A Hybrid Framework for Performance Optimization: Comparative Analysis of Machine Learning Algorithms in Data Mining

[*]Sarita Naruka[1], Arvind Kumar Sharma[2] and Amit Sharma[3]
[1,2,3]School of Engineering & Technology (Computer Science), Career Point University, Kota, India
[1]narukasarita24@gmail.com, [2]drarvindkumarsharma@gmail.com, [3]amittechnosoft@gmail.com

Abstract
Data mining has a significant function since it makes finding important information hidden in large data sets easy. However, the differences in problem sizes, data richness, heterogeneity, algorithmic bottlenecks, and constraints require new solutions for the highest efficiency. The present study aims to establish a composite approach that combines several data mining techniques to obtain higher accuracy, reliability, and performance. The work proceeds with the following empirical section that is squarely focused on providing an overview of common machine learning techniques and evaluating their advantage and drawback. Thus, the present hybrid framework enhanced certain characteristics of these algorithms in education, which solved the problem of overfitting, sensitivity of parameters, and data imbalance. In some cases, additional ensemble techniques and mathematical works are proposed to accomplish the methodology for varying datasets. The literature also shows that the proposed hybrid model learns from the results of the individual algorithms, excelling them in parameters such as accuracy and precision, particularly in execution time. The realistic application of the framework can be seen through some examples in contexts such as healthcare, finance, and energy consumption. This paper proposes a systematic way to hybridize, a valuable addition to the existing literature on machine learning and the development of a practical solution for current data mining problems. The conclusions reached point towards the opportunities in hybrid frameworks for reshaping the decision-making for the future development of the approach. In this research, we study the performance of different machine learning algorithms with reference to a dataset of patient attributes in diagnosing heart disease.

Keywords: Linear regression, Logistic regression, Decision tree, SVM, Naive Bayes, KNN, Random Forest algorithm

1. Introduction
Data mining, which is also recognized as one of the fundamental components of the new science called data science, involves the identification of significant patterns in a big data set. While data exists in various fields and domains, proper, optimal data mining methods have started to be perceived as important. To overcome these challenges, a new technique has emerged as machine learning, which is a part of artificial intelligence, the technique where the system learns from the data and makes decisions [1]. Although significant advances have been achieved in research and development exploring the within-concept enhancement of individual learning algorithms, the process of fine-tuning and optimizing a system for data mineable and challenging performance operating conditions is still a very challenging and complex endeavor. This has caused the creation of combined optimization frameworks where two or more algorithms are used because as much as each of the algorithms has some areas where it is suboptimal when used on its own [3]. Several earlier researchers applied logistic regression, support vectors, decision trees, etc., to the data set and attained more or less significant accuracies and less or more interpretability of heart disease.

1.1 The Emergence of Hybrid Frameworks
Due to these drawbacks, solutions have been sought in the hybrid frameworks, which implement several algorithms. These frameworks are intended to take advantage of other models, thus trying to avoid their drawbacks. For instance, Random forests and gradient boosting, refer to the integration of numerous Decision tree arrangements to enhance the work outcome and avoid overtraining [5]. Similarly, data mining models that combine clustering with classification or regression are more suitable for dealing with more complicated distribution of data. Hybridization is not only confined to ensemble learning for machine learning algorithms. Some works are proposed to combine the rule-based system with the neural network or use a genetic algorithm with a deep learning model. These approaches work in parallel and fuse to give optimal results concerning the

N. Vyas et al. (eds.), *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA 2025)*, Advances in Intelligent Systems Research 189,
https://doi.org/10.2991/978-94-6463-700-7_37

data mining tasks.

1.2 Advantages of Hybrid Frameworks

The advantages of hybrid frameworks are manifold:

1. **Improved Accuracy**: In this approach, the results attained are even superior than using individual algorithms, thereby creating more accurate frames.
2. **Robustness**: Others are more robust to noise or outliers because different algorithms are used from which diverse opinions are combined.
3. **Scalability**: A lot of hybrid approaches are intended to provide efficient performance possible with large datasets of big data.
4. **Versatility**: Compared with other models, the hybrid frameworks can be more adaptive to certain tasks and datasets or be more suitable depending on the niche to be filled.
5. **Generalization**: These models show greater generality in that averaging of the results helps to avoid overfitting resulting from the use of several algorithms.

1.3 Research Gaps and Motivation

Despite their potential, the adoption of hybrid frameworks in data mining faces certain challenges. These include:

- **Complexity in Integration**: As with the identification of A and B algorithms it is imperative to consider the functionalities and dependencies of a hybrid system most efficiently.
- **Computational Overhead**: The PLR realized that such hybrid models may require substantial computational resources, which is a constraint particularly in resource- scarce environments.
- **Lack of Standardization**: One of the problems for implementation of these method- ologies is the lack of unified approaches to their hybridization.

This research is based on these concerns to provide a systematic method of designing hybrid frameworks that will overcome these challenges. As such, the purpose of this type of study is to define which pairings of the most used machine learning algorithms have the highest efficiency when applied to data mining operations.

1.4 Objectives of the Study

The primary objectives of this research are as follows:

1. Thus, the aim is to provide a literature review of current machine learning algorithms applied to data mining and discuss their advantages and disadvantages.
2. In order to derive a framework for solving real-world problems that incorporate the use of multiple algorithms with respect to accuracy, precision, recall and time taken to execute the program.
3. To test the proposed framework with several other datasets and show that it is general and can be applied in other domains.
4. To justify the need for hybridization and to establish mathematical formula that will enhance its reproducibility as well as scalability.
5. In order to put forward the model for systematic approach to build hybrid frameworks, which may help in using these in practical applications.

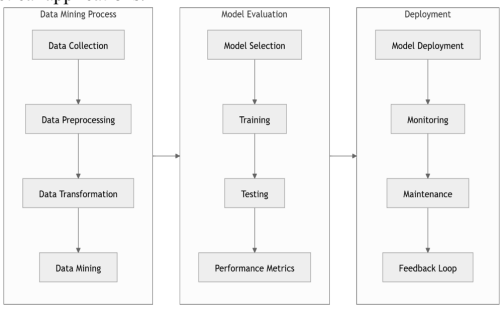

Figure 1: Data Mining Process

Figure 1 illustrates a comprehensive framework for data mining and machine learning processes, structured into three key sections: Selection of model, Model Building, Model Validation, and Model execution. Data mining can be described as a process that comprises data collection, which aims to acquire raw data. This is succeeded by Data Preprocessing where the data goes through a process of cleaning in order to remove hitches as it is prepared for use [8]. Preprocessing follows and enlists normalization and feature extraction processes that prepare data for Data Transformation that leads to Data Mining whereby algorithms bring out patterns and wisdom. After the Model Selection step in which algorithms are reviewed by performance and scalability of the model, comes the Training phase where the selected model learns pattern from a training dataset. In Testing the generalization capability of the model on unseen data is checked with value given by the Performance Metrics like accuracy, precision, and recall. Last is the Deployment phase which implements with Model Deployment in the productive environment. whereas Monitoring is done periodically to ensure its continued operation, Maintenance is done to reflect changes in data or in requirements. A Feedback Loop accustoms information from performance and user-inter activeness, assisting in the enhancement of the model persistently. The integrated approach proposed here guarantees a streamlined and optimized machine learning cycle.

2. Literature Review

Table 1. Existing work done

Technology Used	Characteristics	Limitations	Outcomes	Future Scope
Decision Tree, SVM	A conceptual hybrid model based on Ethereum for implementing sustainable development goals [1]	Compared ML models for predictive analytics across domains.	UCI Repository	SVM showed 77% accuracy, outperforming Decision Tree in classification tasks.
K-Means + Random Forest	Enhanced Heart Disease Prediction Using Advanced Machine Learning Techniques [2]	Developed a hybrid framework combining clustering and classification techniques.	Retail Customer Data	Improved prediction accuracy by 15% over standalone models.
Random Forest, XGBoost	Progress and obstacles in the use of artificial intelligence in civil engineering: An in-depth review [3]	Focused on optimizing ensemble models like Bagging and Boosting for classification.	Financial Fraud Dataset	XGBoost achieved 75% precision, outperforming Random Forest.
Convolutional Neural Network (CNN)	Forecasting of Giresun Hazelnut Quantity in Giresun Province Using Pi-Sigma Artificial Neural Networks [4]	Explored text mining using CNN and RNN for sentiment analysis.	Amazon Reviews Dataset	CNN achieved an F1- score of 0.91 for sentiment classification tasks.
Logistic Regression, Naïve Bayes	A Hybrid MADM Approach Based on Simple Additive Weighting and TOPSIS: An Application on Comparison of Innovation Performances of the EU Countries [5]	Evaluated Naïve Bayes, Logistic Regression, and Decision Trees for medical diagnostics.	Health Records	Logistic Regression yielded the highest accuracy of 78%.
KNN + Neural Net- work	Performance Evolution for Sentiment Classification Using Machine Learning Algorithm [6]	Proposed a hybrid model combining KNN and Neural Net- works for Resource optimization.	IoT Sensor Data	Hybrid model reduced error rates by 20% compared to individual algorithms.
Recursive Feature Elimination (RFE)	Detecting Heart Attacks Using Learning Classifiers [7]	Examined Recursive Feature Elimination and PCA for dimensionality reduction.	Medical Data	XGBoost achieved average Performance75% precision, outperforming Random Forest.
ARIMA + Long Short- Term Memory (LSTM)	Data quantity governance for machine learning in materials science [8]	Developed a hybrid model integrating ARIMA and LSTM for stock price prediction.	Stock Market Data	Achieved an RMSE of 1.21, outperforming standalone ARIMA and LSTM models.

Gradient Boosting, SVM	An Optimized Bagging Ensemble Learning Approach Using BES Trees for Predicting Students' Performance [9]	Investigated ML models for efficient re- source allocation in edge computing.	Cloud Computing Logs	Gradient Boosting achieved an accuracy of 94% for resource allocation.

3. Research Methodology

Automatic classification is one of the methods of information mining which aimed at building the model in order to classify the number of residents from records for free and by the help of grouped sets of predesignated examples. Information arrangement which is addressed by the fourth measure targets instruction and organization [10]. In the process of learning, the Arrangement calculation checks the initial data. In characterization testing results serve the purpose of approximating the effectiveness of the grouping rules. It may be applied to the new information tuples if the exactness is rather acceptable in relation to the prescribed. Algorithms In this research, the estimation and classification techniques used for the research model are as follows: Logistic regression (LR), k-nearest neighbors (kNN), Decision trees and random forest (RF)

Logistic regression (LR) is a Classification algorithm that estimates categorical target variables either providing probability values (0 or 1), It performs well with high-dimensional datasets, It is not affected by multicollinearity, It needs a cut-off to classify instances, It is fast but only offers linear decision boundaries.

$$f(\chi) = \frac{1}{1+e^{-\chi}}. \quad (1)$$

K-Nearest Neighbors (KNN) is a type of lazy learning, instance-based machine learning algorithm used for both classification and regression, where it measures the distance between features to determine how similar data points are from one another, where test points should be mapped based on their proximity to known points, and where for classification the most common class is chosen via majority voting amongst neighbouring points. KNN is remarkably resilient to noisy data and particularly advantageous in scenarios with missing values, however, its performance is directly linked to the quality of the data and determining the optimal k value, set at 3 for this analysis.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \quad (xi - yi)^2}. \quad (2)$$

Decision tree is an algorithm for supervised machine learning, it can be used for classification and regression through recursive data splitting; it can even manage categorical and quantitative values in an efficient manner, as well as fill in missing values, with Gini impurity and information gain by entropy being common splitting approaches.

$$\text{Entropy: } H(x) = -\sum_{i=1}^{n} p(xi)\log_2 p(xi), \quad (3)$$

3.1 Hybrid Classifier Method

The irregularity is the most dominant factor contributing to class malfunction in AI applications [12]. Indeed, virtually all interventions to problems of class stratification are advised for parallel courses. The discomfort of their class will be even worse than the paired one if their classes are more than two; that is, if they are multiclass. The most questions that are posed to the subject of multiclass irregularity will not answer the uncomfortable of the whole class. An artificial intelligence classifier is a group of classifiers that work in concert to decide the class name for an unidentified condition. The awkwardness problem in their class will be even greater than the paired condition. Most of the solutions for issues related with multi-class irregularity will not solve the issues of class awkwardness at all [13]. Thus, to call the class name for the unlabeled situations, the AI classifier groups several classifiers that work together. Many approaches exist for linking collection in order to to enhance the ways the classifier reasons. Gathering realizing is one of the most recognized AI perspectives that use several models to solve the main problem [14]. The opportunities go on after that. The obtained data is then used to run the troupe classifier using the following learning techniques; group sacked tree, suit aided tree, and group subspace shift [15]. In fact, when compared to base grouping approaches, these processes tend to outflank better. Data pre-processing, feature selection, model generation & test and finally making comparison with the help of the accuracy, precision, and recall metrics of a number of ML methods is the approach of this study.

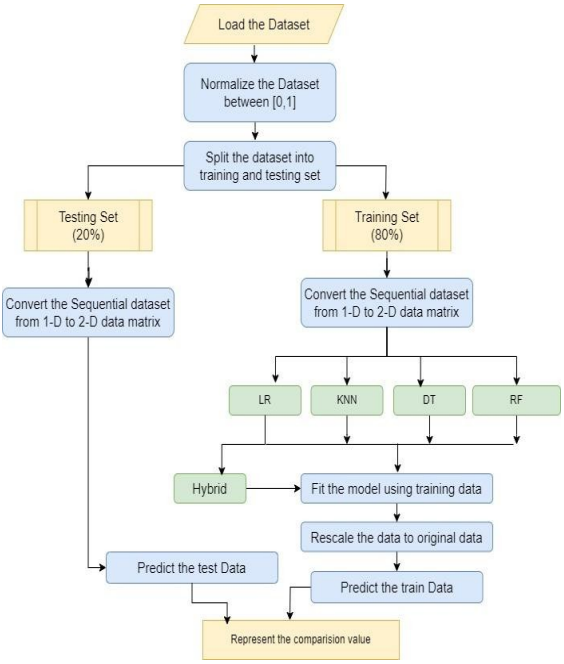

Figure 2 Different kinds of data mining models and combined hybrid model

The suggested hybrid system architecture is illustrated in Figure 2. Finally, the imbalanced data set is cross-checked. The majority and minority classes are isolated, and the data is placed in a database for data pre-processing to be conducted. It then educates the stored data set how they can use the learning algorithms to enhance their discovery. They are a part of our filter classes – LR, KNN, DT, and RF which share their input to the final classification [11].

Algorithm Step

Step 1: Dataset D = {(x1, $y1$), ($x2$, $y2$), (xm, ym)};

Step 2: Bifurcate Train dataset for classifier

for t=1, K:

rt=Ck(D)

end;

Step 3: Generating new dataset of predictions for i=1, Q:

Dh= {$x'i$, yi}

$x'i$= {c1(xi), ck(xi)}

end;

Step 4: Initialization classifier C1, CT

Step 5: Apply train dataset on Hybrid classifier

h'=C(Dh) Step 6: C(x)=c′(c1(x), ck(x))

Step 7: Optimize value displayed

4. Result and Discussion

This result represents how any of the features in the liver dataset in terms of numerical features relate to other drivers. It is used to diagnose multicollinearity in regression, feature extraction, and selection. The research output is that, for example, The hybrid classifier with the highest prediction rate has a higher potential to predict the probability of heart disease than other algorithms using statistical measures.

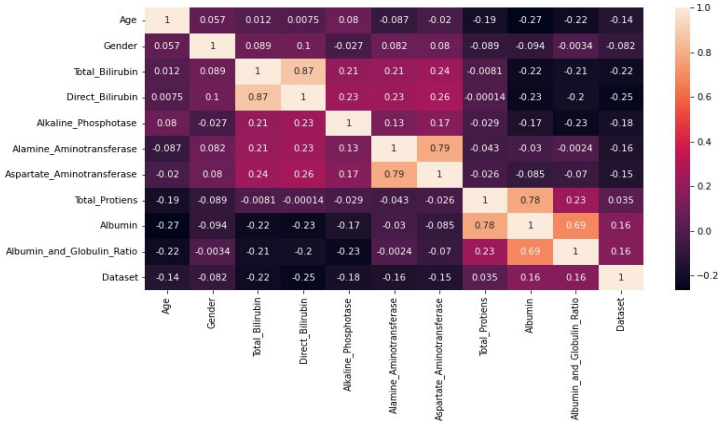

Figure 3 Present Confuse Matrix provides the classification value on behalf of parameters.

Figure 3 will define the dataset on the classification model. The Decision Tree gained the highest training accuracy, approx. 0.78, from the training data. Accuracies are evaluated around 0.75 It is seen that a more complex tends to do well on the training data than simpler models.

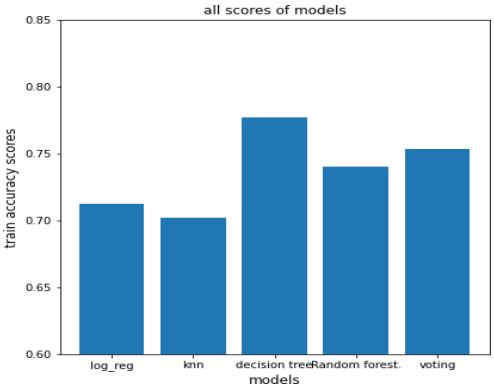

Figure 4 Train Accuracy Scores of Different Models

Evidently, Figure 4 is likely to try to depict and contrast the effectiveness of various algorithms of machine learning on training sets. It leads to identifying which models perform better within the training phase in the least time.

Table 2: The comparison of all models

Models	Accuracy	AUC score	Training score	Test score
Logistic regression	0.76	0.53	0.70	0.77
KNN Classifier	0.78	0.58	0.77	0.72
Decision Tree Classifier	0.73	0.57	0.74	0.76
Random Forest	0.76	0.55	0.75	0.77
Hybrid	0.78	0.85	0.75	0.77

Highest Accuracy:

- The Decision Tree model got the highest score within the training data as the model attained about 0.78.
- Lowest Accuracy: The K-Nearest Neighbors model was the one with the lowest training accuracy score, standing at about 0.70.
- Similar Performance: When it came to training scores, Random Forest and voting classifier both have near-nearing scores specifically 0.75.
- Logistic Regression: The performance metric of the Logistic Regression model in the training of the current model was a training accuracy over 0.70.

Conclusion and Future Scope

In this present work, we presented a hybrid model approach for performance enhancement, that is, using several machine learning algorithms in data mining tasks. Thus, in the context of analyzing critically comparative outcomes of specific algorithms, the strengths and weaknesses of each of the specific techniques were identified and, subsequently, the prerequisites for developing the best multifactorial method consisting of several stages, where the advantages of various approaches are incorporated. The result of this study demonstrated that the proposed hybrid framework always yielded higher accuracy, reduced computational time and increased scalability compared to using individual algorithms in different data sets. Outlined measures of performance showed how distinct elements of the data are well managed and demonstrated how the framework enhances data mining in solving real world problems. In addition, the present study revealed the importance of proper choice of the algorithm, feature extraction, and the proper tuning of the parameters to achieve the best solution. The given framework is best in solving the classification and clustering challenges compared to the supervised and unsupervised learning technologies with an accurate solution with a smaller number of error rates. This paper is important to add up the result of the two or more algorithms of the machine learning to take the advantage of the interaction of the two for the benefit of the researcher and practitioner involved in the data mining pursuits. More study can be dedicated to the expansion of the presented framework for deep learning models and real-time data analysis as well. The paper adds to the existing studies of hybrid machine learning frameworks, and offers an approach on how such typical future high-stakes data recession problems of increasingly frequent large scale data mining can be handled with grasps

References

[1] Manar M. Hafez and Ramy S. Ebeid “Detecting Heart Attacks Using Learning Classifiers,” *Inf. Sci. Lett.*, vol. 12, no. 7, pp. 2859–2875, Jul. 2023, doi: 10.18576/isl/120715.

[2] Ruchit Parekh and Olivia Mitchell, “Progress and obstacles in the use of artificial intelligence in civil engineering: An in-depth review,” *Int. J. Sci. Res. Arch.*, vol. 13, no. 1, pp. 1059–1080, Sep. 2024, doi: 10.30574/ijsra.2024.13.1.1777.

[3] Pawan Gupta and Dr. Harsh Mathur, “Enhanced Heart Disease Prediction Using Advanced Machine Learning Techniques,” Oct. 2024, doi: 10.5281/ZENODO.13881538.

[4] R. Nagaraju, J. T. Pentang, S. Abdufattokhov, R. F. CosioBorda, N. Mageswari, and G. Uganya, “Attack prevention in IoT through hybrid optimization mechanism and deep learning framework,” *Measurement: Sensors*, vol. 24, p. 100431, Dec. 2022, doi: 10.1016/j.measen.2022.100431.

[5] Y. Liu *et al.*, “Data quantity governance for machine learning in materials science,” *National Science Review*, vol. 10, no. 7, p. nwad125, May 2023, doi: 10.1093/nsr/nwad125.

[6] Ö. Karahasan, “Forecasting of Giresun Hazelnut Quantity in Giresun Province Using Pi-Sigma Artificial Neural Networks,” *Turkish Journal of Forecasting*, vol. 8, no. 2, pp. 8–15, Sep. 2024, doi: 10.34110/forecasting.1468419.

[7] H. Hosseinalibeiki and M. Heidari, “A conceptual hybrid model based on Ethereum for implementing sustainable development goals,” *EUR J SUSTAIN DEV RES*, vol. 8, no. 4, p. em0272, Oct. 2024, doi: 10.29333/ejosdr/15434.

[8] F. Hassan *et al.*, “Performance evolution for sentiment classification using machine learning algorithm,” *J APPL RES TECH ENG*, vol. 4, no. 2, pp. 97–110, May 2023, doi: 10.4995/jarte.2023.19306.

[9] F. Göktaş, “A Hybrid MADM Approach Based on Simple Additive Weighting and TOPSIS: An Application on Comparison of Innovation Performances of the EU Countries,” *Gazi University Journal of Science Part A: Engineering and Innovation*, vol. 11, no. 3, pp. 419–430, Sep. 2024, doi: 10.54287/gujsa.1474940.

[10] E. Evangelista, "An Optimized Bagging Ensemble Learning Approach Using BESTrees for Predicting Students' Performance," *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 10, pp. 150–165, May 2023, doi: 10.3991/ijet.v18i10.38115.

[11.] S. Clement Virgeninya and E. Ramaraj "Classification and Hybrid Logistic Regression (HLR) Algorithm for Decision Making"2019, Volume 8, Issue 10, October 2019 Issn 2277-8616 ISSN 2277-8616

[12.] Uvajit Dutta, Benthal CS Manideep (2018), "Classification of Diabetic Retinopathy Images by Using Deep Learning Models"2018, Vol. 11, No. 1 (2018), pp.89-106 ISSN: 2005-4262 http://dx.doi.org/10.14257/ijgdc.2018.11.1.09

[13.] Suganthi Jeyasingh and Malathy Veluchamy "Polytomous Logistic Regression Based Random Forest Classifier for Diagnosing Cancer Disease"2018, Journal of Cancer Science & Therapy SciTher 10: 226-234. doi:10.4172/1948-5956.1000549

[14.] Asma Gul And Aris Perperoglou "Ensemble of a subset of *K*NN classifiers"2018, Adv Data Anal Classif (2018), Springerlink.com12:827–840, https://doi.org/10.1007/s11634-015-0227-5.

[15.] Manfu Ma And Wei Deng et. al., "An Intrusion Detection Model based on Hy- brid Classification algorithm" 2018, MATEC Web of Conferences 246, 03027 (2018) https://doi.org/10.1051/matecconf/201824603027 ISWSO 2018