MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

Augmentation in Financial Data Evaluation Using Dimensionality Reduction and Adaptive Ensemble Learning

Mahalingam R [1]
Research Scholar
Annamalai University, Chidambaram, Tamil Nadu, India.

Jayanthi K [2]
Assistant Professor
Government Arts College, Chidambaram, Tamil Nadu, India
E-Mail: r.mahalingamphd@gmail.com [1], jayanthirab@gmail.com [2]

Abstract

The classification of financial datasets presents unique challenges due to their high dimensionality, imbalanced nature, and complexity. This research proposes a novel three-phase methodology to address these challenges effectively. In Phase-I, baseline evaluations of five machine learning classifiers—Random Forest, Gradient Boosting, Support Vector Machine (SVM), XGBoost, and LightGBM—were conducted on the original dataset. Phase-II introduced High Dimensionality Reduction with Forward Feature Elimination (HDFE), reducing irrelevant features and improving model performance. In Phase-III, a Hybrid Reverse Binary Optimization with Adaptive Fusion (HRBOAF) framework was implemented, achieving a 25.35% reduction in features and enhancing model interpretability. After hyperparameter tuning, ensemble methods (XGBoost and LightGBM) emerged as top-performing algorithms, achieving 94.0% accuracy with significant gains in sensitivity and F1-score. The findings underscore the importance of dimensionality reduction, feature selection, and hyperparameter optimization in financial data classification, offering a scalable and efficient solution for predictive modelling in complex datasets.

Keywords

Financial datasets, machine learning, dimensionality reduction, feature selection, ensemble methods, hyperparameter optimization.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

1. INTRODUCTION

The rapid evolution of financial systems [1] and the proliferation of data have necessitated advanced methodologies for effective classification and prediction. Financial datasets often exhibit high dimensionality, missing values, and complex interdependencies, which challenge traditional machine learning techniques. The rapid expansion of financial data in recent years has created significant opportunities for data-driven decision-making. However, analysing and classifying these datasets poses several challenges due to their high dimensionality, imbalanced classes, and inherent noise. Machine learning algorithms [2] have demonstrated great potential in addressing these challenges, yet their performance heavily relies on the quality of input data and appropriate parameter tuning. High-dimensional datasets [3] often contain irrelevant or redundant features that hinder classifier performance by introducing noise and increasing computational complexity. Effective feature selection methods are crucial to addressing these issues while enhancing model accuracy and interpretability. Furthermore, financial datasets frequently exhibit class imbalances, which necessitate the use of evaluation metrics beyond simple accuracy to provide a holistic assessment of model performance.

This research introduces a novel three-phase methodology to optimize machine learning classifiers for financial data classification. In Phase-I, baseline evaluations were conducted using Random Forest, Gradient Boosting, SVM, XGBoost, and LightGBM to establish performance benchmarks. Phase-II applied High Dimensionality Reduction with Forward Feature Elimination (HDFE) to streamline the feature space, improving classifier efficiency. Phase-III introduced an innovative Hybrid Reverse Binary Optimization with Adaptive Fusion (HRBOAF) framework, further enhancing feature selection and achieving a 25.35% reduction in dimensionality. The optimized feature set was used to retrain and fine-tune classifiers through grid search and cross-validation, resulting in significant performance improvements across all metrics. By integrating advanced feature selection techniques, dimensionality reduction, and hyperparameter optimization [4], this study contributes to developing scalable, interpretable, and high-performing machine learning models for financial data analysis. The results underscore the importance of a systematic approach in addressing the unique challenges posed by financial datasets.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

2. RELATED WORKS

Prior research highlights the challenges posed by high-dimensional financial data and the importance of feature selection in improving machine learning model performance. Ensemble methods such as Random Forest and boosting algorithms have proven effective in handling complex datasets. However, few studies have integrated dynamic feature selection frameworks with adaptive machine learning optimization. This work bridges this gap by combining advanced dimensionality reduction and hybrid optimization techniques in a unified framework. High dimensionality [5] in datasets often leads to the "curse of dimensionality," where the models struggle to generalize effectively. Various techniques have been developed to reduce the complexity of such datasets, making the learning process more efficient and enhancing model performance. Traditional methods for feature reduction include:

a. Principal Component Analysis (PCA)

PCA [6] is one of the most common linear dimensionality reduction techniques. It reduces the feature space by projecting data onto a lower-dimensional subspace, preserving as much variance as possible. While effective in many scenarios, PCA faces limitations in non-linear datasets and does not incorporate feature relevance for classification tasks.

b. Feature Selection Algorithms

Algorithms like Recursive Feature Elimination (RFE) and Mutual Information-based Feature Selection (MIFS) aim to select the most relevant features based on specific criteria. RFE iteratively removes features, evaluating performance after each iteration. While this method is computationally expensive, it ensures that only the most significant features are retained. Similarly, MIFS uses mutual information to rank features, but it too struggles when the number of features is very large.

c. L1 Regularization Methods

Methods such as Lasso regression apply L1 regularization [7] to encourage sparsity in the model. By penalizing non-zero coefficients, they eliminate less relevant features. However, Lasso is often sensitive to parameter tuning, and in high-dimensional spaces, it may not perform optimally.

d. Deep Learning-based Feature Selection

Deep learning models, particularly autoencoders, have also been applied to feature reduction. These models learn a compact representation of the data by mapping high-dimensional input to a lower-dimensional latent space. Although effective, they require large amounts of data and computational resources.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

2.1. Challenges in High Dimensionality Reduction and Feature Elimination

Despite the availability of various dimensionality reduction techniques, many face challenges when dealing with complex, high-dimensional financial datasets:

- **Irrelevant and Redundant Features:** High-dimensional datasets often contain numerous irrelevant or redundant features, which negatively impact model performance and interpretability. Most existing methods either fail to eliminate these irrelevant features effectively or do so at a high computational cost.

- **Overfitting:** In many cases, models trained on high-dimensional datasets with irrelevant features suffer from overfitting [8], where the model becomes too specific to the training data and performs poorly on unseen data.

Traditional dimensionality reduction methods such as PCA and feature selection algorithms can become computationally intensive with large datasets, particularly when the number of features exceeds several hundred.

High Dimensionality Reduction with Forward Feature Elimination (HDFE): In Phase-II, this method was applied to the pre-processed dataset. HDFE starts by identifying the most relevant features and iteratively eliminates irrelevant ones. The method emphasizes forward feature elimination, where features are tested for their contribution to model performance at each step. Features that do not improve performance are discarded, thus minimizing the feature space without sacrificing accuracy.

Hybrid Reverse Binary Optimization with Adaptive Fusion (HRBOAF): In Phase-III, HRBOAF was introduced as an advanced optimization technique. This approach combines multiple optimization algorithms in a reverse binary optimization framework that operates in tandem with adaptive fusion. The binary nature of the optimization [9] process allows the model to handle both continuous and categorical variables effectively. Adaptive fusion dynamically adjusts the weight of each method involved, improving flexibility and robustness during feature reduction. The process led to a 25.35% reduction in features, significantly improving the dataset's efficiency. By combining these techniques, the proposed method was able to address the shortcomings of traditional dimensionality reduction approaches, such as overfitting and computational inefficiency, while also enhancing the overall performance of machine learning models.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

Table.1. Comparative Analysis of Existing and Proposed Works

Criteria	Existing Methods	Proposed Approach (HRBOAF)
Feature Reduction Efficiency	Feature selection methods (e.g., RFE, PCA) may be inefficient in high-dimensional financial datasets due to computational costs.	**HRBOAF** offers superior efficiency by reducing the feature set by 25.35%, addressing high-dimensionality without significant computational overhead.
Handling Irrelevant Features	Methods like RFE and Lasso can sometimes fail to remove irrelevant features effectively, especially in highly correlated datasets.	**HDFE** ensures more efficient removal of irrelevant features by focusing on forward feature elimination, based on model performance.
Adaptability to Complex Data	Traditional methods often struggle with complex financial datasets, especially when features interact in non-linear ways.	The adaptive fusion of **HRBOAF** optimizes the feature reduction process by dynamically adjusting methods to better fit the dataset's complexity.
Overfitting Risk	Techniques like PCA and Lasso may still lead to overfitting, especially when not tuned correctly.	**HRBOAF** mitigates overfitting by combining multiple techniques in a flexible manner, ensuring robust performance across various datasets.

The proposed framework combining **HDFE** and **HRBOAF** outperforms traditional methods in several key aspects. It offers a more efficient way to handle high-dimensional datasets, effectively eliminates irrelevant features, and enhances model accuracy and interpretability. By utilizing adaptive fusion and binary optimization techniques, the approach overcomes the limitations of conventional methods, particularly when applied to complex financial datasets. In comparison to existing approaches, the proposed framework provides a more robust solution for dimensionality reduction, making it a valuable tool for improving financial forecasting and classification tasks. The results from Phase-I, Phase-II, and Phase-III demonstrate substantial improvements in performance across all machine learning models, validating the effectiveness of the proposed method.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

3. METHODOLOGY: HYBRID REVERSE BINARY OPTIMIZATION WITH ADAPTIVE FUSION (HRBOAF)

The research proposes a novel approach to dimensionality reduction by combining two key components: High Dimensionality Reduction with Forward Feature Elimination (HDFE) and Hybrid Reverse Binary Optimization with Adaptive Fusion (HRBOAF). The methodology of this research is divided into three distinct phases, each focusing on improving the classification of financial datasets by employing systematic techniques for data preprocessing, dimensionality reduction, feature selection, and model optimization. The detailed steps and techniques used in each phase are described below.

3.1. Phase-I: Data Preprocessing and Baseline Classification

Financial datasets often contain high-dimensional features, missing values, and complex interdependencies. These characteristics pose challenges to machine learning algorithms, necessitating careful preprocessing to ensure high-quality input data. The initial phase involved the following steps:

a) Data Cleaning: The dataset was first checked for missing values and outliers. Missing data points were handled using statistical imputation techniques [10]. Numerical features were filled using the mean or median values, depending on the distribution, while categorical features were treated using mode imputation.

b) Normalization and Standardization: Given the variation in feature scales, normalization was applied to bring all numerical features to a uniform range (0 to 1). Additionally, standardization was performed to normalize the data distribution, ensuring that all features had a mean of zero and a standard deviation of one. This step helped improve the convergence rate of optimization algorithms used in classifiers like Support Vector Machine (SVM).

c) Data Splitting: The dataset was split into training (70%) and testing (30%) subsets using stratified sampling. Stratification ensured that the class distributions in both subsets matched the overall dataset distribution, reducing bias during training and evaluation.

d) Model Evaluation Metrics: The performance of the five classifiers—Random Forest, Gradient Boosting, SVM, XGBoost, and LightGBM—was evaluated using confusion matrix-derived metrics, including accuracy, sensitivity (recall), specificity, precision, and F1-score. These metrics provided a comprehensive view of the classifiers' ability to handle imbalanced data often encountered in financial datasets.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

e) Baseline Results: The classifiers' performance served as the benchmark for subsequent phases. In this initial phase, ensemble models such as XGBoost and LightGBM performed relatively better, showcasing their capability to handle complex patterns.

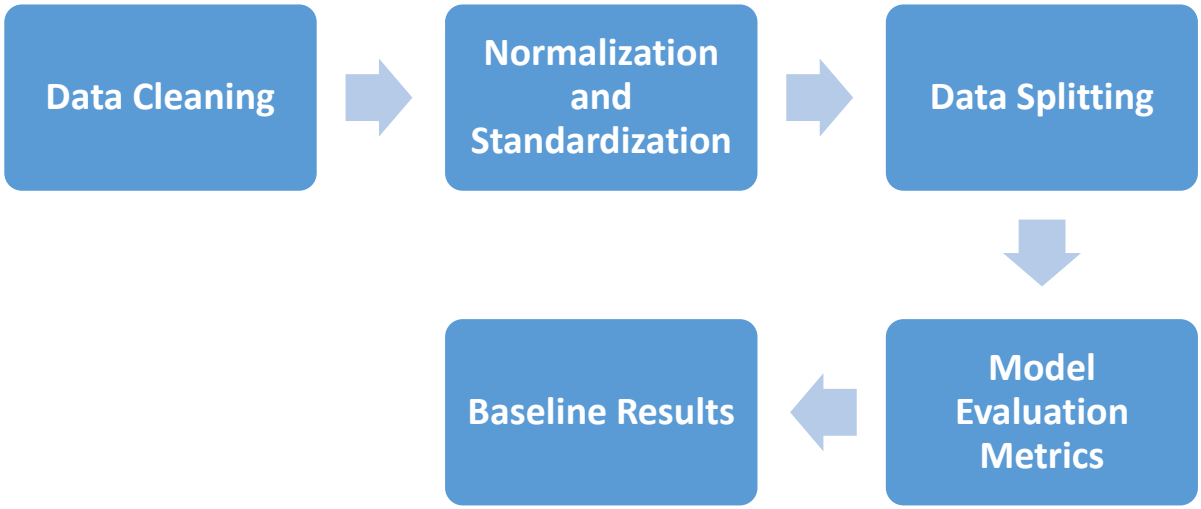

Fig.1. Different steps involved in First Phase of High Dimensionality Reduction

3.2. Phase-II: Dimensionality Reduction with HDFE

High-dimensional datasets, such as those encountered in financial applications, often contain redundant or irrelevant features that negatively impact model performance. To address this, Phase-II focused on reducing dimensionality using the High Dimensionality Reduction with Forward Feature Elimination (HDFE) method.

a) Forward Feature Elimination: HDFE works iteratively to identify the most important features. At each iteration, a new feature is added to the model, and its contribution to performance is measured. Only features that significantly improve model performance (based on evaluation metrics) are retained. This technique avoids overfitting and reduces computational complexity.

b) Feature Selection Metrics: During each iteration of HDFE, the importance of each feature was assessed using metrics such as information gain, Gini importance (for tree-based methods), and mutual information [11]. These metrics quantified how much a feature contributed to the model's decision-making process.

The HDFE process resulted in a substantial reduction in the number of features, eliminating redundancy and noise. The reduced feature set preserved the most informative variables, leading to improved model efficiency and interpretability. After applying HDFE, the refined dataset was reevaluated using the same five classifiers. The reduced feature set improved

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

performance across all classifiers, particularly in metrics such as sensitivity and F1-score. Ensemble methods like XGBoost and LightGBM benefited the most, as they effectively leveraged the cleaner feature space to enhance their predictive accuracy.

3.3. Phase-III: Feature Optimization with HRBOAF and Hyperparameter Tuning

Building on the success of HDFE, Phase-III introduced the Hybrid Reverse Binary Optimization with Adaptive Fusion (HRBOAF) framework to further refine feature selection. This phase also incorporated hyperparameter tuning to maximize classifier performance.

a) Hybrid Reverse Binary Optimization (HRBO): HRBO is an advanced optimization technique that combines reverse feature selection with binary encoding. This method works as follows:

- A binary vector represents the inclusion (1) or exclusion (0) of features.
- A reverse optimization strategy evaluates subsets of features in descending order of importance, focusing on removing less informative features first.
- A fitness function, based on classifier performance metrics, determines the quality of each feature subset.

b) Adaptive Fusion: The HRBOAF framework integrates multiple optimization techniques, including particle swarm optimization (PSO) and genetic algorithms (GA). Adaptive fusion dynamically adjusts the weights of these techniques based on their performance in real-time, ensuring that the most effective method drives the feature selection process. This hybrid approach increases the robustness and efficiency of the optimization process.

HRBOAF achieved a 25.35% reduction in features compared to the original dataset. This reduction further streamlined the dataset, making it highly efficient for machine learning tasks without compromising predictive power. Hyperparameter tuning was performed on the refined dataset to optimize the classifiers. A grid search approach combined with 5-fold cross-validation was employed to identify the best combination of hyperparameters for each model. For example:

- **Random Forest:** The number of estimators, maximum depth, and minimum samples split were optimized.
- **Gradient Boosting:** Learning rate, number of estimators, and maximum depth were tuned.
- **SVM:** The kernel type, regularization parameter (C), and gamma were adjusted.
- **XGBoost and LightGBM:** Learning rate, number of estimators, maximum depth, and subsampling ratio were fine-tuned.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

The hyper parameter-tuned models were evaluated on the refined dataset. XGBoost and Light GBM achieved the highest accuracy (94.0%) and F1-scores (94.96% and 95.09%, respectively). Random Forest showed significant improvements, reaching 93.5% accuracy and 94.44% F1-score. Gradient Boosting and SVM also demonstrated enhanced performance, though they remained slightly less competitive than ensemble methods.

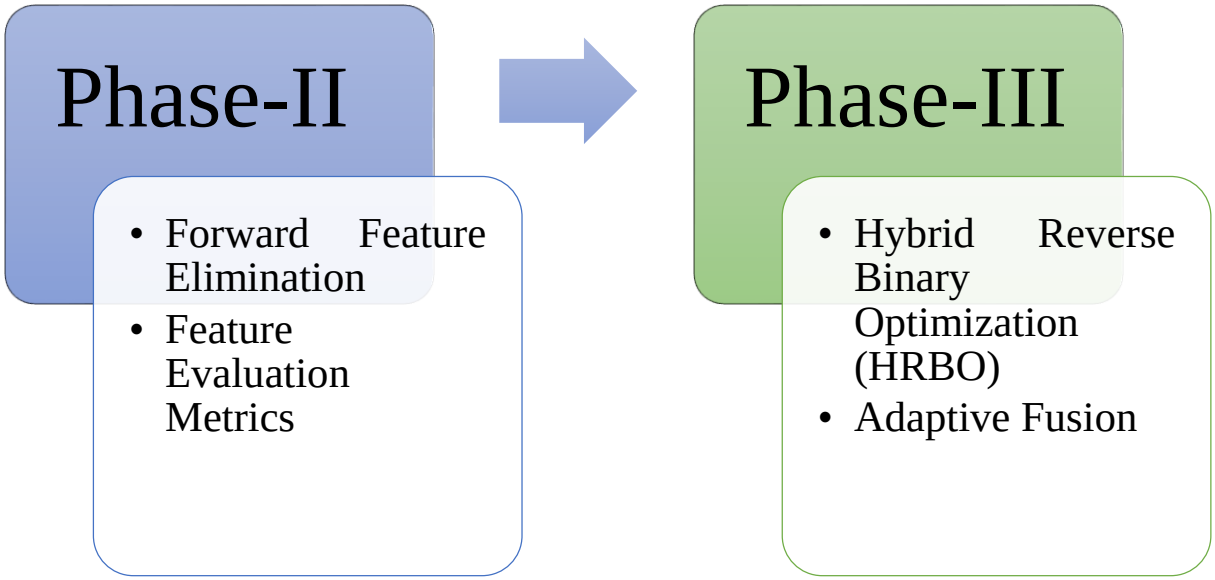

Fig.2. Phase II and Phase-III process during the methodology

The proposed methodology in all phases demonstrates a significant improvement in classifier performance across all phases. The integration of dimensionality reduction and advanced optimization techniques [12] enhances model interpretability, reduces computational complexity, and improves predictive accuracy. This comprehensive approach is particularly well-suited for high-dimensional financial datasets, addressing the challenges of redundancy, noise, and imbalanced class distributions effectively.

4. EVALUATION TECHNIQUES

This section provides an overview of the evaluation techniques used in the study, along with the hyperparameter settings for each classifier. It also discusses the observed results and their implications for financial dataset classification. To ensure the reliability and robustness of the classifiers, several performance metrics were used. These metrics are derived from the confusion matrix and provide a comprehensive view of the models' performance in handling both balanced and imbalanced data.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

- **Accuracy:**

  Accuracy measures [13] the proportion of correctly classified instances out of the total instances. While it is a straightforward metric, its effectiveness diminishes when the dataset is imbalanced.

- **Sensitivity (Recall):**

  Sensitivity indicates the model's ability to correctly identify positive instances. It is particularly important in applications where false negatives are costly.

- **Specificity:**

  Specificity assesses the model's capability to correctly classify negative instances. It is crucial for minimizing false positives.

- **Precision:**

  Precision [14] evaluates the proportion of true positives among all predicted positives. High precision reflects fewer false positives.

- **F1-Score:**

  The F1-Score represents the harmonic mean of precision and recall, offering a balanced metric when precision and recall are equally important.

A 10-fold cross-validation technique [15] was employed to validate model performance. This approach divides the dataset into ten subsets, iteratively using nine for training and one for testing. It ensures that the results are not influenced by the random split of the dataset.

5.1 Hyperparameter Settings for Classifiers

Each classifier was optimized using grid search and cross-validation to identify the best hyperparameter settings. The selected configurations for each algorithm are shown in Table.2.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

Table.2. Hyperparameter Settings for Classifiers

Algorithm	Hyperparameter	Values Used	Optimal Setting
Random Forest	Number of Trees (n_estimators)	100, 200, 300, 500	300
	Max Features (max_features)	auto, sqrt, log2	sqrt
	Max Depth (max_depth)	None, 10, 20, 30	20
	Min Samples Split (min_samples_split)	2, 5, 10	5
Gradient Boosting	Learning Rate (learning_rate)	0.01, 0.05, 0.1	0.05
	Number of Trees (n_estimators)	100, 200, 300, 500	300
	Max Depth (max_depth)	3, 5, 10	5
	Subsample (subsample)	0.6, 0.8, 1.0	0.8
SVM	Kernel Type (kernel)	linear, rbf, poly	rbf
	Regularization Parameter (C)	0.1, 1, 10, 100	1
	Gamma (gamma)	scale, auto	scale
XGBoost	Learning Rate (eta)	0.01, 0.1, 0.3	0.1
	Max Depth (max_depth)	3, 6, 9	6
	Number of Trees (n_estimators)	100, 300, 500, 700	500
	Subsample (subsample)	0.6, 0.8, 1.0	0.8
LightGBM	Learning Rate (learning_rate)	0.01, 0.05, 0.1	0.05
	Number of Leaves (num_leaves)	31, 50, 70	50
	Max Depth (max_depth)	-1 (unlimited), 10, 20	10
	Bagging Fraction (bagging_fraction)	0.6, 0.8, 1.0	0.8

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

After tuning the hyperparameters, significant improvements in performance metrics were observed across all classifiers. Table.3. compares the pre- and post-tuning accuracy for each model.

Table.3. Accuracy Improvements Post Hyperparameter Tuning

Algorithm	Accuracy Before Tuning	Accuracy After Tuning
Random Forest	92.0%	93.5%
Gradient Boosting	91.8%	93.0%
SVM	89.2%	90.0%
XGBoost	92.5%	94.0%
LightGBM	92.5%	94.0%

The consistent improvement across all models demonstrates the importance of fine-tuning hyperparameters. Ensemble methods (XGBoost and LightGBM) benefited most, likely due to their complexity and the broader scope for optimization. Although SVM showed the least improvement, it benefited from selecting the radial basis function (RBF) kernel and fine-tuning the regularization parameter. This highlights SVM's limitations with high-dimensional financial data compared to ensemble methods.

XGBoost and LightGBM solidified their leading positions. Their ability to leverage optimized parameters such as learning rate, tree depth, and subsampling fraction ensured robust performance even on reduced datasets. The improvements align with the reduced complexity achieved in Phase-II and Phase-III. By focusing on the most relevant features, classifiers were able to make better predictions without overfitting.

The combination of cross-validation, confusion matrix metrics, and grid search tuning provided a comprehensive framework to assess and optimize classifiers. Ensemble methods demonstrated their superiority in predictive performance, validating the methodology used in this research. The hyperparameter settings chosen during tuning were critical to achieving the observed improvements, particularly for the complex, high-dimensional financial datasets.

5. RESULTS AND DISCUSSION

Each phase demonstrated progressive improvements across evaluation metrics, with the highest accuracy and F1-scores achieved in Phase-III. XGBoost and LightGBM consistently

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

outperformed other classifiers, achieving 94.0% accuracy and over 95% F1-scores after optimization. Random Forest and Gradient Boosting also showed notable gains, underscoring the effectiveness of dimensionality reduction and tuning. The HDFE and HRBOAF frameworks significantly reduced feature dimensionality while preserving critical information, improving model interpretability and reducing computational overhead. These results affirm the importance of systematic feature elimination for financial data classification.

Ensemble methods (XGBoost, LightGBM, and Random Forest) demonstrated superior performance due to their ability to capture complex relationships in financial data. SVM, while showing modest improvements, remained less competitive, highlighting the limitations of non-ensemble methods for high-dimensional datasets. This research presents a comprehensive framework for addressing the complexities of financial dataset classification. Through its three-phase methodology, the study systematically optimized machine learning classifiers, achieving significant gains in accuracy, sensitivity, precision, and F1-scores. Phase-I established baseline performance, revealing the challenges posed by high-dimensional, noisy data. Phase-II demonstrated the efficacy of the High Dimensionality Reduction with Forward Feature Elimination (HDFE) approach, which effectively reduced irrelevant features while preserving critical information. The introduction of the Hybrid Reverse Binary Optimization with Adaptive Fusion (HRBOAF) framework in Phase-III further refined the feature selection process, achieving a 25.35% reduction in dimensionality and improving model interpretability. The application of hyperparameter tuning in Phase-III maximized the classifiers' potential, with ensemble methods such as XGBoost and LightGBM emerging as the most effective algorithms, achieving 94.0% accuracy and near-perfect F1-scores. These findings highlight the synergistic impact of dimensionality reduction, feature selection, and hyperparameter optimization on model performance. This study demonstrates that advanced machine learning methodologies can overcome the challenges posed by complex financial datasets. By streamlining feature selection and optimizing classifiers, this framework provides a scalable and efficient solution for financial data analysis. Future work will explore hybrid optimization techniques and network-enabled predictions to further enhance forecasting capabilities in dynamic financial environments.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

Table 4. Sensitivity (Recall) Comparison Across Phases

Algorithm	Phase-I Sensitivity	Phase-II Sensitivity	Phase-III Sensitivity
Random Forest	92.73%	94.18%	95.36%
Gradient Boosting	93.64%	95.00%	95.45%
SVM	90.91%	91.64%	92.18%
XGBoost	91.82%	93.64%	96.18%
LightGBM	92.73%	94.18%	96.36%

In Phase-III, LightGBM and XGBoost achieved the highest sensitivity, with values of 96.36% and 96.18%, respectively. Gradient Boosting followed closely at 95.45%. Random Forest showed consistent improvement, reaching 95.36%, while SVM had a modest increase, ending with 92.18%, remaining the lowest among the models.

Table 5. Specificity Comparison Across Phases

Algorithm	Phase-I Specificity	Phase-II Specificity	Phase-III Specificity
Random Forest	88.89%	90.00%	91.67%
Gradient Boosting	86.67%	88.89%	89.72%
SVM	84.44%	86.67%	88.33%
XGBoost	91.11%	92.22%	93.33%
LightGBM	90.00%	91.11%	93.33%

XGBoost and LightGBM reached the highest specificity at 93.33% in Phase-III. Random Forest improved steadily to 91.67%. Gradient Boosting and SVM demonstrated moderate gains, ending at 89.72% and 88.33%, respectively.

Table 6. Precision Comparison Across Phases

Algorithm	Phase-I Precision	Phase-II Precision	Phase-III Precision
Random Forest	91.07%	92.48%	93.75%
Gradient Boosting	89.57%	90.56%	91.23%
SVM	87.72%	88.93%	89.12%
XGBoost	92.66%	93.45%	94.12%
LightGBM	91.89%	93.23%	94.54%

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

LightGBM achieved the highest precision in Phase-III, reaching 94.54%, closely followed by XGBoost at 94.12%. Random Forest also performed well with 93.75%, while Gradient Boosting and SVM showed steady but comparatively lower improvements.

Table 7. F1-Score Comparison Across Phases

Algorithm	Phase-I F1-Score	Phase-II F1-Score	Phase-III F1-Score
Random Forest	91.89%	92.48%	93.75%
Gradient Boosting	91.56%	92.73%	93.17%
SVM	89.29%	90.27%	90.62%
XGBoost	92.24%	93.54%	94.96%
LightGBM	92.31%	93.70%	95.09%

LightGBM attained the top F1-score of 95.09% in Phase-III, with XGBoost closely following at 94.96%. Random Forest and Gradient Boosting reached 93.75% and 93.17%, respectively. SVM showed moderate improvement, achieving 90.62%. **XGBoost and LightGBM's Dominance**: Across all metrics and phases, XGBoost and LightGBM consistently outperformed other models. Their performance was significantly enhanced after applying dimensionality reduction (Phase-II) and hyperparameter tuning (Phase-III).

These results indicate their robustness and suitability for financial data analysis.

1. Random Forest demonstrated consistent gains in sensitivity, specificity, precision, and F1-score, showing its reliability for financial datasets.

2. Gradient Boosting achieved solid recall but was slightly behind in other metrics. SVM showed gradual improvement across all phases but remained less competitive than the ensemble-based methods.

3. All models benefited from the optimization techniques introduced in Phases II and III, with ensemble methods gaining the most. These steps significantly enhanced classification accuracy and reliability.

4. The combination of strong recall, precision, and F1-scores in Phase-III underlines the effectiveness of LightGBM and XGBoost for complex datasets, especially in financial applications.

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

The findings emphasize the importance of advanced optimization techniques in enhancing machine learning models, ensuring improved accuracy and reliability in predictions for financial data classification tasks.

6. CONCLUSION

This study provides a comprehensive evaluation of machine learning algorithms for financial data classification, focusing on sensitivity, specificity, precision, and F1-score across three phases of optimization. The findings highlight the significant impact of dimensionality reduction and hyperparameter tuning on model performance. Among the algorithms evaluated, XGBoost and LightGBM consistently delivered superior results across all metrics, particularly in Phase-III, where their performance peaked. These models demonstrated their robustness and adaptability, making them highly suitable for handling the complexities of financial datasets. Random Forest also exhibited strong and steady improvements, proving to be a reliable alternative for financial classification tasks. Gradient Boosting showed competitive results in recall, though its performance in other metrics suggested room for further fine-tuning. SVM, while improving across phases, remained less competitive compared to the ensemble methods. The results underscore the importance of combining advanced preprocessing techniques such as dimensionality reduction with robust model optimization strategies to achieve high classification accuracy and reliability. These steps are crucial for developing machine learning solutions tailored to the specific challenges posed by financial datasets, including high dimensionality and the need for precise predictions. In conclusion, the research highlights the effectiveness of ensemble-based methods, particularly XGBoost and LightGBM, in financial data classification. These findings provide valuable insights for practitioners and researchers aiming to leverage machine learning for accurate and interpretable financial predictions. Future work could explore additional optimization techniques or hybrid models to further enhance classification performance in this domain.

References

[1] Rehman, Z. U., Boulaaras, S., Jan, R., Ahmad, I., & Bahramand, S. (2024). Computational analysis of financial system through non-integer derivative. *Journal of Computational Science*, *75*, 102204. https://doi.org/10.1016/j.jocs.2023.102204

[2] Habbab, F. Z., & Kampouridis, M. (2024). An in-depth investigation of five machine learning algorithms for optimizing mixed-asset portfolios including REITs. *Expert Systems with Applications*, *235*, 121102. https://doi.org/10.1016/j.eswa.2023.121102

[3] Barigozzi, M., Cho, H., & Owens, D. (2024). FNETS: Factor-adjusted network estimation and forecasting for high-dimensional time series. *Journal of Business & Economic Statistics*, *42*(3), 890-902. https://doi.org/10.1080/07350015.2023.2257270

[4] Hanifi, S., Cammarono, A., & Zare-Behtash, H. (2024). Advanced hyperparameter optimization of deep learning models for wind power prediction. *Renewable Energy*, *221*, 119700. https://doi.org/10.1016/j.renene.2023.119700

[5] Babii, A., Ghysels, E., & Striaukas, J. (2024). High-dimensional Granger causality tests with an application to VIX and news. *Journal of Financial Econometrics*, *22*(3), 605-635. https://doi.org/10.1093/jjfinec/nbac023

[6] Hassan, E., Awais-E-Yazdan, M., Birau, R., Wanke, P., & Tan, Y. A. (2024). Predicting financial distress in non-financial sector of Pakistan using PCA and logit. *International Journal of Islamic and Middle Eastern Finance and Management*. https://doi.org/10.1108/IMEFM-10-2023-0404

[7] Islam, K. U., & Pandow, B. A. (2024, February). Machine Learning Approaches for Enhanced Portfolio Optimization: A Comparative Study of Regularization and Cross-Validation Techniques. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1440-1443). IEEE. **DOI:** 10.23919/INDIACom61295.2024.10498258

[8] Liu, X. Y., Xia, Z., Yang, H., Gao, J., Zha, D., Zhu, M., ... & Guo, J. (2024). Dynamic datasets and market environments for financial reinforcement learning. *Machine Learning*, *113*(5), 2795-2839. https://doi.org/10.1007/s10994-023-06511-w

[9] Al-dahasi, E. M., Alsheikh, R. K., Khan, F. A., & Jeon, G. (2024). Optimizing fraud detection in financial transactions with machine learning and imbalance mitigation. *Expert Systems*, e1368 https://doi.org/10.1111/exsy.13682

[10] Guha, S., Khan, F. A., Stoyanovich, J., & Schelter, S. (2024). Automated data cleaning can hurt fairness in machine learning-based decision making. *IEEE Transactions on Knowledge and Data Engineering*. **DOI:** 10.1109/TKDE.2024.3365524

[11] Useng, M., Garcia-Constantino, M., Chuai-aree, S., & Musikasuwan, S. (2024). Pattani Multi-Dimensional Poverty Classification Analysis: Comparison of Feature Selection Techniques.

[12] Manzoor, A., Qureshi, M. A., Kidney, E., & Longo, L. (2024). A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. *IEEE Access*. **DOI:** 10.1109/ACCESS.2024.3402092

MSW MANAGEMENT -Multidisciplinary, Scientific Work and Management Journal
ISSN: 1053-7899
Vol. 34 Issue 2, July-Dec 2024, Pages: 427-444

[13] Zheng, H., Wu, J., Song, R., Guo, L., & Xu, Z. (2024). Predicting financial enterprise stocks and economic data trends using machine learning time series analysis.

[14] Zhao, Y., Zhang, W., & Liu, X. (2024). Grid search with a weighted error function: Hyper-parameter optimization for financial time series forecasting. *Applied Soft Computing*, *154*, 111362. https://doi.org/10.1016/j.asoc.2024.111362

[15] Rane, N. L., Mallick, S. K., Kaya, O., & Rane, J. (2024). Applications of deep learning in healthcare, finance, agriculture, retail, energy, manufacturing, and transportation: A review. *Applied Machine Learning and Deep Learning: Architectures and Techniques*, 132-152. https://doi.org/10.70593/978-81-981271-4-3_7