRSVM-RF to handle massive volumes of financial and market data. The proposed techniques achieve better performance when compared to the baseline methods, precision (94.82%), recall (93.77%), accuracy (95.62%) F1 score (96.10%). The DRSVM-RF achieves better error performance with an MSE of 1.012, and R^2 *of* 0.98. It is constrained by the quality and availability of structured financial information, which could potentially influence the validity of both conventional and machine learning-driven analyses. The interpretability of certain sophisticated ML models is also an issue for financial decision-making. In the future, incorporating unstructured big data sources like social media and market sentiment, as well as explainable AI methods, could add strength and transparency to financial statement analysis to facilitate more timely and informed decision-making processes.

Transparency:

The author confirms that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] M. Akhtar, K. Yusheng, M. Haris, Q. U. Ain, and H. M. Javaid, "Impact of financial leverage on sustainable growth, market performance, and profitability," *Economic Change and Restructuring*, vol. 55, pp. 737–774, 2022. https://doi.org/10.1007/s10644-021-09321-z

[2] B. A. Zelalem and A. A. Abebe, "Balance sheet and income statement effect on dividend policy of private commercial banks in Ethiopia," *Cogent Economics & Finance*, vol. 10, no. 1, p. 2035917, 2022. https://doi.org/10.1080/23322039.2022.2035917

[3] T. Achmad, I. Ghozali, and I. D. Pamungkas, "Hexagon fraud: Detection of fraudulent financial reporting in state-owned enterprises Indonesia," *Economies*, vol. 10, no. 1, p. 13, 2022. https://doi.org/10.3390/economies10010013

[4] C.-C. Lee, "Analyses of the operating performance of information service companies based on indicators of financial statements," *Asia Pacific Management Review*, vol. 28, no. 4, pp. 410-419, 2023. https://doi.org/10.1016/j.apmrv.2023.01.002

[5] T. Achmad, I. Ghozali, M. R. A. Helmina, D. I. Hapsari, and I. D. Pamungkas, "Detecting fraudulent financial reporting using the fraud hexagon model: Evidence from the banking sector in Indonesia," *Economies*, vol. 11, no. 1, p. 5, 2023. https://doi.org/10.3390/economies11010005

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 7: 238-259, 2025
DOI: 10.55214/25768484.v9i7.8563

[6] M. P. Sari, E. Mahardika, D. Suryandari, and S. Raharja, "The audit committee as moderating the effect of hexagon's fraud on fraudulent financial statements in mining companies listed on the Indonesia stock exchange," *Cogent Business & Management*, vol. 9, no. 1, p. 2150118, 2022. https://doi.org/10.1080/23311975.2022.2150118

[7] G. D. Carnegie, P. Ferri, L. D. Parker, S. I. L. Sidaway, and E. E. Tsahuridu, "Accounting as technical, social and moral practice: The monetary valuation of public cultural, heritage and scientific collections in financial reports," *Australian Accounting Review*, vol. 32, no. 4, pp. 460-472, 2022. https://doi.org/10.1111/auar.12371

[8] F. S. Shiyyab, A. B. Alzoubi, Q. M. Obidat, and H. Alshurafat, "The impact of artificial intelligence disclosure on financial performance," *International Journal of Financial Studies*, vol. 11, no. 3, p. 115, 2023. https://doi.org/10.3390/ijfs11030115

[9] B. A. F. Jarah, M. A. AL Jarrah, M. A. A. Al-Zaqeba, and M. F. M. Al-Jarrah, "The role of internal audit to reduce the effects of creative accounting on the reliability of financial statements in the Jordanian Islamic banks," *International Journal of Financial Studies*, vol. 10, no. 3, p. 60, 2022. https://doi.org/10.3390/ijfs10030060

[10] J. Prather-Kinsey, F. De Luca, and H.-T.-P. Phan, "Improving the global comparability of IFRS-based financial reporting through global enforcement: A proposed organizational dynamic," *International Journal of Disclosure and Governance*, vol. 19, no. 3, pp. 330-351, 2022. https://doi.org/10.1057/s41310-022-00145-5

[11] A. Tsang, K. T. Wang, Y. Wu, and J. Lee, "Nonfinancial corporate social responsibility reporting and firm value: International evidence on the role of financial analysts," *European Accounting Review*, vol. 33, no. 2, pp. 399-434, 2024. https://doi.org/10.1080/09638180.2022.2094435

[12] H. A. Hamad and K. Cek, "The moderating effects of corporate social responsibility on corporate financial performance: Evidence from OECD countries," *Sustainability*, vol. 15, no. 11, p. 8901, 2023. https://doi.org/10.3390/su15118901

[13] X. Chen, Y. H. Cho, Y. Dou, and B. Lev, "Predicting future earnings changes using machine learning and detailed financial data," *Journal of Accounting Research*, vol. 60, no. 2, pp. 467-515, 2022. https://doi.org/10.1111/1475-679X.12429

[14] D. Ngoc Hung, V. T. Thuy Van, and L. Archer, "Factors affecting the quality of financial statements from an audit point of view: A machine learning approach," *Cogent Business & Management*, vol. 10, no. 1, p. 2184225, 2023.

[15] K. Wang, "Multifactor prediction model for stock market analysis based on deep learning techniques," *Scientific Reports*, vol. 15, no. 1, p. 5121, 2025. https://doi.org/10.1038/s41598-025-88734-6

[16] R. Liu, J. Liang, H. Chen, and Y. Hu, "Analyst reports and stock performance: Evidence from the Chinese market," *Asia-Pacific Financial Markets*, pp. 1-26, 2025. https://doi.org/10.1007/s10690-025-09522-w

[17] K. Mabelane, W. T. Mongwe, R. Mbuvha, and T. Marwala, "An analysis of local government financial statement audit outcomes in a developing economy using machine learning," *Sustainability*, vol. 15, no. 1, p. 12, 2023. https://doi.org/10.3390/su15010012

[18] T. H. H. Aldhyani and A. Alzahrani, "Framework for predicting and modeling stock market prices based on deep learning algorithms," *Electronics*, vol. 11, no. 19, p. 3149, 2022. https://doi.org/10.3390/electronics11193149

[19] A. Bakumenko and A. Elragal, "Detecting anomalies in financial data using machine learning algorithms," *Systems*, vol. 10, no. 5, p. 130, 2022. https://doi.org/10.3390/systems10050130

[20] X. Shi, Y. Zhang, M. Yu, and L. Zhang, "Deep learning for enhanced risk management: A novel approach to analyzing financial reports," *PeerJ Computer Science*, vol. 11, p. e2661, 2025. https://doi.org/10.7717/peerj-cs.2661

[21] S. K. Hashemi, S. L. Mirtaheri, and S. Greco, "Fraud detection in banking data by machine learning techniques," *IEEE Access*, vol. 11, pp. 3034-3043, 2023. https://doi.org/10.1109/ACCESS.2022.3232287

[22] Z. Zhao and T. Bai, "Financial fraud detection and prediction in listed companies using SMOTE and machine learning algorithms," *Entropy*, vol. 24, no. 8, p. 1157, 2022. https://doi.org/10.3390/e24081157

[23] Y. Song and R. Wu, "The impact of financial enterprises' excessive financialization risk assessment for risk control based on data mining and machine learning," *Computational Economics*, vol. 60, no. 4, pp. 1245-1267, 2022. https://doi.org/10.1007/s10614-021-10135-4

[24] Y. Fan, "Research and empirical evidence of machine learning based financial statement analysis methods," *Scalable Computing: Practice and Experience*, vol. 25, no. 6, pp. 4693–4701, 2024. https://doi.org/10.12694/scpe.v25i6.3284

[25] P. Chakri, S. Pratap, Lakshay, and S. K. Gouda, "An exploratory data analysis approach for analyzing financial accounting data using machine learning," *Decision Analytics Journal*, vol. 7, p. 100212, 2023. https://doi.org/10.1016/j.dajour.2023.100212