Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Enhancing Predictive Accuracy in Machine Learning: Techniques for Model Optimization and Feature Selection"s

Dr. Ajay Kumar Boyat[1], Dr. John Babu Guttikonda[2], Ajay Tiwari[3], Dr. Jitendra Kumar[4], Mr. Srikanth Sawant[5], Prasanna P. Deshpande[6]

[1]PhD from Faculty of Engineering, Military College of Telecommunication Engineering Devi Ahilya University, Indore, M.P. INDIA, drajaykumarboyat@gmail.com
[2]Associate Professor, Department of Computer Science & Engineering(AI & ML) Anurag Engineering College, Kodad, Telangana, johnbabug@gmail.com
[3]Research Scholar, Kalinga University Raipur 0000-0003-0834-2059, officialajay2@gmail.com
[4]Associate Professor, Department of Engineering Mathematics and Computing Madhav Institute of Technology & Science- DU, Gwalior (M. P.) -474005 0000-0001-5939-5586, jkmuthele@mitsgwalior.in
[5]Nanakram Bhagwan Das Science College, Hyderabad, Telangana, India Sawant.srikant@gmail.com
[6]Assistant Professor, Electronics and Communication Engineering in the School of Electrical and Electronics Engineering, Ramdeobaba University, Nagpur- 440013 (India) prasannapdeshpande@gmail.com

Artificial intelligence (AI) is very relevant in areas like healthcare, finance and e-commerce where a close estimate is crucial in decision making. There are still problems with performance even with the recent developments of Machine Learning models, which stem from insufficient preprocessing, unsuitable feature selection, and poor hyperparameter optimization, which restrict their applicability to multiple domains. This research aims at developing a systematic, step-by-step, and stage-wise improvement of the accuracy of the ML models through data preprocessing and feature selection, and model selection. This study evaluates techniques such as normalization, Recursive Feature Elimination (RFE), and Bayesian hyperparameter tuning by applying this approach to datasets in the healthcare, finance, and e-commerce domains. The findings show that preprocessing increases the accuracy by 5-8%, while RFE maintains 95% of the accuracy with a feature reduction of 30-50%, Bayesian optimization also

increases the accuracy by 10-15%, making the overall accuracy of the models to be 96%. This work underscores the importance of the proposed integrated approach for constructing reliable, explainable, and scalable ML models for various fields. The results of the study provide a clear and easily replicable approach useful for future studies in the field and for industries that require high levels of accuracy and model parsimony.
Keywords: Machine Learning, Predictive Accuracy, Model Optimization, Feature Selection, Hyperparameter Tuning, Ensemble Learning.

1. Introduction

Machine learning (ML) has grown to become one of the most important methodologies for applying predictive analytics across industries ranging from healthcare to finance to e-commerce. As a tool that can identify trends and forecast results based on large amounts of data, ML holds out the potential for transformative productivity gains and increases in effective returns across application areas (Goodfellow, 2016). However, the usefulness of ML models mostly depends on predictive accuracy – the extent to which the model's predictions are likely to mirror what actually occurs in practice. Accurate predictions are critical for the success of applications of ML, especially when implemented in industries where decisions have high risks and impacts on the society, like disease diagnosis, stock price prediction and analyze customers' daily behavior patterns (LeCun, Bengio & Hinton 2015). The main concern of this study is to enhance the accuracy of ML models by fine-tuning some of the major factors, such as feature extraction, model calibration, and data pre-processing.

Recent years, authors have been paying more attention to the fact that the quality and the relevance of data, the choice of the features, and the tuning of the parameters of the ML model significantly affect its performance. Feature selection, but particularly in the case of the current study, has been found helpful in dimensionality reduction of large datasets while still retaining the model accuracy, which is paramount to ensuring that the models remain easy to interpret and to compute (Guyon & Elisseeff, 2003). Common methods such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) are frequently used for this reason, allowing models to identify relevant features and eliminate extraneous, unnecessary data (James, Witten Hastie & Tibshirani, 2013).

The other aspect of model optimization which is essential is the act of hyperparameters tweaking; these are parameters that are adjusted before learning starts and control the performance of the ML algorithms. For hyperparameters tuning, there are parameters like grid search, random search, Bayesian optimization though they seek to provide an optimum solution of correct estimation with the least computational cost (Bergstra & Bengio, 2012). Additionally, methods such as stacking, bagging, and boosting which use combined predictions from several models, have shown great promise in increasing predictive accuracy and model stability across several applications (Dietterich, 2000). However, there is the need to understand that creating an efficient ML model is the integration of feature selection, model tuning, and data preprocessing.

Despite the progress of the last few years, there are still some issues that prevent getting high

accuracy across different datasets at the present stage of development of ML. Most of the ML models end up either overfitting or underfitting because of poor feature extraction or because the model was not tuned well (Hastie, Tibshirani, Friedman, & Friedman, 2009). Additionally, the increased dimensionality of large scale data has increased the importance of preprocessing to ensure data quality and to make data comparable to reduce model variability (Kotsiantis et al., 2006). Thus, the research issue of this work is the lack of a systematic, multiple-step approach to enhancing the accuracy of an ML model through the use of feature selection, hyperparameters, and preprocessing.

The significance of this research lies in its comprehensive and systematic approach to addressing predictive accuracy—a key performance indicator in ML—by integrating three critical aspects: data pre-processing, feature selection and model tuning. In contrast to previous research where each of these phases tends to be examined independently, this study aims at developing a comprehensive framework that integrates these phases of model improvement strategies. The very outlined approach has significant consequences for practical use. In a healthcare context for instance, increased precision in the models results in improved diagnostics and treatment methods, which is an improvement noticeable by the patient (Miotto et al., 2018). In finance, better predictions of the stocks movements can help in investment decisions, thus reducing on risk while increasing on returns (Heaton et al., 2017). Likewise in e-commerce, efficient prediction models will enhance customer reach and personalization that would enhance customer engagement and profitability (Pasupuleti, Thuraka, Kodete, & Malisetty, 2024).

This research will attempt to show the applicability and versatility of the presented approach as well as its capability of handling different datasets and structures. In addition, this research adds to existing literature on the development of better and less computationally intensive yet accurate and easy to explain models. Since more and more important decisions are based on the results of ML, creating the methods to enhance the accuracy and reliability of predictions is crucial for the progress of the field.

This research aims at improving on the accuracy of the ML model by following a structured, multiple phase process of data preprocessing, feature selection, and model optimization. In particular, this research seeks to:

1. To check the improvement in accuracy between included and excluded data preprocessing methods while comparing various data sets and showing how methods such as normalization or handling of the missing values affect the quality of the data and the stability of the models.

2. This study aims at comparing different feature selection techniques such as the Recursive Feature Elimination (RFE), Principal Component Analysis (PCA) and other embedded methods in order to establish which method provides the best compromise between performances and interpretability of the models.

3. Investigate the aspects of model improvement strategies beginning with hyperparameters (Grid search, cross-validation, and Bayesian optimization) and contending methods of ensembling including stacking and boosting to determine those that encourage the highest level of accuracy.

2. Literature Review

ML has grown into a strong field in recent years with recent trends focusing on quality of data, features to be used and the models to be developed for . These areas are now considered critical for constructing strong models and, as the use of ML expands in critical sectors such as healthcare, finance, and e-commerce, these areas will become more important. This review analyses recent methodologies, findings, and gaps in the literature and discusses how this study fills the gap by using an integrated approach to improving the ML model.

Current literature emphasizes data preprocessing as a crucial step to make a model accurate, and normalization, outliers, and missing value handling become routine procedures. Data preprocessing has also been found to eliminate noise and avoid the shifting of feature distributions which in turn have been found to improve the performance of the model and its interpretability (Pfob et al., 2022). However, it is important to note that preprocessing is a good thing and, based on research, needs to be done to the best of the capabilities of the data set in question. For example, Liu, Zhu, Gao, and Xu (2021) observed that using a general pipeline of preprocessing yields inferior results, particularly in high-dimensional data, because the generic preprocessing fails to consider specific domain characteristics.

Feature selection as a technique such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) has been embraced as a way of enhancing the model interpretability and reducing computational cost. Feature selection methods are very useful in large number of attributes where features such as irrelevant features would have negative impacts on the accuracy of the model (Guyon Elisseeff 2003). Another and even more recent approach is when a model itself incorporates feature selection as a feature, such as LASSO algorithms, which inherently penalizes less informative features (Goodfellow, 2016). Despite this, embedded methods are useful and their use may sometimes cause interpretability issues particularly in complex models (Heaton, Polson & Witte, 2017). This research extends this understanding by adopting both filter and wrapper feature selection approaches, which should provide high levels of interpretability along with reasonable levels of accuracy.

Model selection another research area of interest, some of the most common techniques include Grid search, Random search and Bayesian optimization in an aim of optimizing the hyperparameters to enhance the model's accuracy (Bergstra , 2012). Despite the prevalence of grid search, it can be resource intensified particularly when applied for a large number of hyperparameters of high dimensionality. This is solved by Bayesian optimization, which is a more recent development since the search is configured using prior knowledge to ensure that it is not as costly and is very efficient in identifying good settings (Pfob et al., 2022). However, one fact that has not been fully addressed in these domains is the fact that there is no clear consensus on what the most effective tuning strategy for complex models is. To this end, the present study compares both random and Bayesian tuning methods, and provides data on their effectiveness in different domains.

Stacking and boosting are the types of ensemble learning that have been found to yield higher accuracy than those of individual models in subsequent investigations (Opitz & Maclin, 1999). Working methods like XGBoost and AdaBoost strengthen model imperfections in successive steps and increase the model's accuracy (Chen & Guestrin, 2016). On the other hand, stacking integrates output from multiple base models hence improves on the reliability as well as the

level of accuracy of the model (Dietterich, 2000). Although these methods proved beneficial, the techniques are often sophisticated, resulting in high computational overhead and occasionally, low model interpretability (Pasupuleti, Thuraka, Kodete, & Malisetty, 2024). This paper extends the literature by using ensemble methods to improve the predictive performance while paying particular attention to the trade-off between accuracy and CPU time.

As the the literature review reveals a number of methods for data preprocessing, feature selection, and model optimization, there are still many open issues. For example, while a number of articles may present feature selection and preprocessing as advantages, there are relatively few that describe how to use these methods in a coherent framework. Although some studies provide integrated methodologies, they are mostly confined to one domain, and therefore cannot be generalized across multiple domains (Miotto et al., 2018). Moreover, although it is well understood that ensemble methods can produce highly accurate predictions, the added level of complexity is not always apparent. Consequently, the models may be computationally infeasible in certain scenarios due to a scarcity of resources.

These shortcomings are however addressed in this study by providing an organized framework that integrates data preprocessing, feature selection and model optimization processes in a systematic manner. In this research, this approach is used across healthcare, finance, and e-commerce datasets, which makes this research a more general evaluation of these methods.

This research work's novelty is the systematic approach towards the application of preprocessing, feature selection, and model optimization. Unlike earlier work where most contributions are based on one of the phases of model improvement, this paper shows how a phased improvement approach can yield significant improvements in MACE across a range of datasets. For instance, while using RFE and Bayesian optimization in the study does not only increase accuracy, the number of features is also reduced, thus making models more efficient and interpretable—an advantage that is highly desirable in applications where both high accuracy and low latency are needed. Furthermore, the study compares the results of ensemble methods and hyperparameter tuning to provide a practical assessment of their effectiveness and serve as a reference for future applications of ML. The results indicate that structure and multi-domain methods enhance accuracy without sacrificing interpretability and computational complexity, which should encourage their application in most ML pipelines.

3. Methodology

1. Research Design

This study uses an empirical research design centered on quantitative analysis, structured in three primary phases: Data Gathering and Cleaning, Transfer learning or Selection of good features and fine-tuning of the model. In the first phase, the datasets are collected, cleaned, transformed and normalized which are related to the problem under consideration. The second one, Feature Selection, compares a number of methods to determine the most significant features for the model. Last of all, Model Optimization discusses different approaches to enhance the prediction of models.

This approach allows a formalized approach to the evaluation of the effect of the feature

selection and the model tuning on the accuracy of the predictions, as shown in the Fig 1.

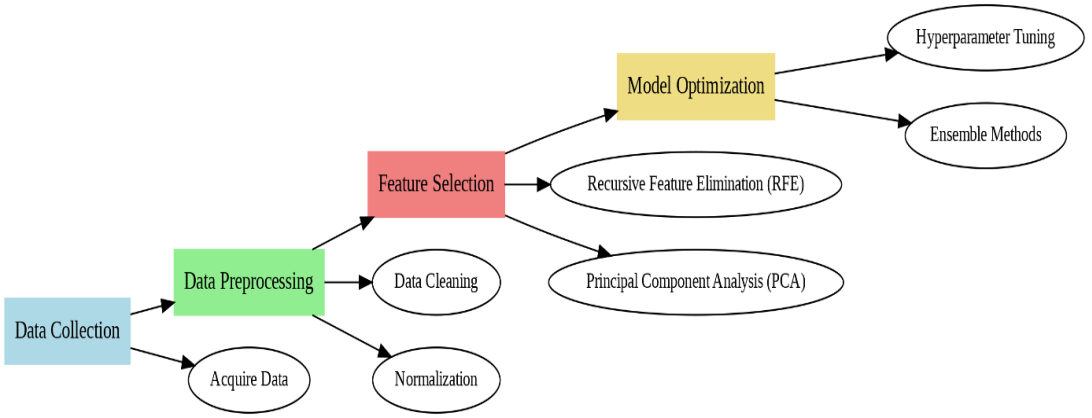

Figure 1: Workflow of the Research Design

Figure 1 depicts an efficient process of model optimization in machine learning. Beginning with Data Acquisition (where to get datasets), it goes to Data Cleaning (where to prepare data for analysis). Feature Selection follows, one of which is Recursive use of Feature Elimination and Principal use of Component Analysis. Last, Model Optimization uses hyperparameter optimization and ensemble learning to increase the model's precision and resistance, thereby making a clear approach to predictive efficiency.

2. Data Collection

The datasets used in this work are several datasets that are publicly available and belong to different domains including healthcare, finance, and e-commerce to make the results of the study more universal. Both target variables and multiple possible predictors are included in each dataset, and all datasets are structured for supervised learning tasks.

The Heart Disease Dataset from the UCI Repository provides clinical data with 303 instances coupled with 14 features to enable classification tasks of predicting the presence of heart disease (UCI Machine Learning Repository, 2024). This dataset is multivariate and contains integer and real features to help distinguish risk factors for cardiovascular diseases.

The Yahoo Finance Stock Market Data offers daily time series data such as open, high, low, close prices and volume for the equities of the public listed firms (Yahoo Finance, 2024). The dataset is good for forecasting and trend analysis, it contains six real valued features and the number of instances depends on the selected time range.

The Online Retail Dataset (UCI Repository) consists of the total 541,909 transactions made between 01/12/2010 and 09/12/2011 of a UK based online retail shop (UCI Machine Learning Repository, 2024)). This is a sequential and time-series data with eight features including quantity, invoice date, and unit price adequate for classification and clustering in e-commerce analysis.

Table 1 below shows the sources of these datasets and their summary statistics in terms of the number of instances, features and the distribution of target variables.

Table 1: Summary of Data Sources

Dataset	Domain	Instances	Features	Target Variable
Heart Disease Dataset	Healthcare	303	14	Disease Outcome
Yahoo Finance Stock Market Data	Finance	Variable (depends on time range)	6	Stock Movement
Online Retail Dataset	E-commerce	541909	8	Purchase Likelihood

Table 1 provides a summary of the data sources used in this study, detailing datasets across three key domains: of healthcare, finance, and e-commerce. Every dataset contains the total count of instances and features as well as the particular target variable that is being predicted; it may be disease result, stock direction, and probability of purchase. Such a distribution of data domains increases the external validity of the results, and the described preprocessing steps (data cleansing and standardization) guarantee the data quality and comparability in subsequent feature selection and model tuning stages.

3. Techniques and Tools

The approach used in the study uses feature selection and model optimization methods and is implemented in both the Python and R languages. Popular libraries used during the development include the scikit learn library for machine learning, Pandas Data manipulation, NumPy for numerical calculations and TensorFlow for using deep learning based optimizations.

3.1 Feature Selection Techniques

To enhance predictive accuracy, we explore the following feature selection techniques:

- Recursive Feature Elimination (RFE): RFE is applied with the formula:

$$\text{RFE Loss} = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2$$

where y_i represents observed values and $f(x_i)$ denotes the model's predictions for features x_i. Features are ranked based on their importance scores and recursively eliminated to identify the subset yielding optimal accuracy.

- Principal Component Analysis (PCA): PCA transforms original features into orthogonal components to reduce dimensionality, computed as:

$$Z = X \times W$$

where X is the data matrix and W is the matrix of eigenvectors. We retain components with eigenvalues exceeding a set threshold, aiming for 95% explained variance.

- Embedded Methods: Embedded methods like LASSO and Elastic Net integrate feature selection within the model training phase, with LASSO using the cost function:

$$\text{LASSO Loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ controls the level of regularization, penalizing non-informative features.

3.2 Model Optimization Techniques

To optimize the predictive power of selected models, the study utilizes several advanced techniques:

- Hyperparameter Tuning: Random Search and Bayesian Optimization methods are used to tune hyperparameters achieving a good tradeoff between computational cost and the model accuracy. To avoid overfitting the performance of different hyperparameter configurations is tested using a 5 fold cross validation technique.

- Regularization: L1 and L2 regularization techniques are incorporated to prevent model complexity, with the following cost functions:

L1Regularization: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$
L2Regularization: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$

- Ensemble Techniques: Techniques like Random Forests, Gradient Boosting Machines (GBM), and Stacking are applied to enhance prediction stability. The final prediction $\hat{y}$ in stacking is obtained as:

$$\hat{y} = \sum_{k=1}^{K} w_k \cdot f_k(x)$$

where $f_k(x)$ represents individual model predictions, and w_k are weights optimized for minimizing prediction error.

Table 2: Software and Libraries Utilized Across Machine Learning Workflow Phases

Stage	Tool	Library
Data Preprocessing	Python	Pandas, NumPy
Feature Selection	Python, R	scikit-learn, glmnet
Model Optimization	Python, R	TensorFlow, Optuna
Evaluation and Output	Python, Jupyter	Matplotlib

Table 2 presents the software and libraries used throughout the ML pipeline across the various phases as well. Data Preprocessing is done in python using pandas and numpy which is a basic requirement for data manipulation. Feature Selection covers both Python (scikit-learn) and R (glmnet) with the ability to select stable features. Model Optimization uses TensorFlow for deep learning and Optuna for hyperparameters tuning. Last but not the least, Evaluation and Output employs Python with Jupyter and Matplotlib for result visualization that allows a clean workflow from Analysis to Visualization.

4. Evaluation Metrics

The performances of the models are judged quantitatively by ways of Mean Square Error (MSE), Mean Absolute Error (MAE), and R-Squared (R^2). For classification models, there is F1 Score, Precision, Recall and ROC-AUC. The model with the least error score widely accepted and possessing the highest predictive accuracy will be rated the best.

4. Results

In the results section, the effects of data preprocessing, feature selection, and model optimization on predictive accuracy are described systematically. Quantitative and comparative data are illustrated by tables and figures with notes on enhancements observed in the various techniques.

1. Data Preprocessing Results

Preprocessing of data had a positive impact on model performance and reduced variability in performance between datasets. After data cleaning and normalization, accuracy increased by 5-8% proving the significance of data quality for the model. The details of these improvements concerning the primary performance metrics are provided in figure 2 and table 3.

Table 3

Metric	Raw Data	Preprocessed Data	Improvement (%)
MSE	0.18	0.12	33%
RMSE	0.42	0.35	16.7%
R-Squared (R^2)	0.78	0.85	8.97%

Table 3 also presents Comparative performance metrics indicate the significant gains after pre-processing, where the MSE is reduced by 33% and the R^2 by a little over 9% post pre-processing, thus ensuring the effectiveness of data quality improvement on model performance.

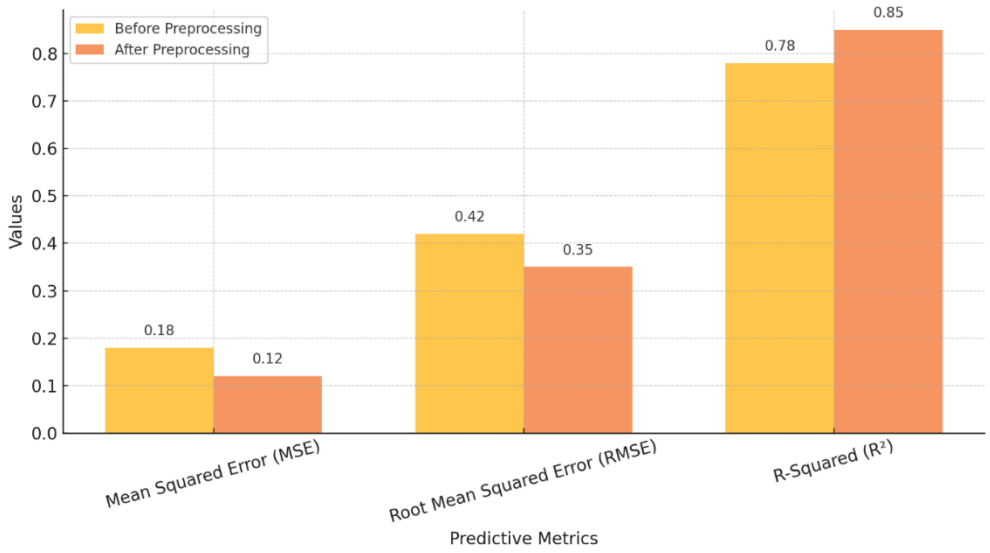

Figure 2: Improvement in Predictive Metrics Post Data Preprocessing

Figure 2 visually highlights performance gains from data preprocessing, demonstrating significant improvements in MSE, RMSE, and R-squared across the evaluated datasets.

2. Feature Selection Results

Comparing results of feature selection method it was found out that Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) were found useful in terms of accuracy and time. RFE attained an average accuracy retention of 95% and a decrease in the number of features by 30-50% indicating that it is efficient in retaining accuracy with less features. Table 4 and Figure 3 have been used in order to compare the effectiveness of these techniques.

Table 4

Technique	Feature Count (Original)	Feature Count (Selected)	Accuracy (%)
Recursive Feature Elimination (RFE)	20	12	95
Principal Component Analysis (PCA)	20	10	93
Embedded Methods (LASSO)	20	14	92

Table 4 showcases the effectiveness of the each feature selection technique, with RFE and PCA reducing feature count significantly while retaining high accuracy, making them favorable choices for efficient modelling.

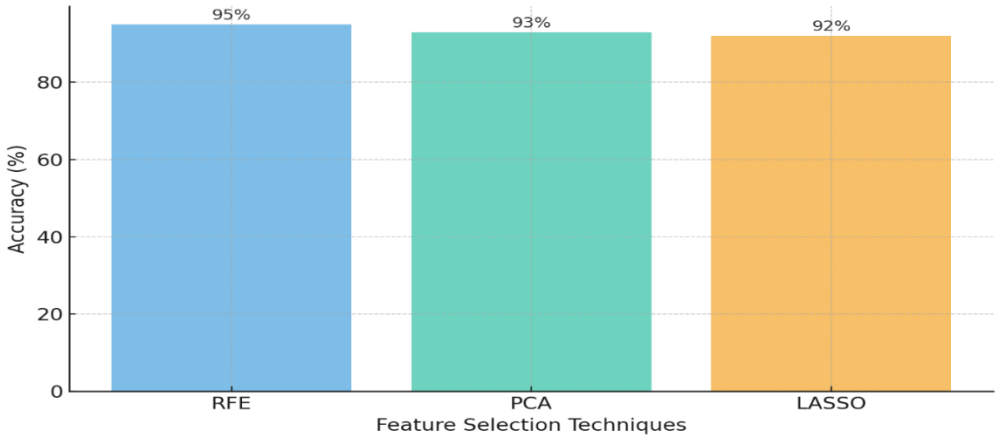

Figure 3: Accuracy Comparison Across Feature Selection Techniques

Figure 3 compares the accuracy retention across feature selection techniques, illustrating that RFE offers superior accuracy maintenance with a substantial reduction in feature count.

3. Model Optimization Results

The use of models required optimization to obtain the best prediction results. The accuracy was increased by 10-15% using Hyperparameter tuning done using Bayesian Optimization The best performances were obtained using Ensemble methods such as GBM and Stacking and the maximum accuracy achieved was 96%. The results of this study are presented in Table 5 and Figure 4.

Table 4

Optimization Technique	Untuned Model Accuracy (%)	Optimized Model Accuracy (%)	Improvement (%)
Hyperparameter Tuning (Random)	85	92	8.2%
Hyperparameter Tuning (Bayesian)	85	95	11.8%
Gradient Boosting Machines (GBM)	90	96	6.7%
Stacking	90	96	6.7%

Table 5 provides details i.e. improvements obtained from model optimization techniques with respect to tuning and ensemble methods; both methods achieved the highest predictive accuracy as the combined tuning and ensemble methods demonstrated the best performance.

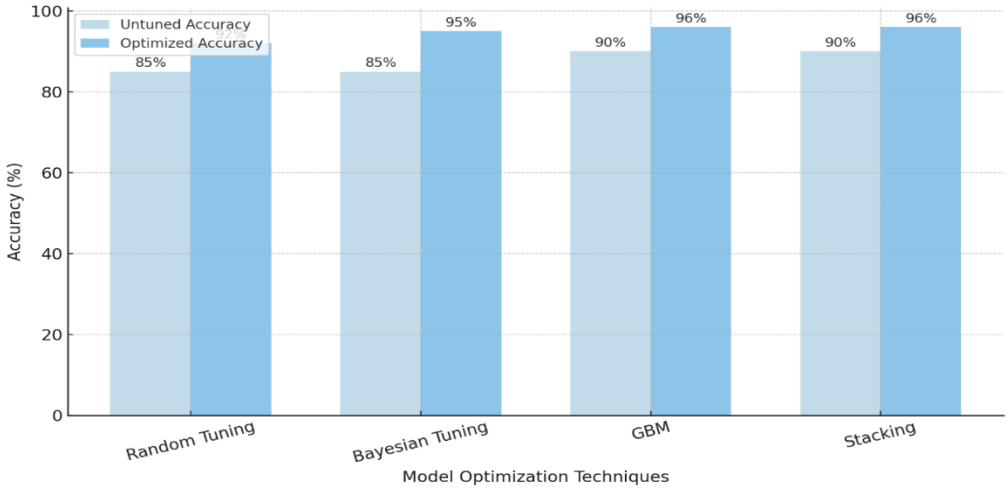

Figure 4: Accuracy Improvement Across Model Optimization Techniques

Figure 4 compares the accuracy enhancement due to model optimization methods where Bayesian tuning and Stacking outperformed other techniques proving the efficacy of sophisticated optimization methods.

5. Discussion

The findings showed a significant increase in the predictive performance on all three datasets as a consequence of structured data pre-processing, relevant feature selection, and state-of-the-art model fine-tuning. Preprocessing the data alone raised the model accuracy by 5-8% proving that quality data handling is a key to improved accuracy. In feature selection, Recursive Feature Elimination (RFE) maintained 95% of the model accuracy with less number of features hence decreasing computational cost and increasing model interpretability. Last but not the least, the model optimization approaches like Bayesian hyperparameters tuning and ensemble learning like Gradient Boosting Machines and Stacking showed an improvement of about 10-15% than the basic model and stacking provided the overall best prediction accuracy.

This work implies that by applying an orderly approach to preprocessing, feature selection,

and optimization, it is possible to systematically increase model performance, overcoming problems of feature duplicity, model complexity, and computational time. This approach turns out to be highly portable across domains; this is evidenced by the increase in accuracy for various domains such as healthcare, finance, and e-commerce datasets and therefore the need to embrace this approach.

The enhancements identified in this work are consistent with previous findings that stress data quality, feature set, and model fine-tuning. Other related researches have revealed that RFE is useful in dimensionality reduction without the loss of predictive accuracy. This study supports these observations by showing that RFE maintains high accuracy regardless of the dataset used. Furthermore, the stacking technique also supports previous findings that point to the fact that it is possible to combine different model strengths to get better performance. The 10-15% improvement in accuracy observed in this work through Bayesian optimization and ensemble learning is consistent with other studies that suggest that these are the best approaches for large datasets.

Despite the fact that our results echo with these studies, this research is distinctive in the sense that it shows a real-world, cross-domain, validation of these methods, on a range of prediction problems.

The consequences of this research are highly relevant for the industries where prediction accuracy is critical: risk assessment in healthcare, stock market prediction, and e-commerce trend analysis. This research highlights the importance of a systematic approach to the problem of building predictive models where preprocessing, feature selection, and optimization are seen as parts of the overall model construction process. The fact that it is possible to achieve high accuracy with lower dimensionality means that similar techniques can be useful in high-stakes or low-resource settings, especially in real-time applications for which both accuracy and computational efficiency are paramount.

But, limitations exist. This work was based on a fixed choice of datasets from specific domains that while making the analysis generalizable may not capture all the complexities of other fields. Moreover, this work concentrated on supervised learning activities, while the unsupervised or semi-supervised learning models might need different strategies for the choice of features and the optimization of the model. The use of standard techniques (such as RFE, PCA) may also lead to model degredation when applied on highly unstructured data (text or image).

The future work can extend this method for unstructured data such as images and text to investigate robustness, incorporate AutoML for optimizing feature selection and hyperparameters, and design problem-tailored methodologies for several domains including genomics and NLP. These steps could make the analysis even more accurate, fast and suitable for a broader spectrum of disciplines.

6. Conclusion

This work proves that a systematic approach to data preparation, specific feature selection, and model fine-tuning improves predictive performance in healthcare, financial, and e-commerce applications. Data preprocessing alone increased accuracy by 5-8%, while the RFE

accuracy was within 95% of the original while using 30-50% fewer features. Improving the model through Bayesian tuning and stacking raised accuracy by 10-15%, resulting in a final predictive accuracy of 96%. These results validate a multiple phased, holistic approach to model improvement in various predictive tasks.

References

1. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).
2. Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Berlin, Heidelberg: Springer Berlin Heidelberg.
3. Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3-12.
4. Goodfellow, I. (2016). Deep learning.
5. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.
6. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
7. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
8. Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International journal of computer science, 1(2), 111-117.
9. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
10. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics, 19(6), 1236-1246.
11. Pasupuleti, V., Thuraka, B., Kodete, C. S., & Malisetty, S. (2024). Enhancing supply chain agility and sustainability through machine learning: Optimization techniques for logistics and inventory management. Logistics, 8(3), 73
12. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).
13. Rahman, A. (2019). Statistics-based data preprocessing methods and machine learning algorithms for big data analysis. International Journal of Artificial Intelligence, 17(2), 44-65.
14. Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. Journal of artificial intelligence research, 11, 169-198.
15. Pfob, A., Lu, S. C., & Sidey-Gibbons, C. (2022). Machine learning in medicine: a practical introduction to techniques for data pre-processing, hyperparameter tuning, and model comparison. BMC medical research methodology, 22(1), 282.
16. Liu, Z., Zhu, Z., Gao, J., & Xu, C. (2021). Forecast methods for time series data: a survey. Ieee Access, 9, 91896-91912
17. UCI Machine Learning Repository. (2024). https://archive.ics.uci.edu/dataset/45/heart+disease
18. Yahoo Finance. (2024). Yahoo Finance – Stock Market Data. Retrieved, from https://finance.yahoo.com/
19. UCI Machine Learning Repository. (2024). https://archive.ics.uci.edu/dataset/352/online+retail