Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multidimensional Feature Optimization and Machine Learning-Based Predictive Framework for Automated Wine Quality Assessment

Snehit Shaprapawad
Tech Mahindra
Premkumar Borugadda
premkumar.borugadda@becbapatla.ac.in
Bapatla Engineering College
Vengalarao Pachava
NALSAR University of Law
N. Vinoda
Acharya N.G Ranga Agricultural University
Phaniendra Alugoju
Palacky University

Research Article

Keywords: Wine Quality Assessment, Machine Learning, Feature Selection, Dimensionality Reduction, Hybrid Ensemble Learning

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11466347/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multidimensional Feature Optimization and Machine Learning-Based Predictive Framework for Automated Wine Quality Assessment

[1]Snehit Shaprapawad, [1]Premkumar Borugadda[*], [3]Vengalarao Pachava, [4]N. Vinoda, [5]Phaniendra Alugoju

[1]Tech Mahindra, India
[2]Dept. of CSE (AI & ML), Bapatla Engineering College, Bapatla-522102, India.
[3]Dept. of Management Studies, NALSAR University of Law, Hyderabad, India22102,
[4]Dept. of Food Process Engineering, Acharya N.G Ranga Agricultural University, Lam, Andhra Pradesh, India
[5]Dept. of Biophysics, Faculty of Science, Palacky University, Slechtitelu 27, Olomouc, 77900, Czech Republic.

***Corresponding Author: premkumar.borugadda@becbapatla.ac.in**

Abstract

Assessment of wine quality is an essential step in the wine industry, where traditional evaluation methods mainly depend on expert sensory analysis and physicochemical testing. However, these conventional approaches are often time-consuming, expensive, and influenced by subjective human judgment. In this context, wine quality evaluation by alternative computer based-methods is gaining substantial interest to overcome traditional assessment related challenges. In this context, the present study aimed at investigating the wine quality by machine learning based approaches using available wine datasets of different origin and to promote the speed of automation in wine industry. The proposed framework integrates both feature selection methods (e.g., multiple dimensionality reduction techniques, including filter-, wrapper-, and embedded-based) and feature transformation techniques (e.g., principal component analysis (PCA) and Linear Discriminant Analysis (LDA)). Furthermore, we employed other machine learning and ensemble models such as Logistic Regression, K-Nearest Neighbour, Random Forest, Support Vector Classifier, Extra Trees Classifier, Gradient Boosting, Voting Ensemble, Stacking Ensemble, and a Hybrid Ensemble models, for the assessment of wine quality classification. Experimental results demonstrate that the correlation-based filter feature selection method achieved the best overall performance when combined with the proposed Hybrid Ensemble model, obtaining an accuracy of 89.08% and an F1-score of 88.93%. The findings indicate that dimensionality reduction combined with ensemble learning significantly improves prediction performance while reducing feature redundancy and computational complexity. The proposed approach provides an efficient and reliable framework for automated wine quality assessment and decision support in the wine industry.

Keywords: Wine Quality Assessment; Machine Learning; Feature Selection; Dimensionality Reduction; Hybrid Ensemble Learning

Introduction

Assessment of wine quality, which is a basis of the world-wide wine market, with 2021 exports valued at over 40.7 billion dollars (International Organisation of Vine and Wine, 2022; Jain et al., 2023), Wine quality has typically been guaranteed using a dual system, which consists of expert sensory assessments of the wine alongside physicochemical lab analyses . Yet, this standard approach has high operational costs, is time-consuming, and is ultimately ineffective as human judgment is subjective; furthermore, the non-linearity of relationships among chemical properties and sensory quality makes it complex (Bhardwaj et al., 2022; Kumar et al., 2020).

Application of machine learning (ML) based models for the evaluation of wine quality has gained gaining significant interest owing to their ease of automation and relatively lower testing costs compared to traditional methods. Machine learning algorithms have high efficiency in modeling complex correlations in multivariate datasets, ultimately predicting based on states learned from the corresponding physicochemical compositions (Dahal et al., 2021; Mohana et al., 2023). Indeed, such data dependent ML -based methods not only enhance the prediction accuracy and but also enable winemakers to decide on how to improve the wine production and uphold its quality. However, prediction of wine quality using ML based methods still a challenging as most of the wine quality datasets are high-dimensional with a large number of physicochemical properties and this sets a challenge in terms of computational difficulty in the form of curse of dimensionality (Anowar et al., 2021). Also, the class imbalance affects quality grades, whereby only few quality grades can be observed, as well as the existing redundant or irrelevant feature can drastically reduce the model effectiveness (Chen et al., 2022). In order to overcome these challenges and improve computational efficiency, it is important to include efficient and advanced pre-processing filtering steps (e.g., dimensionality reduction) must be employed.

The methods can be grouped into feature selection methods, which choose these best predictors and discard the less informative ones, and feature transformation methods, which perform a transformation of the initial variables, reducing noise and redundancy (Lee et al., 2021). Feature selection methods, such as filter, wrapper and embedded methods, maintain interpretability since they act on original features, while transformation methods such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) generate novel features which aim to summarize maximal variance or class separation (Zhang et al., 2021).

The present ML based wine quality prediction study involves a systematic embedding of dimensionality reduction methods to ensure simplification of data into ML algorithms. The objectives of the current investigation are as follows. Firstly, the evaluation and comparison of three feature section methods such as Filter-based correlation, Wrapper-based algorithms, and Embedded methods, that are applied on large wine sample datasets with multiple physicochemical features. This study also performed comparative assessment of feature transformation using PCA and LDA methods to ensure for ideal dimensionality reduction while preserving the important information. Further, we compared two ensemble classifiers including Voting and Stacking ensemble classifiers. Finally, implementation of a hybrid ensemble model by integrating both Voting and Stacking ensemble classifiers to maximize prediction accuracy. We hypothesise that proposed ML models help in advancing the existing automated systems for wine quality assessment and provide significant real-time insights in a practical yet interpretable manner.

Literature Review

Over the last two decades, the use of machine learning for wine quality prediction has changed considerably from the early examples of Cortez et al . (2009) who showed that wine quality can be predicted form the chemical data using support vector machines and regression models. Accumulating research data showed that sensory characteristic-based traditional wine quality assessment can be replaced by ML-based automated prediction techniques which are based on wine' physicochemical properties. This foundation has since been broadened significantly, further research followed. Gupta (2018) pushed the field further forward with the application of advanced feature selection approaches along with neural networks and support vector machines in wine quality prediction to emphasize that feature engineering is the key to win in wine quality prediction. The current research demonstrated that cautious selection of physicochemical features not just reduce the complexity but also significantly affect the model performance.

Several previous studies provided strong evidence on the potential inclusion of different ensemble-based algorithms in ML-based prediction research. Dahal et al. (2021) demonstrated the superiority of ensemble methods over individual models. Their study found that gradient boosting regressor model as a remarkably efficient ML model with best performance scores of mean squared error (MSE = 0.3741) and coefficient of determination (R-squared = 0.6057). Jain et al. (2023) extended this comparative study with Decision Trees, Random Forest,

AdaBoost, Gradient Boost, and XGBoost and successfully demonstrated classification accuracies >90% with modern ensemble techniques.

Recently, there is a growing interest in the application of ML based models to overcome challenges linked to wine quality estimation. Bhardwaj et al. (2022) produced a New Zealand Pinot noir wines dataset with limited size, improving model generalization through data augmentation via synthetic data generation using SMOTE. Kumar et al(2020) and Quinta-Frois et al. (2020) states that hyperparameter optimization was an important task, since the results pointed to a vast positive improvement in the prediction accuracy for red wine quality assessment.

2.2 Dimensionality Reduction Techniques-Applications

The wine quality datasets are usually highly dimensional in nature often due to several physicochemical features of wine. Therefore, it is highly desired to include a crucial pre-processing dimensionality reduction steps. Lee et al. (2021) provided a comprehensive theoretical and empirical comparison of dimensionality reduction algorithms, evaluating PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, and t-SNE across various datasets. This study's' results strongly suggest that the ML performance can be substantially influenced by type of dimensionality reduction technique used. Additionally, previous studies showed that the selection of supervised and unsupervised dimensionality reduction methods can significantly affect the downstream ML performance. For instance, Zhang et al. (2021) conducted a detailed comparison of PCA and LDA on wine datasets, revealing that LDA's supervised nature allows it to better preserve class-discriminatory information. LDA particularly optimises for class separability, which makes it possibly more appropriate for classification tasks where retaining quality level distinctions is crucial, whereas PCA maximises general variance in the data.

These techniques have formed the basis of critical insights from practical approaches. According to Verleysen and François (2020), PCA only works out for the data that is lined together but with LDA, it depends on the classes being normally distributed. However, these assumptions should be taken into account while using these techniques (for wine quality data, the nonlinear relationship among the chemical properties and perceived quality can be complex).

It is also important to note that researchers have employed other intricate dimensionality reduction methods such as Kernel methods (e.g., Kernel PCA) and Manifold learning methods

(e.g., t-SNE). Particularly, Kernel PCA has been applied model the wine quality data, while t-SNE has been applied to visualize the data. However, only the Kernel PCA , but not the t-SNE, can be scalable to larger datasets to models the wine quality data.

2.3 Feature Selection Methodologies in Wine Quality Analysis

Undoubtedly, feature selection methods (filter, wrapper and embedded) have become an important factor in the design of efficient ML based models for the wine quality prediction. However, these feature selection methods are still associated with several pros and cons in the context of wine quality assessment.

Filter methods are computationally less expensive and not model-specific, therefore they are usually are applied on high-dimensional datasets. Pawar et al. (2019) also found that a number of filter methods were effective applying for wine quality prediction, after feature selection, the accuracy of 87.33% was acquired using Random Forest. These kinds of techniques, such as mutual information and correlation-based feature selection, evaluate features based on their statistical characteristics without requiring model-fitting. It has been discovered that correlation levels between 0.5 and 0.7 are useful for finding redundant characteristics in wine quality datasets with minimal information loss.

Wrapper methods explore feature interaction methods and are the resulting feature subset tends are better for downstream ML performance. However, these methods are computationally expensive. Borugadda et al. (2023) proved that in different fields the forward and backward selection and recursive feature eliminations were effective. For wine quality prediction, the Boruta algorithm, developed by Kursa and Rudnicki (2010), results particularly promising. Boruta is able to precisely identify the truly significant factors without creating erroneous cor relations by comparing genuine features with shadow features after shuffling. This is especial ly crucial because wine chemical properties are known to interact intricately.

Embedded methods are both efficient and effective since the feature selection is embedded in the model construction. Zhou et al. (2021) managed to use decision tree-based feature importance and the LASSO regularization successfully, in a way that these methods could find significant predictors, but still be interpreted. According to conventional enological principles, various research confirm that alcohol concentration, volatile acidity, sulphates, and citric acid are important characteristics.

2.4 Ensemble Learning and Advanced Modeling Approaches

Growing research evidence showed the potential of ensemble learning as the most robust method for the prediction of wine grades with improved generalization error, which can be achieved by averaging different models. One such ensemble method is Random Forest (RF) which builds many decision tress and combines their outputs. Khilari et al. (2022) that Random Forest generated high predictive accuracy and useful feature ranking comparisons, so it was well-suited to extract drivers for wine quality. This technique is appropriate for the complexity of wine quality data since it can be used to nonlinear connections and feature interactions.

Boosting methods have also gained significant importance in current research. Zeng (2023) independently combined models through stacking ensemble learning, and reported better accuracy (87.22%) over traditional framework models like XGBoost (86.5%), Random Forest (85.12%) and SVM (68.27%). It has been shown that the stacking method, in which a meta learner is created to aggregate base model outcomes, is highly effective in this regard for combining the advantages of several algorithm approaches.

Mohana et al. (2023) have taken the ensemble performance one step ahead by introducing a complex scheme that delivered 98.36% accuracy for red wine quality prediction. Their research focused deeply into choosing and evaluating several base learners and showed that careful ensemble design could result in remarkable improvement in ML performance. Similarly, Supriatna et al. (2023) investigated voting classifiers, evidencing that, in general, soft voting, depending on probability outputs instead of hard classifications, performed better.

Hybrid techniques also emerged from the development of ensemble-based methodologies. Bagging, boosting, and stacking ensembles have been examined in recent works where multi-level ensembles are created by utilising ensembles as foundation learners for higher-level meta-models. Although these intricate structures are computationally costly, they have shown promise as a means of achieving cutting-edge performance..

2.5 Deep Learning and Advanced Techniques

Despite potential application of ML techniques in wine quality prediction, researchers also focus on the implication of DL methods for the same purpose. Di and Yang (2023) applied 1D-Convolutional Neural Networks to represent the non-linear feature interactions, and showed superior accuracy than typical neural networks. However, the majority of wine quality datasets are flat and tiny, which limits the advantages of deep learning.

Zhang (2024) also observed for standard wine quality datasets involving hundreds to thousands of samples, well-tuned ensemble methods usually achieve comparable or even better classification performance as compared to deep learning with less systematic resources and interpretability. These findings have real-time practical application in wine industries as the predictions are accurate.

Advanced methods are emerging continuously to overcome the challenges associated with high-dimension datasets. Chen et al. (2022) investigated the cost-sensitive learning and SMOTE to deal with class imbalance so that the rare-quality ratings can be given enough focus in model training. These methods have proven particularly helpful for datasets with under-represented extreme ratings (extremely high and very low).

2.6 Integration of Domain Knowledge and Interpretability

Scientific studies also highlight the significance of combining enological information with ML approaches. Kulasiri et al. (2024) developed a comprehensive pipeline for Pinot Noir wines that incorporated vineyard data, climatic conditions, and chemical analyses, demonstrating that context-aware modeling could significantly improve prediction accuracy and practical utility.

The understanding of ML model derived predictions is essentially an important factor that needs to be consider for further implementation of these models in wine industries. Török (2023) applied SHAP (Shapley Additive Explanations) values to quantify feature contributions, providing winemakers with actionable insights about which chemical properties most influence quality predictions. Such strategies connects link the gap between black-box predictions and practical winemaking decisions.

Yavas et al. (2023) further advanced interpretable modeling by developing systems that could suggest specific production adjustments. Their ML models help in the detection of controllable variables including acidity levels or alcohol content and enhance the wine quality prediction. This prescriptive capability represents a significant advance in making machine learning practically useful for winemakers.

2.7 Challenges and Perspectives

Despite significant ongoing advancements, there are still some limitations linked to wine quality prediction by ML. The differential wine quality assessment by different experts may result in rating same wine differently, adding ambiguity to training data. Another challenge in developing robust ML models for wine prediction is that regional specific expectations and

taste requirements add additional complexity. Data standardisation and availability are persistent problems. Large-scale research has been made possible by data like the Portuguese Vinho Verde collection, but there are currently few large, mixed, and standardised data sets available, which limits the model's usefulness. Future research will include transfer learning techniques, which may help alleviate data scarcity by transferring knowledge from one type or location of wine to another.

Another new trend is the blending of many modalities. Combining chemical results with sensory ratings, production process statistics, and extensive environmental data is probably going to result in more comprehensive models. Furthermore, real-time quality prediction systems that can support decision-making during wine production rather than just calculating quality at the end of fermentation in an a posteriori operation could be very beneficial to the sector. It is still unclear how to strike a balance between interpretability and model complexity. Simpler and more transparent models often produce more insightful results about what truly drives quality, even while complicated models may result in a somewhat better match. The development of targeted modelling techniques in the future should continue to be directed by trade-offs between such research directions, where a thorough grasp of the precise objectives and purposes as well as the target audience of such frameworks should continue to be a research priority.

1. Methodology

In this section, we provided a complete detail of methods used the prediction of wine quality. The methodology involves different sequential steps, such as initial data collection, pre-processing the data by feature scaling and data transformation, and building of multiple ML algorithms. In order to improve the accuracy of the ML models this study applied dimensionality reduction techniques and ensemble approaches. Figure 1 depicts a complete workflow architecture of the present investigation.

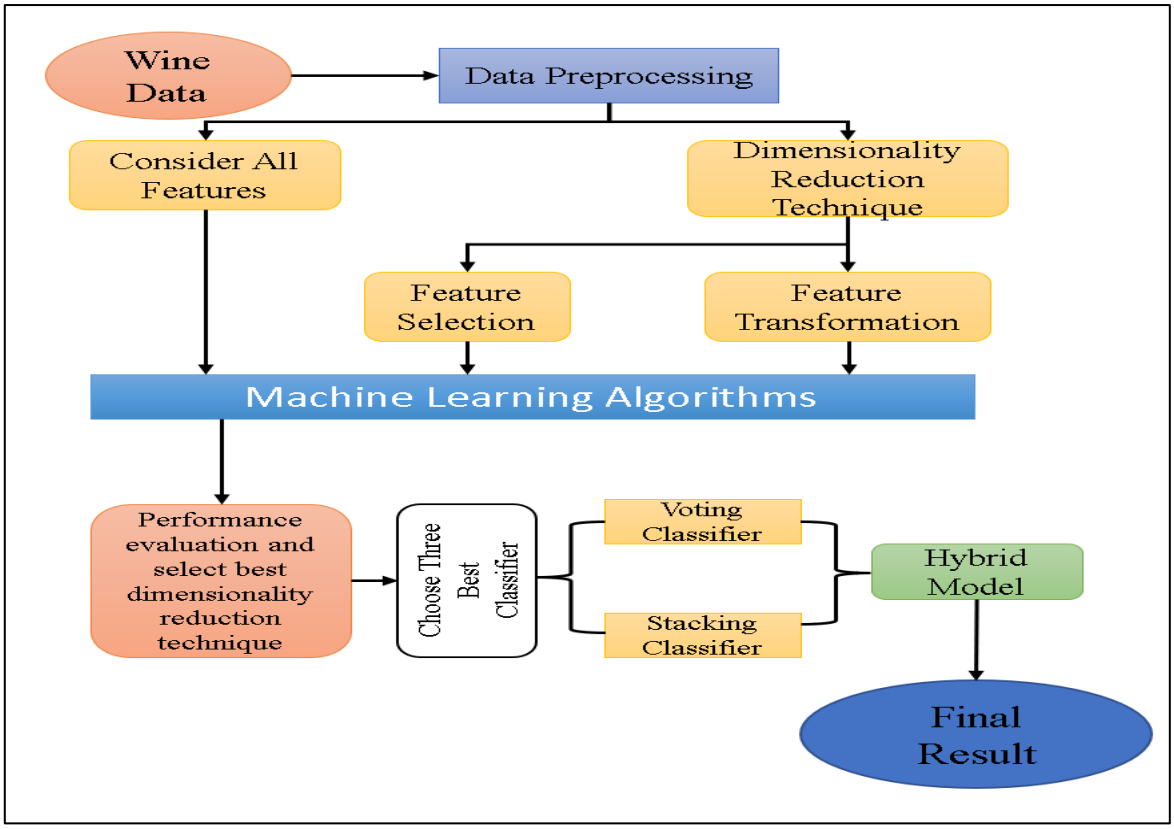


Figure 1. Proposed workflow

3.1) Data Description

The Kaggle database (UCI Machine Learning Repository) was used for the collection of data which contains 6497 instances and 13 features.

Description of data is shown in the following figure 2.

```
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   type                  6497 non-null   object
 1   fixed acidity         6487 non-null   float64
 2   volatile acidity      6489 non-null   float64
 3   citric acid           6494 non-null   float64
 4   residual sugar        6495 non-null   float64
 5   chlorides             6495 non-null   float64
 6   free sulfur dioxide   6497 non-null   float64
 7   total sulfur dioxide  6497 non-null   float64
 8   density               6497 non-null   float64
 9   pH                    6488 non-null   float64
 10  sulphates             6493 non-null   float64
 11  alcohol               6497 non-null   float64
 12  quality               6497 non-null   int64
```

Figure 2. Description of the dataset

3.2) Data Pre-processing

ML relies heavily on preprocessing to obtain the most accurate information. There are several subdivisions of data pre-processing at this stage called Data Cleaning, Data Transformation, Data Normalization (Rissanen, 2001).

In this work, the data is preprocessed before machine learning is used to build the prediction model. Compared to similar datasets, this one requires less time and effort to prepare because it has been thoroughly cleaned and preprocessed, making it easier to access.

3.3) Data Sampling

As shown in below figure, there was a class imbalance problem in the target feature (‘quality), which was overcome by using a data sampling technique known as Synthetic Minority Oversampling Technique (SMOTE) (Fernández, Garcia, Herrera, & Chawla, 2018) distribution is shows in following figure 3.

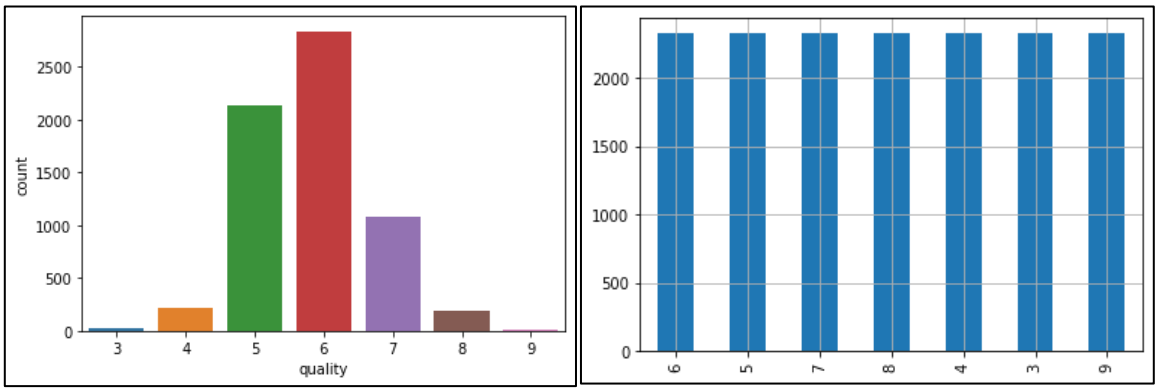


Figure 3. Data Distribution Before and after Data sampling (SMOTE)

As shown in Figure 3, the allocation of data before and after sampling was achieved by SMOTE generates unique examples that are similar to the existing examples in order to produce synthetic examples of the minority class. The process entails choosing an example of a minority class and figuring out who its k nearest neighbor is. One can generate new instances by interpolating between the original example and its k-nearest neighbour in feature space.

3.4) Dimensionality Reduction

To minimize the dataset complexity, this study employed a dimension reduction technique to reduce the number of variables or features, while still retaining the dataset's significant features. This not only eliminates redundant features or irrelevant features and reduce computational complexity but also improves the efficiency of ML algorithms, and redundant features or irrelevant features are eliminated. When dealing with large datasets or when there are more features than observations, this process can be particularly useful (Borugadda, Lakshmi, & Sahoo, 2023). Dimensionality reduction can classified into feature selection and feature transformation techniques.

3.4.1) Feature Selection

Feature selection identifies the most relevant variables from the original dataset by retaining informative features and removing less useful ones. Common approaches include filter, wrapper, and embedded methods. In this study, these techniques are applied to select the most relevant features, as summarized in Table 1.

Table 1. Comparative Analysis of Feature Selection Techniques and Selected Features Across Filter, Wrapper, and Embedded Methods

Feature Selection Technique	Methods	Threshold Value	No. of Features Selected	Selected Feature names
Filter	Univariate	Constant Features	0	All
		Quasi Constant Features	0	All
		Duplicated Features	0	All
	Correlation	0.5	5	'alcohol', 'density', 'total sulfur dioxide', 'citric acid', 'volatile acidity'
		0.6	3	'alcohol', 'volatile acidity', 'total sulfur dioxide'
		0.7	2	'alcohol', 'total sulfur dioxide'
	Forward Feature Selection	_	11	Type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur

Wrapper				dioxide, density, pH, sulphates, alcohol
	Backward Feature Selection	_	7	volatile acidityresidual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol
	Boruta Feature Selection	_	11	Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol
Embedded	Decision Tree	Top 4 Features	4	Density, pH, sulphates, alcohol
		Top 5 Features	5	total sulfur dioxide, density, pH, sulphates, alcohol
	Random Forest	Top 4 Features	4	Chlorides, free sulfur dioxide, total sulfur dioxide, alcohol
		Top 5 Features	5	fixed acidity, chlorides, free sulfur, dioxide total sulfur dioxide, alcohol
	Extra tree	Top 4 features	4	volatile acidity, free sulfur dioxide, total sulfur dioxide, alcohol
		Top 5 Features	5	volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, alcohol

304

305 **A) Filter Method**

306 The filter approach is a feature selection technique that, regardless of a particular algorithm
307 used in the machine learning process, considers the inherent qualities of individual features
308 when determining their importance. A feature set is chosen from a set of selected features based
309 on a certain statistical measure or score that does not consider the interaction between features

or how they depend on one another (Sánchez-Maroño, Alonso-Betanzos, & Tombilla-Sanromán, 2007).

1) Univariate Feature Selection

A univariate technique, a kind of feature selection filter, assesses each feature's relevance by examining its relationship with the target variable, which determines whether each item merits consideration in terms of its relevance. As a result of this method, each feature is treated independently, and its importance is assessed without considering the interactions between the features (Jović, Brkić, & Bogunović, 2015).

2) Correlation feature selection

Correlation features, such as Pearson correlation coefficients, are selected based on a correlation coefficient. There are a range of -1 to 1 for this coefficient. Correlations close to 1 indicate strong positive correlates, negative correlations close to -1 indicate strong negative correlations, and strong correlations close to 0 indicate weak or no correlations (Blessie & Karthikeyan, 2012) which is showed in below figure 4.

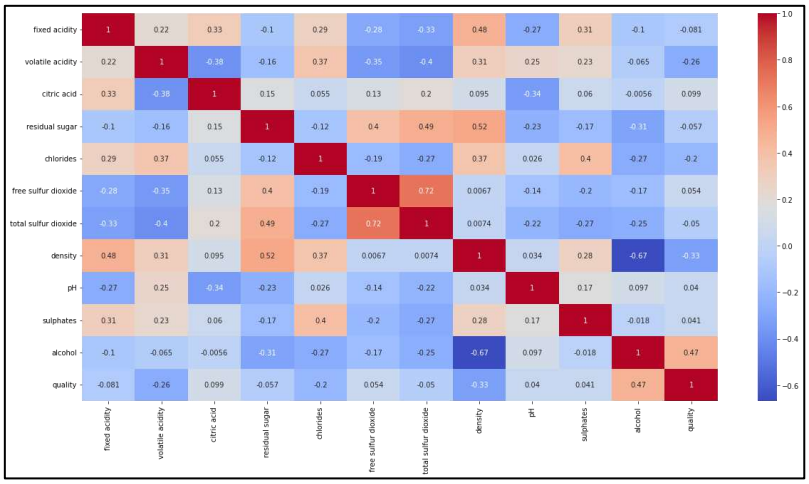


Figure 3. Correlation analysis of data

B) Wrapper Method

The Wrapper technique, which use various feature subsets to assess a machine learning model's performance, is used to analyse its performance. This approach views feature selection as an optimisation problem, where features are chosen according to how they affect model performance.

1) Forward feature selection

An approach called forward feature selection begins with an empty set of features and adds each feature iteratively based on its performance. To begin, each feature of the machine learning model is evaluated individually and the one which performs the best is selected (Ververidis & Kotropoulos, 2005).

The graph of forward feature selection is shown in figure 4.

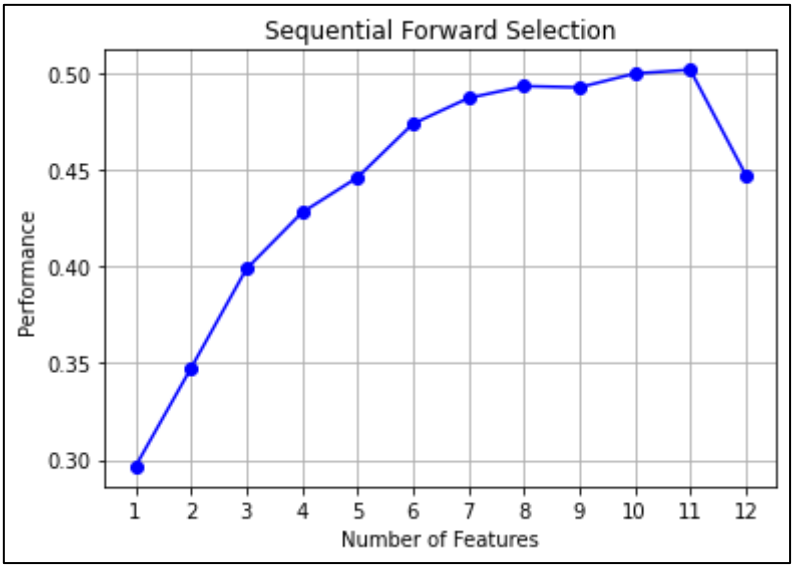


Figure 4. Forward Feature Selection

2) Backward Feature Selection

It starts with a collection of all the features available and iteratively eliminates one feature at a time, then evaluating the model's performance each time. Backward feature selection is a wrapper method for feature selection. After reaching a specified number of features or when further removals are no longer beneficial, the flow continues until a specified stopping criterion is reached (Zhou et al., 2021).

The following figure 5 shows the process of backward feature elimination

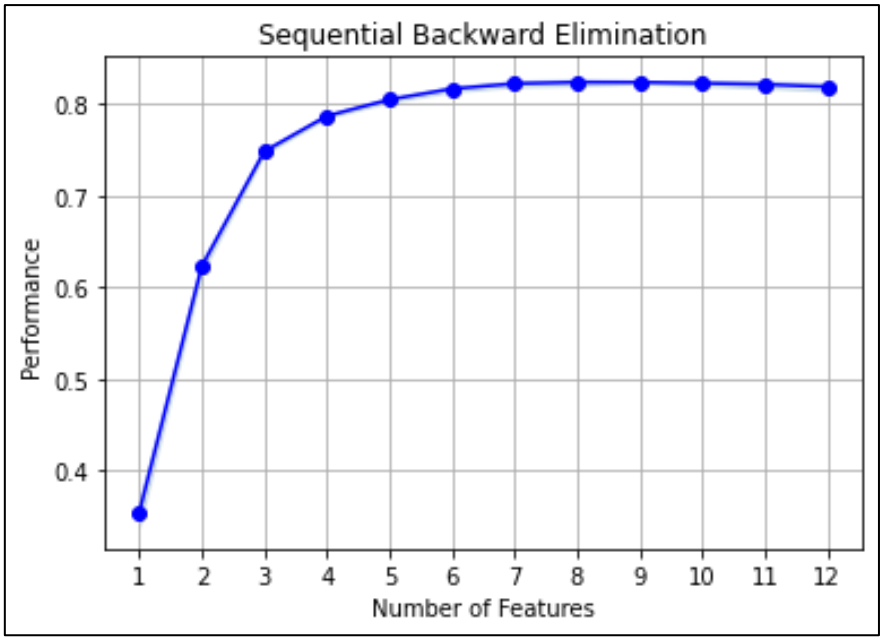


Figure 5. Backward Feature Elimination

3) Boruta Feature Selection

As one of the most important methods used by the wrapper method category, Boruta is a feature selection algorithm that is designed to identify the most relevant features in a dataset by comparing their importance with the importance of random noise variables within the dataset (Kursa & Rudnicki, 2010). Boruta algorithm iterations are shown in Figure 6.

```
BorutaPy finished running.

Iteration:      9 / 100
Confirmed:      11
Tentative:      0
Rejected:       1
```

Figure 6. Boruta Algorithm

C) Embedded Method

Embedded methods directly incorporate feature selection into training process, thereby minimize the dimension of ML models. Unlike separate selection approaches, embedded methods take into account the significance of the features during the model's training. This is one of their unique characteristics. Using algorithms such as Decision Tree, Random Forest, and Extra Tree, the feature importance technique, an embedded method calculated the most important features (Figure 7) that are affected to the target variable.

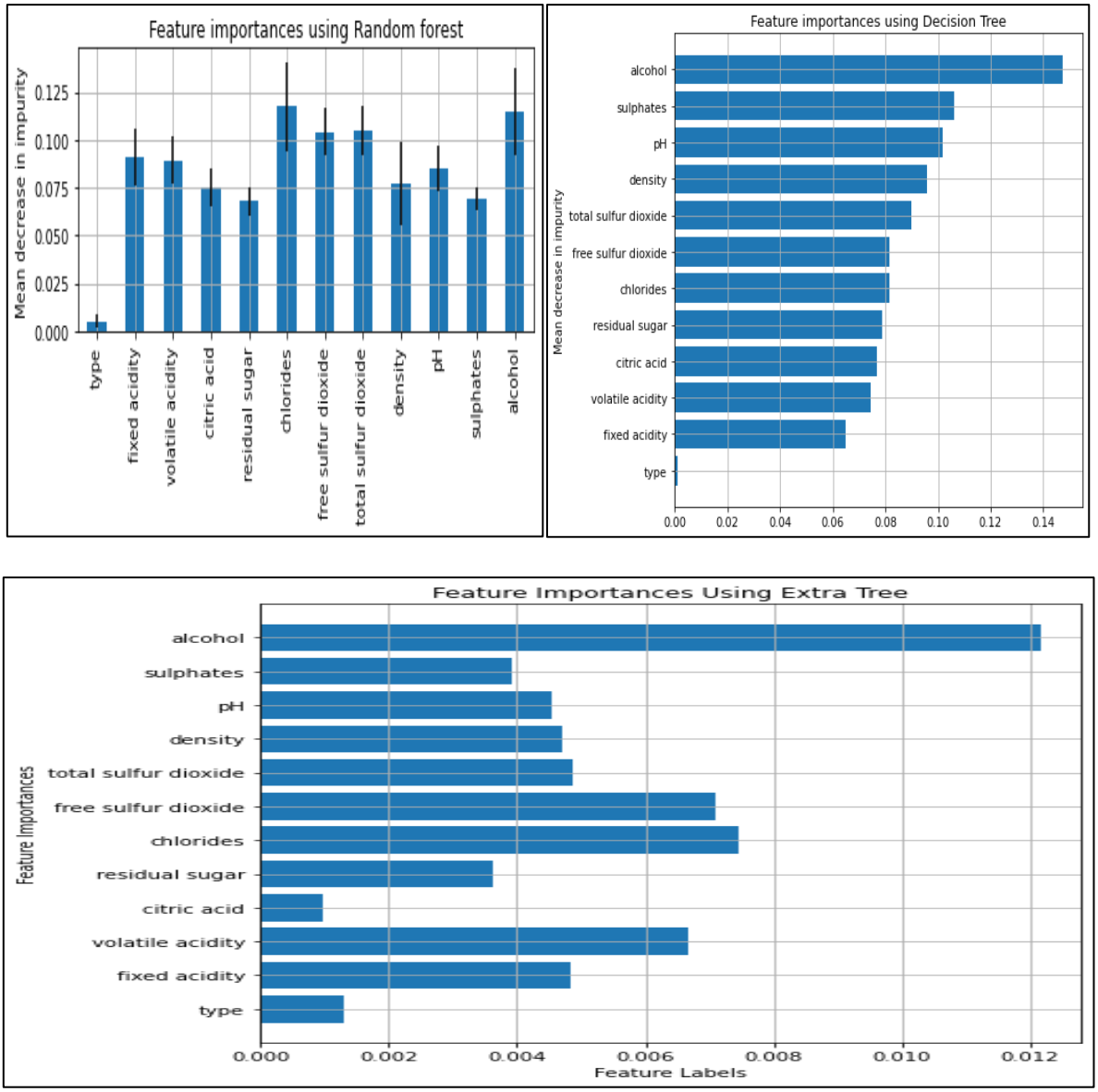


Figure 7. Embedded Method Using DTC, ETC and RFC

3.4.2) Feature Transformation Technique

Feature transformation techniques aim to modify or project the original features into a new space that better captures the underlying patterns and relationships in the data. The details of the applied transformation methods and the corresponding number of components are presented in Table 2.

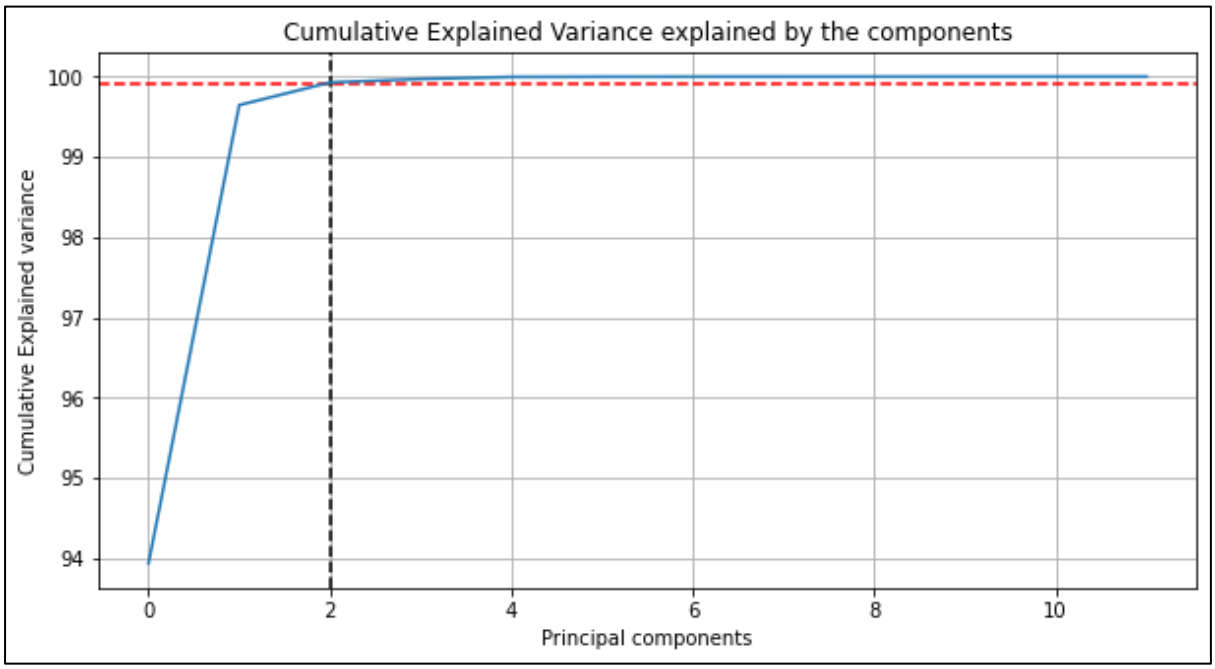
Table 2. Summary of Feature Transformation Techniques and Selected Components Based on Explained Variance

Feature Transformation Technique	Explained Variance Ratio	No. of components
Principal Component Analysis (PCA)	<99	3
	<95	4
Linear Discriminant Analysis (LDA)	No. of classes – 1 (7 - 1)	6

373

374 **I) Principal Component Analysis (PCA)**

375 PCA is a widely used dimensionality reduction technique in data analysis and ML mainly to
376 reduce dimensionality and extract features. Data that is high-dimensional are transformed into
377 lower-dimensional representations using PCA, preserving the most important information
378 within the dataset. The PCA graph with cumulative frequency explained by components is
379 displayed in Figure 8.



380

381 Figure 8. Cumulative explained variance by components

382

383 **II) Linear Discriminant Analysis (LDA)**

384 LDA is frequently used to reduce the dimensionality and extract features from the data in ML
385 and pattern recognition tasks. By using LDA, the goal is to create a representation of the data

that minimises the number of dimensions and maximizes the ability to separate classes and discriminate them (Yu & Yang, 2001). n LDA, the number of components equals the number of classes in the target variable minus one, as shown in Figure 9 below.

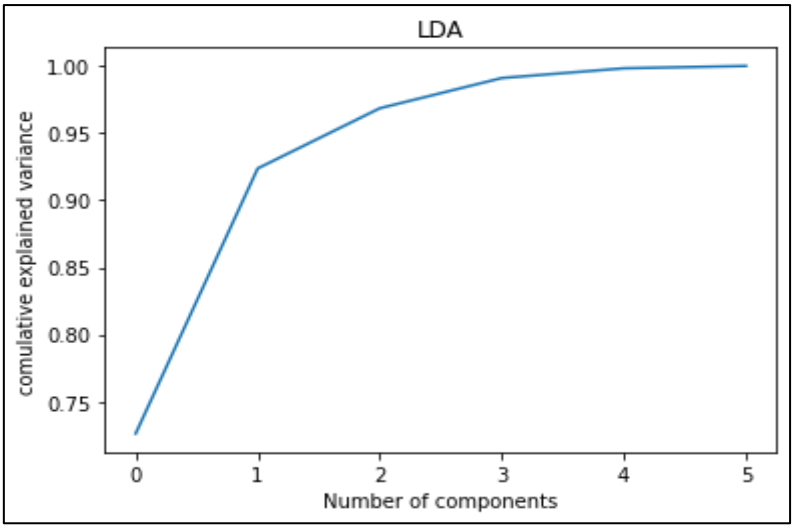


Figure 9. LDA

3.5) Performance Measures

The study employed a hybrid model (Voting + Stacking) to predict wine quality using RFC, KNN, LR, SVC, GB, XGboost, Extra Tree Classifier, DTC, Gradient Boosting, Adaboost Classifier, stacking ensemble, and voting ensemble algorithms. The following subsection briefly discusses these matrices before using the ML algorithms: the confusion matrix, the receiver operating characteristics (ROC), the area under the curve (AUC) and the precision-recall curve (Hossin & Sulaiman, 2015).

3.5.1) Confusion Matrix

A table pattern called a confusion matrix (Figure 10) can be utilised to predict how a classification problem might develop and how a prediction might turn out. While the anticipated values are determined by the model based on its predictions, the actual values are determined by the given data.

		PREDICTED CLASSES				
ACTUAL CLASSES	CLASSES	C_1	C_2	...	Cn	
	C_1	TP_1				AVC_1
	C_2		TP_2			AVC_2
	...			...		...
	C_n				TP_n	AVC_n
		PVC_1	PVC_2	...	PVC_n	

Figure 10. The Confusion Matrix

Following are some examples of the confusion matrix:

True Positive: The model's prediction was matched by the actual value, which was yes.

True Negative: Both the actual and projected values were negative.

False Positive: The actual value was Yes and model prediction was No this is called Type1 error.

False Negative: The actual value was No and model prediction was Yes this is called Type2 error.

There are several Evaluation Metrics that are used for evaluation machine learning algorithms:

A) Accuracy Score

The classification accuracy score is determined by the proportion in which the prediction makes an accurate prediction based on all input data points.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

B) Precision

The precision ratio is the sum of the predicted positives divided by the number of true positives.

$$Precision = \frac{TP}{TP + FP}$$

C) Recall

The proportion of True Positives to all other Positives is known as the recall rate.

$$TRP/Recall/Sensitivity = \frac{TP}{TP + FN}$$

D) F1_score

Precision and recall together make up the F1 score. Only when recall and precision are high is the F1 score high.

$$\text{F1Score} = 2*\frac{Precision*Recall}{Precision+Recall}$$

3.6) ML Algorithms

A) Logistic Regression (LR)

It is a type of statistical modelling that estimates the likelihood that an event will occur, such as voting or not voting, based on a set of independent variables (Koshiga, Borugadda, & Shaprapawad, 2023).

$$LR = \frac{1}{1 + e^{-(Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots\cdots\beta_nX_n)}}$$

B) K-Nearest Neighbor (KNN)

KNN is a lazy learner among all ML algorithms. KNN classifier is different from a linear, supervised learning classifier in that it uses proximity to make decisions about a data point's grouping based on the classifications or predictions that it makes (Zhang et al., 2006).

$$Euclidean = \sqrt{\sum_{i=0}^{n}(X_i - Y_i)^2}$$

C) Random Forest

To increase prediction accuracy, the RF classifier aggregates the predictions from many decision trees. RF uses bagging to combine multiple decision trees to make predictions accurate (Parmar, Katariya, & Patel, 2019).

447

448 **D) Support Vector Machine (SVM)**

449 Using SVC, the most significant margin between two classes is located by finding the
450 hyperplane. In many cases, it is referred to as a maximum-margin classifier (Anguita et al.,
451 2005).

452 **E) Extra Tree Classifier (ETC)**

453 The extra trees are used to create decision trees as part of an ensemble-supervised ML
454 method. A decision tree is constructed over the entire dataset using the Extra Trees algorithms
455 (Peng et al., 2021).

456 **F) Gradient Boosting (GB)**

457 By combining weak models into one the GB Classifier generates a robust classifier. As each
458 model is trained, residual errors from the previous model are addressed. Based on the
459 weighted contributions from each model, the final prediction is made (Natekin & Knoll,
460 2013).

461 **G) Decision Tree**

462 Decision trees are machine learning techniques that can be used with numbers and categories.
463 They are designed for forecasting a target variable using basic decision-making principles
464 (Quinlan, 1996).

465 Decision trees can easily and quickly be constructed by using recursive top-down DAC
466 (divide and conquer) approach. Additionally, irrelevant samples are also removed from D by
467 using tree pruning. The entropy of a system is calculated using the following equation:

468 $$Entropy = \sum_{i=0}^{n} P_{ij} log_2 P_{ij}$$

469 **H) Voting Ensemble**

470 Voting classifiers can be divided into hard voting classifiers and soft voting classifiers. It
471 combines the predictions from several machine learning models for output using the ensemble
472 approach (Mahabub, 2020).

Hard: The final class prediction among the ML base models is decided by a majority vote using the ML base model's most frequent prediction.

Soft: The average probability derived from all of the ML base model's predictions is used to select the final class.

The workflow for voting model shows in figure 11.

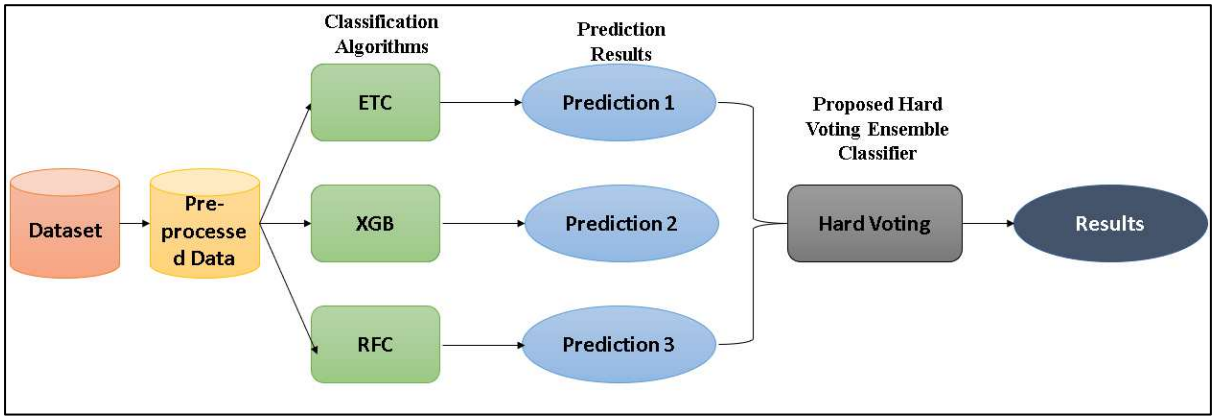


Figure 11. Workflow of Voting Ensemble

I) Stacking ensemble

The process of merging the predictions of many individual models, also known as base models, into a final prediction that is based on the predictions of multiple individual models is known as stacking in ensemble learning. The meta-model, which learns to combine these forecasts to generate the final prediction, and the base models, which are trained on the same set of data, are the two types of stacking algorithms. Or the aggregator, which learns to combine the predictions made by the layers to make the final prediction (Wang et al., 2020). The following figure 12 shows the workflow for stacking ensemble model.

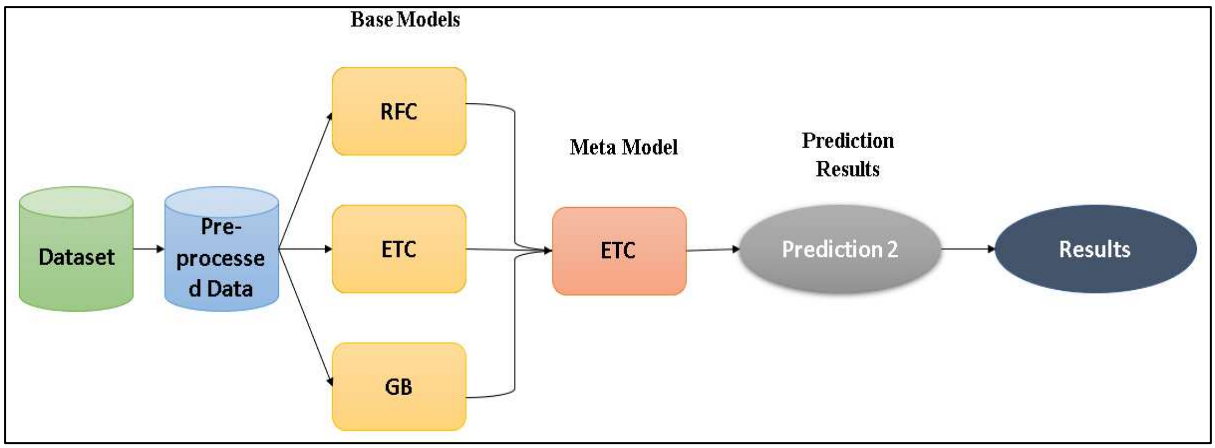


Figure 12. Workflow of Stacking Ensemble Model

J) Stacking + Voting

By utilising the advantages of both strategies, it has been demonstrated that combining stacking and voting is a successful method for developing hybrid models. The final prediction in this hybrid ensemble is obtained by integrating the predictions from each base model through voting after the several base models are sequentially trained via stacking. The suggested hybrid model approach is depicted in figure 13.

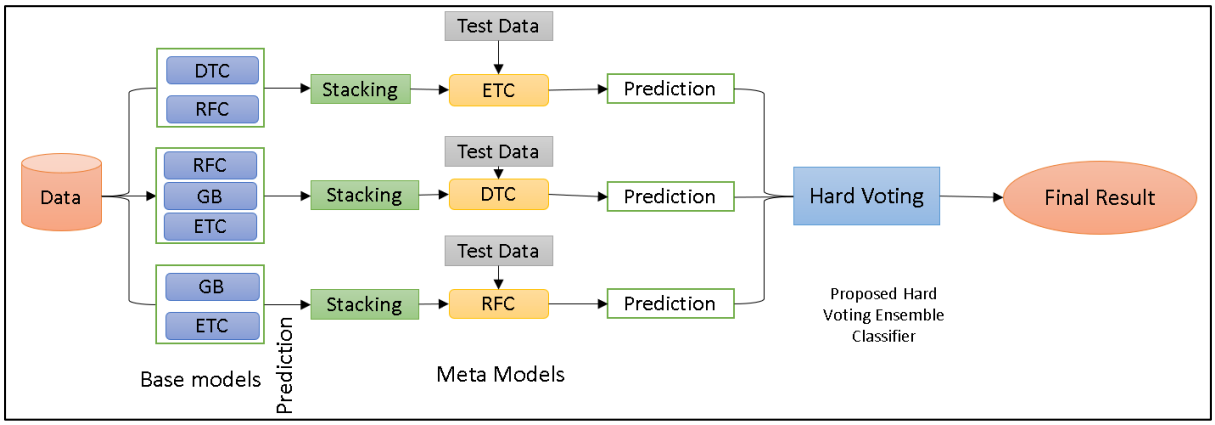


Figure 13. Proposed Hybrid Model Workflow

Experimental data analysis and ML algorithms

4.1) Experimental setup and hyperparameter tuning

The current experiment developed machine learning techniques in Python 3.16 with Jupiter notebook support. Because hyperparameters can control how the model behaves during the learning process, the model learns better when hyperparameters are set before the model is permitted to start learning. In order to improve a model's performance, hyperparameters are set before the model starts learning. To obtain the ideal values, a number of hyperparameters may need to be adjusted. Each hyperparameter is given a range of values, to which the GridsearchCV method is used with a CV of 10. The results are displayed in table 3.

4.1) Hyperparameter Tuning

Hyperparameters are predefined settings that are specified before the training process begins and are not learned directly from the data. They play an important role in controlling how a model learns and performs. Common examples include the number of layers, the number of nodes in each layer, and the strength of regularization. The optimized hyperparameter settings for machine learning and ensemble models considered in this study are presented in Table 3. In practice, hyperparameter tuning involves exploring different combinations of these parameters and evaluating model performance on a validation set to identify the most effective configuration for a given task, such as classification or regression. Techniques such as grid search, random search, and Bayesian optimization are widely used to perform this optimization efficiently.

Table 3. Optimal Hyperparameter Settings for Machine Learning and Ensemble Models

Sr.no.	Model	Hyper parameters	Optimal Values
1	Logistic Regression	Penalty	L1
		tol	1e-05
		C	1
		fit_intercept	True
		solver	liblinear
		multi_class	auto
2	KNN	n_neighbors	5
		weights	Uniform
		algorithm	Auto
		p	4
		n_jobs	-1
3	Random Forest Classifier	n_estimators	250
		criterion	Entropy
		max_features	50
		n_jobs	-1
4	Support Vector Classifier	C	9
		Kernel	Rbf
		Degree	2
		Gamma	Auto
5	Extra Tree Classifier	Criterion	Gini

		Max_features Min_sample_leaf N_estimators N_jobs	None 15 150 -1
6	Decision Tree classifier	Criterion Splitter Min_sample_split	Log_loss Random 5
7	Gradient Boosting Classifier	Loss Learning_rate N_estimators Criterion Max_depth Max_features	Log_loss 001 200 Squared_mse 15 auto
8	Voting Ensemble Classifier	Estimators voting	(RFC, Extra Tree Classifier, Gradient Boosting) Hard
9	Stacking Ensemble classifier	Base estimators Final Estimator CV	RFC, ETC, DTC GBC 10
10	Voting with stacking	Estimators voting	Stacking 1, stacking 2 hard

5) Results and discussion

This section presents the comparative performance analysis of different machine learning, ensemble, and hybrid models for wine quality prediction after applying various dimensionality reduction techniques. Logistic Regression (LR), K-Nearest Neighbour (KNN), Random Forest Classifier (RFC), Support Vector Classifier (SVC), Extra Trees Classifier (ETC), Decision Tree Classifier (DTC), Gradient Boosting Classifier (GBC), Voting Ensemble, Stacking Ensemble, and the proposed Hybrid Ensemble model combining Voting and Stacking techniques are among the models assessed in this study.

The performance of these models was evaluated using accuracy, precision, recall, and F1-score. Different dimensionality reduction strategies were investigated, including filter-based, wrapper-based, embedded feature selection methods, and feature transformation techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

5.1 Filter-Based Feature Selection Results

Table 4 presents the comparative analysis of filter-based feature selection techniques using univariate and correlation-based methods. Among the univariate methods, the Stacking Ensemble model achieved the highest accuracy of 86.92%, while the proposed Hybrid model achieved an accuracy of 86.02%. The ETC model also demonstrated strong performance with an accuracy of 85.11%.

For correlation-based feature selection, different threshold values of 0.5, 0.6, and 0.7 were evaluated. The results indicate that selecting highly correlated and informative features significantly improved model performance. Among all the evaluated configurations, the correlation threshold of 0.6 produced the best overall results. The proposed Hybrid Ensemble model achieved the highest accuracy of 89.08%, with corresponding precision, recall, and F1-score values of 89.11%, 88.91%, and 88.93%, respectively. Similarly, the Hybrid model achieved accuracy of 87.98% and 88.83% for correlation thresholds of 0.5 and 0.7, respectively.

The results demonstrate that correlation-based feature selection effectively removes redundant and irrelevant features while preserving the most informative attributes related to wine quality prediction. Figures 14 and 15 further illustrate the comparative accuracy analysis of the univariate and filter-based feature selection methods across different machine learning models. The ensemble and hybrid models consistently outperformed traditional classifiers such as LR and SVC, indicating the effectiveness of ensemble learning in handling complex nonlinear relationships within the wine dataset.

Table 4. Comparative Analysis of Filter-Based Feature Selection Methods for Wine Quality Prediction Using Machine Learning Models

Method	Threshold Value	Model	Accuracy	Precision	Recall	F1-Score
	Quasi constant	LR	31.83	29.71	31.97	29.31
		KNN	74.69	71.96	74.14	72.36

Univariate		RFC	84.25	83.35	84.31	83.67
		SVC	36.25	32.42	35.72	30.14
		ETC	85.11	85.46	86.15	85.60
		DTC	75.83	75.38	75.39	75.36
		GBC	73.01	71.75	72.35	71.84
		Voting	84.05	83.23	83.59	83.23
		Stacking	86.92	87.18	86.94	87.03
		Hybrid (Voting + Stacking)	86.02	86.36	86.07	86.19
	Constant	LR	31.83	29.71	31.97	29.31
		KNN	74.69	71.96	74.14	72.36
		RFC	84.25	83.35	84.31	83.67
		SVC	36.25	32.42	35.72	30.14
		ETC	85.11	85.46	86.15	85.60
		DTC	75.83	75.38	75.39	75.36
		GBC	73.01	71.75	72.35	71.84
		Voting	84.05	83.23	83.59	83.23
		Stacking	86.92	87.18	86.94	87.03
		Hybrid (Voting + Stacking)	86.02	86.36	86.07	86.19
	Duplicated Features	LR	31.83	29.71	31.97	29.31
		KNN	74.69	71.96	74.14	72.36
		RFC	84.25	83.35	84.31	83.67
		SVC	36.25	32.42	35.72	30.14
		ETC	85.11	85.46	86.15	85.60
		DTC	75.83	75.38	75.39	75.36
		GBC	73.01	71.75	72.35	71.84
		Voting	84.05	83.23	83.59	83.23
		Stacking	86.92	87.18	86.94	87.03
		Hybrid	86.02	86.36	86.07	86.19

		(Voting + Stacking)				
Correlation	0.5	LR	31.53	25.57	31.40	25.16
		KNN	63.07	60.47	62.27	61.01
		RFC	80.12	78.88	79.63	79.16
		SVC	32.88	25.78	33.55	27.75
		ETC	79.26	78.05	78.81	78.38
		DTC	67.98	66.85	67.33	67.02
		GBC	61.72	60.00	61.34	60.37
		Voting	79.02	77.62	78.48	77.97
		Stacking	79.63	79.12	79.17	79.13
		Hybrid (Voting + Stacking)	87.98	88.24	87.79	87.82
	0.6	LR	34.11	29.77	33.65	27.44
		KNN	63.80	61.15	62.96	61.63
		RFC	83.19	82.04	82.65	82.14
		SVC	36.56	40.38	37.57	33.10
		ETC	81.79	84.05	84.48	84.10
		DTC	72.52	71.40	71.95	71.64
		GBC	68.71	67.99	68.20	67.97
		Voting	84.05	83.23	83.59	83.23
		Stacking	86.01	85.90	85.61	85.75
		Hybrid (Voting + Stacking)	**89.08**	**89.11**	**88.91**	**88.93**
	0.7	LR	37.79	31.59	37.00	31.36
		KNN	64.54	62.19	63.71	62.52
		RFC	84.79	83.78	84.33	83.91
		SVC	36.56	40.38	37.57	33.10
		ETC	85.89	85.14	85.58	85.22
		DTC	71.41	71.01	70.92	70.87
		GBC	71.29	70.17	70.70	70.35

		Voting	84.29	83.27	83.84	83.49
		Stacking	85.52	85.70	85.13	85.37
		Hybrid (Voting + Stacking)	88.83	88.56	88.93	88.66

The accuracy results for various machine learning models employing the univariate feature selection strategy are compared in Figure 14. The results indicate that ensemble models, particularly the Stacking Ensemble and Hybrid model, achieved better prediction performance compared with traditional classifiers such as LR and SVC.

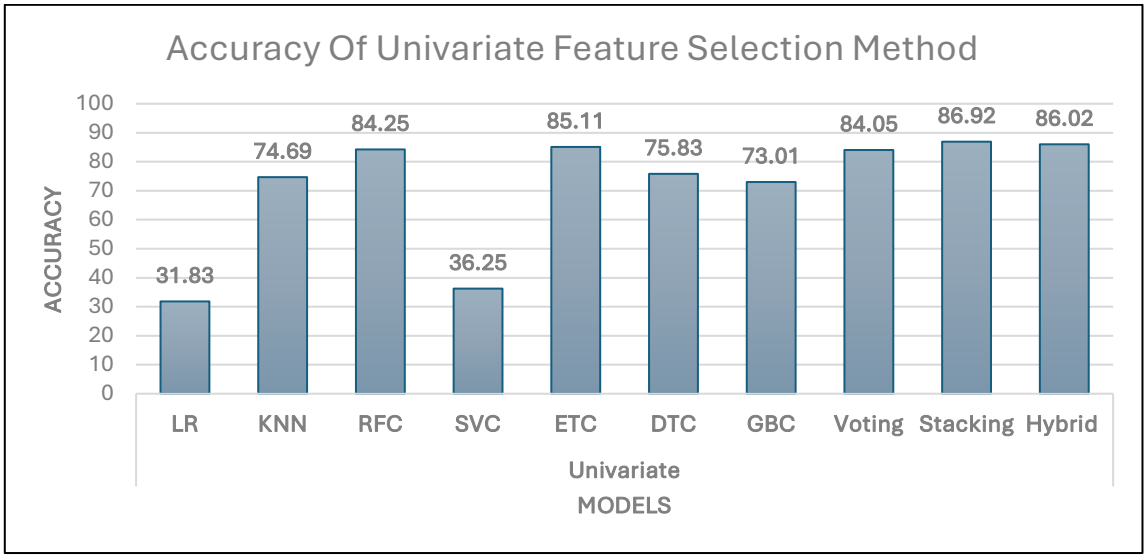


Figure 14. Comparison of Accuracy for different ML models of Univariate methods

Figure 15 presents the comparison of accuracy scores for different machine learning models using filter-based feature selection methods. The results show that the correlation-based feature selection technique, particularly with a threshold value of 0.6, achieved the highest prediction accuracy using the proposed Hybrid Ensemble model, outperforming the traditional machine learning classifiers.

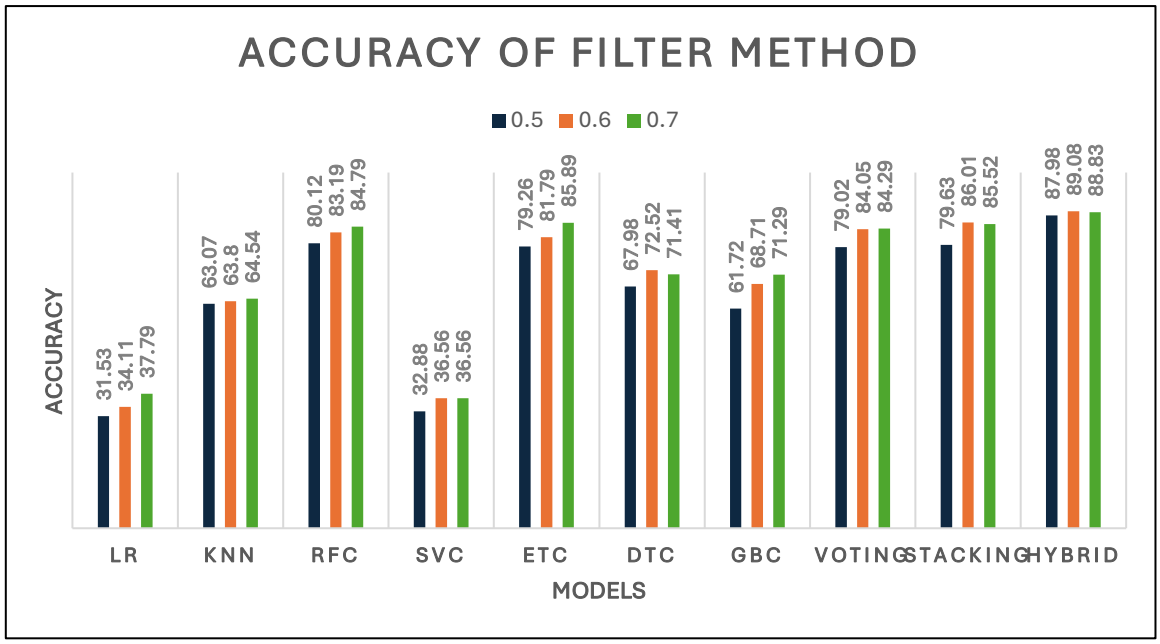


Figure 15. Comparison of Accuracy for different ML models of Filter methods

5.2 Wrapper-Based Feature Selection Results

Table 5 presents the performance analysis of wrapper-based feature selection techniques, including Forward Feature Selection, Backward Feature Elimination, and Boruta Feature Selection.

Among the wrapper methods, Boruta Feature Selection produced the best classification results. The ETC model achieved the highest accuracy of 87.62%, followed by the Voting Ensemble model with 86.99% accuracy and the RFC model with 86.26% accuracy. The Hybrid Ensemble model also demonstrated competitive performance with an accuracy of 85.52%.

Backward Feature Elimination showed slightly better performance than Forward Feature Selection in most cases. The ETC classifier achieved an accuracy of 86.99% using Backward Feature Elimination, whereas the highest accuracy under Forward Feature Selection was 85.89% using the Hybrid model.

The Boruta algorithm effectively identified the most relevant physicochemical features influencing wine quality, leading to improved classification performance. Figure 16 illustrates the comparison of classification accuracy across different wrapper-based feature selection methods and machine learning models. Overall, the results confirm that wrapper-based

methods can improve predictive performance by selecting optimal feature subsets based on model-specific learning behavior.

Table 5. Comparative Performance Analysis of Wrapper-Based Feature Selection Techniques Using Machine Learning and Ensemble Models

Method	Model	Accuracy	Precision	Recall	F1 Score
Forward Feature Selection	**LR**	40.49	38.29	36.06	39.83
	KNN	71.17	68.89	69.04	70.66
	RFC	85.28	84.23	84.43	84.96
	SVC	54.85	44.88	43.26	45.02
	ETC	85.52	84.57	84.79	85.18
	DTC	73.37	72.10	72.36	72.88
	GBC	71.66	70.54	70.68	71.27
	Voting	83.68	83.06	83.18	83.44
	Stacking	83.19	82.04	82.14	82.65
	Hybrid (Voting + Stacking)	85.89	85.14	85.22	85.58
Backward Feature Elimination	**LR**	32.88	29.07	29.97	32.75
	KNN	73.01	70.11	70.57	72.60
	RFC	85.77	84.91	85.04	85.48
	SVC	35.46	33.55	29.77	34.75
	ETC	86.99	86.29	86.37	86.77
	DTC	73.87	72.83	73.07	73.43
	GBC	73.01	72.12	72.19	72.41
	Voting	83.68	83.06	83.18	83.44
	Stacking	84.29	83.27	83.49	83.84
	Hybrid (Voting + Stacking)	85.52	84.23	84.43	84.96
Boruta Feature Selection	**LR**	34.11	30.30	30.71	33.94
	KNN	76.44	73.16	73.78	75.74
	RFC	86.26	85.18	85.37	85.85

	SVC	36.93	36.38	31.90	36.34
	ETC	87.62	86.93	87.25	87.06
	DTC	74.72	73.51	73.74	74.11
	GBC	75.09	74.26	74.49	74.36
	Voting	86.99	86.29	86.37	86.77
	Stacking	83.68	83.06	83.18	83.44
	Hybrid (Voting + Stacking)	85.52	84.57	84.79	85.18

Figure 16 illustrates the comparison of accuracy scores for different machine learning models using wrapper-based feature selection techniques. The results indicate that the Boruta feature selection method achieved better classification performance compared with Forward Feature Selection and Backward Feature Elimination, with the ETC and ensemble models producing the highest prediction accuracies.

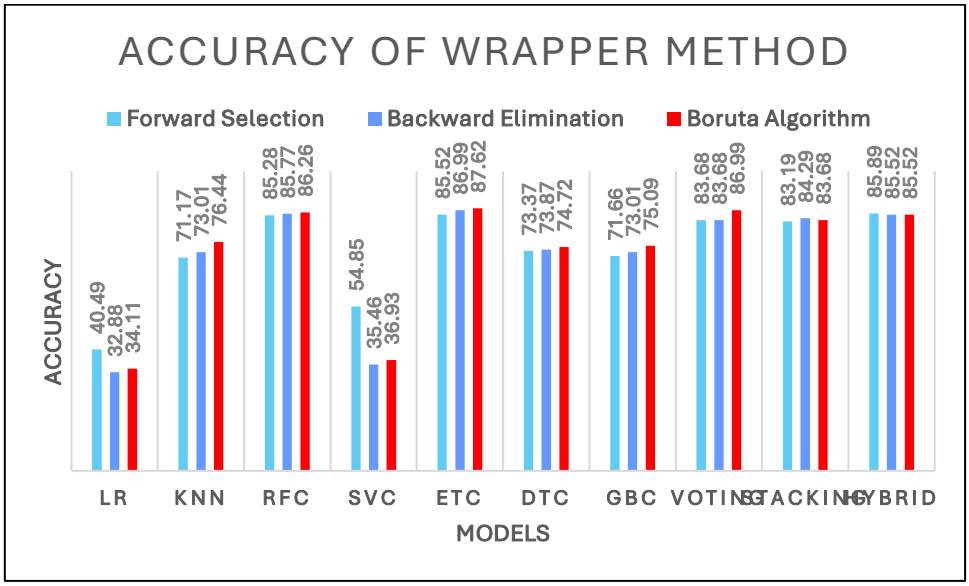


Figure 16. Comparison of Accuracy for different ML models of Wrapper methods

5.3 Embedded Feature Selection Results

Table 6 summarizes the performance of embedded feature selection techniques using feature importance derived from DTC, RFC, and ETC models.

The embedded methods showed moderate performance compared with filter and wrapper methods. Among all embedded approaches, the ETC (Top 5 Features) combined with the Stacking Ensemble achieved the highest accuracy of 84.05%, while the Hybrid model achieved 82.49% accuracy. Similarly, RFC (Top 5 Features) achieved an accuracy of 80.86% using the Hybrid model.

The results indicate that selecting only a limited number of important features reduces computational complexity; however, excessive reduction may also lead to information loss, thereby affecting prediction performance. Nevertheless, embedded methods demonstrated stable performance while maintaining model interpretability and computational efficiency.

Traditional classifiers such as LR and SVC consistently produced lower accuracies across all embedded feature selection methods, indicating their limited ability to capture complex nonlinear interactions among wine physicochemical properties.

Table 6. Comparative Performance Analysis of Embedded Method Feature Selection Techniques Using Machine Learning and Ensemble Models

Method	Model	Accuracy	Precision	Recall	F1 Score
DTC (Top 4 Features)	**LR**	31.29	29.49	29.43	33.01
	KNN	54.23	53.77	54.05	55.21
	RFC	70.55	69.80	70.24	71.22
	SVC	31.17	29.13	27.05	31.89
	ETC	71.41	77.6 2	72.13	71.00
	DTC	62.82	62.71	63.23	63.95
	GBC	52.52	50.75	51.33	53.46
	Voting	69.57	68.84	69.25	70.27
	Stacking	71.04	70.63	71.02	71.72
	Hybrid (Voting + Stacking)	71.04	70.58	71.59	70.82
DTC (Top 5 Features)	**LR**	26.87	23.55	23.90	27.00
	KNN	56.69	55.36	55.77	57.39
	RFC	78.04	76.82	77.19	78.34
	SVC	29.33	36.36	24.19	30.16

	ETC	79.20	78.83	80.14	79.20
	DTC	69.20	69.03	69.44	70.01
	GBC	60.49	59.89	60.12	61.41
	Voting	77.91	76.83	77.24	78.18
	Stacking	77.91	78.01	78.07	78.28
	Hybrid (Voting + Stacking)	77.30	77.14	77.16	77.56
ETC (Top 4 Features)	**LR**	27.24	26.65	24.45	27.53
	KNN	67.48	64.80	65.69	68.09
	RFC	80.61	79.97	80.28	80.95
	SVC	35.34	28.89	35.27	28.76
	ETC	82.21	81.64	82.59	82.00
	DTC	71.78	71.61	71.79	72.23
	GBC	65.15	64.81	65.07	65.76
	Voting	80.37	79.84	80.15	80.68
	Stacking	83.31	83.26	83.43	83.64
	Hybrid (Voting + Stacking)	81.35	81.58	81.59	81.69
ETC (Top 5 Features)	**LR**	27.24	26.65	24.45	27.53
	KNN	67.48	64.80	65.69	68.09
	RFC	80.61	79.97	80.28	80.95
	SVC	35.34	28.89	28.76	35.27
	ETC	82.45	81.94	82.85	82.21
	DTC	71.78	71.67	71.88	72.24
	GBC	65.28	64.94	65.20	65.88
	Voting	80.74	80.11	80.40	81.06
	Stacking	84.05	84.14	84.24	84.37
	Hybrid (Voting + Stacking)	82.49	81.96	82.26	82.13

RFC (Top 4 Features)	**LR**	25.03	19.91	25.04	21.32
	KNN	67.12	64.67	65.49	67.71
	RFC	78.40	77.82	78.07	78.73
	SVC	35.09	28.37	35.01	28.66
	ETC	80.10	79.28	82.91	79.60
	DTC	71.04	71.27	71.46	71.31
	GBC	60.25	59.65	61.04	59.88
	Voting	80.02	79.24	80.27	79.50
	Stacking	79.26	79.04	79.58	79.22
	Hybrid (Voting + Stacking)	80.12	80.49	80.45	80.32
RFC (Top 5 Features)	**LR**	27.73	25.93	27.28	24.72
	KNN	71.29	68.57	71.82	69.35
	RFC	79.39	78.70	79.79	78.96
	SVC	34.97	27.52	34.86	28.32
	ETC	82.21	81.61	82.53	81.84
	DTC	71.90	71.71	72.40	71.99
	GBC	62.94	62.46	63.72	62.57
	Voting	80.61	79.89	80.87	80.09
	Stacking	83.07	83.18	83.30	83.23
	Hybrid (Voting + Stacking)	80.86	81.36	81.11	81.15

Figures 17–19 present the comparative accuracy analysis of embedded feature selection methods using Decision Tree, Extra Trees, and Random Forest feature importance techniques, respectively. The results demonstrate that ensemble and hybrid models consistently achieved better prediction performance compared with traditional machine learning models. Among the embedded methods, the Extra Trees-based feature selection approach produced comparatively higher accuracies, particularly when combined with Stacking and Hybrid ensemble models, indicating the effectiveness of tree-based feature importance methods in selecting the most relevant features for wine quality prediction.

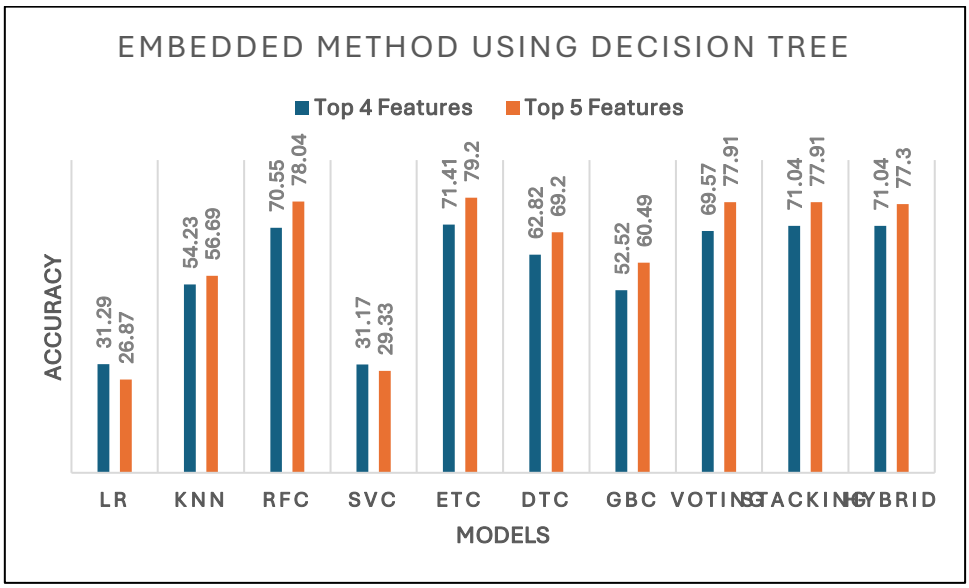


Figure 17. Comparison of Embedded methods using Decision trees

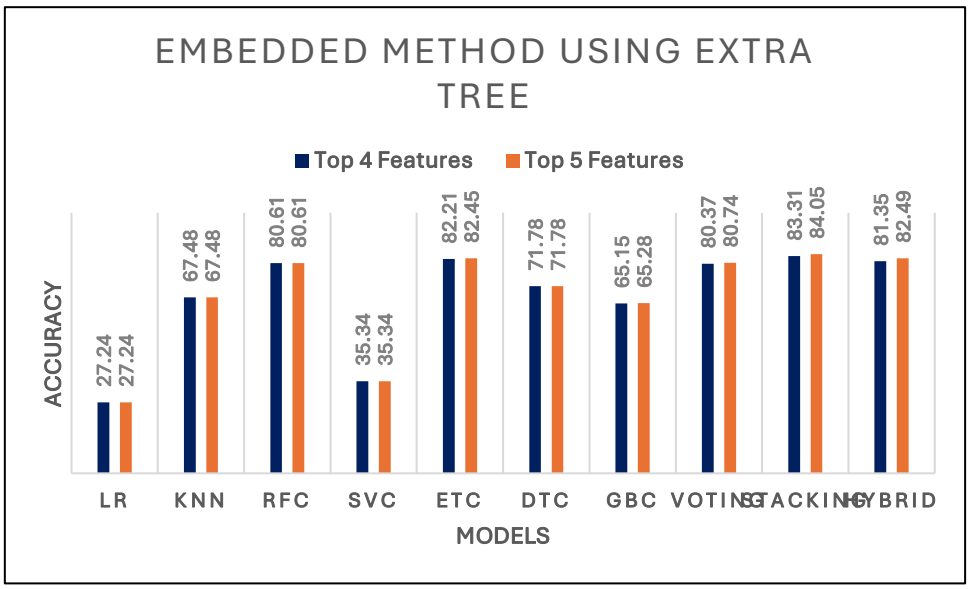


Figure 18. Comparison of Embedded methods using Extra trees

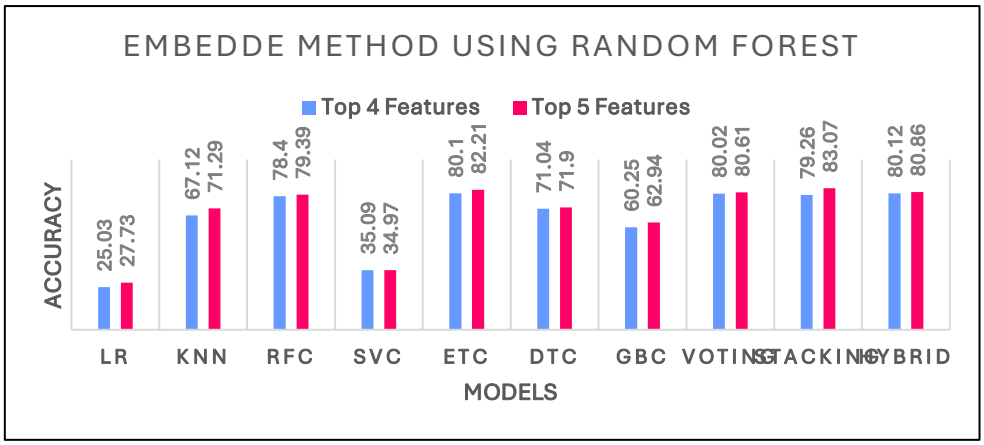


Figure 19. Comparison of Embedded methods using Random Forest

5.4 Feature Transformation Technique Results

Table 7 presents the performance evaluation of feature transformation techniques using PCA and LDA.

For PCA with 3 components, the Hybrid model achieved an accuracy of 72.61%, while ETC achieved 73.13%. When the number of PCA components increased to 4, classification performance improved significantly. The Hybrid Ensemble model achieved an accuracy of 81.47%, and the ETC classifier also achieved 81.47% accuracy.

Among all transformation techniques, LDA produced the best overall results. The proposed Hybrid Ensemble model achieved the highest accuracy of 83.19%, followed by ETC with 81.84% and RFC with 81.23%. The superior performance of LDA can be attributed to its supervised nature, which maximizes class separability while reducing dimensionality.

The results suggest that PCA preserves overall variance within the dataset, whereas LDA better captures discriminative information required for multiclass wine quality classification. However, the overall performance of feature transformation techniques remained lower than the best-performing filter and wrapper feature selection methods.

Table 7. Performance Evaluation of PCA and LDA Feature Transformation Techniques for Wine Quality Prediction

PCA (3 components)	**LR**	33.50	30.70	33.35	30.80
	KNN	69.69	66.58	69.20	672.2
	RFC	71.17	69.26	70.95	69.89
	SVC	35.95	30.92	35.11	29.84

	ETC	73.13	70.93	72.88	71.55
	DTC	68.34	67.05	68.11	67.48
	GBC	53.23	49.18	67.48	49.52
	Voting	73.74	71.55	73.35	71.97
	Stacking	72.64	71.12	72.49	71.63
	Hybrid (Voting + Stacking)	72.61	72.33	71.51	71.82
PCA (4 components)	**LR**	33.37	30.46	33.41	30.03
	KNN	73.50	70.77	73.04	71.31
	RFC	77.91	76.02	77.46	76.55
	SVC	32.39	26.84	33.63	28.67
	ETC	81.47	79.59	81.03	80.03
	DTC	72.88	71.61	72.56	72.04
	GBC	60.61	57.87	59.88	58.16
	Voting	79.14	77.11	78.72	77.52
	Stacking	80.04	78.65	79.60	79.03
	Hybrid (Voting + Stacking)	81.47	80.92	80.55	80.57
LDA	**LR**	51.29	49.79	51.61	50.51
	KNN	64.79	63.15	64.30	63.49
	RFC	81.23	78.69	79.59	79.01
	SVC	41.50	31.76	42.15	35.85
	ETC	81.84	80.35	81.31	80.55
	DTC	71.90	70.98	71.35	71.11
	GBC	64.17	62.32	63.68	62.67
	Voting	80.12	79.92	80.73	80.17
	Stacking	75.95	75.91	75.36	75.62
	Hybrid (Voting + Stacking)	83.19	82.79	82.80	82.79

Figures 20 and 21 present the comparative performance analysis of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for wine quality prediction using different machine learning models. The results show that LDA achieved better classification

performance than PCA due to its ability to maximize class separability, while PCA preserved the overall variance of the dataset. Among all models, the ensemble and hybrid approaches produced higher prediction accuracies compared with traditional classifiers under both feature transformation techniques.

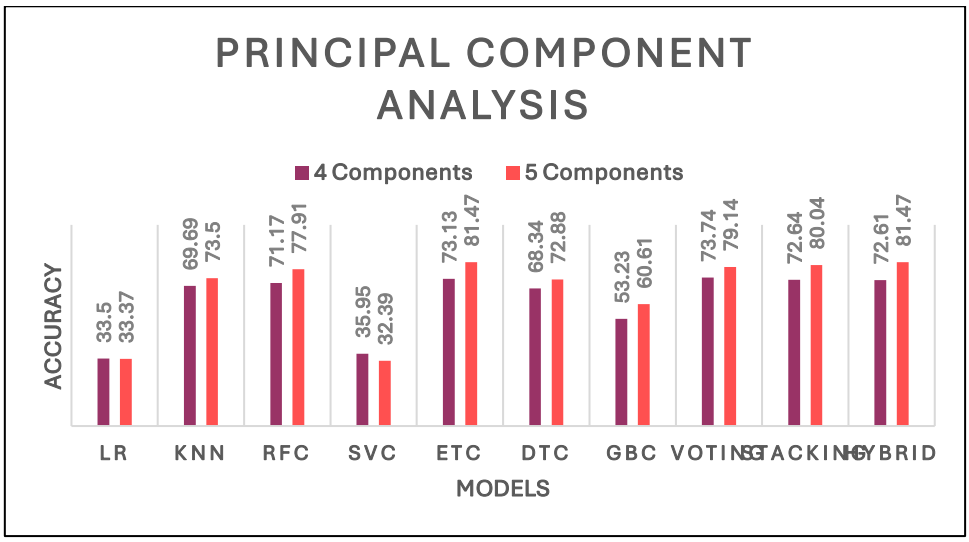


Figure 20. Comparison of Principal Component Analysis

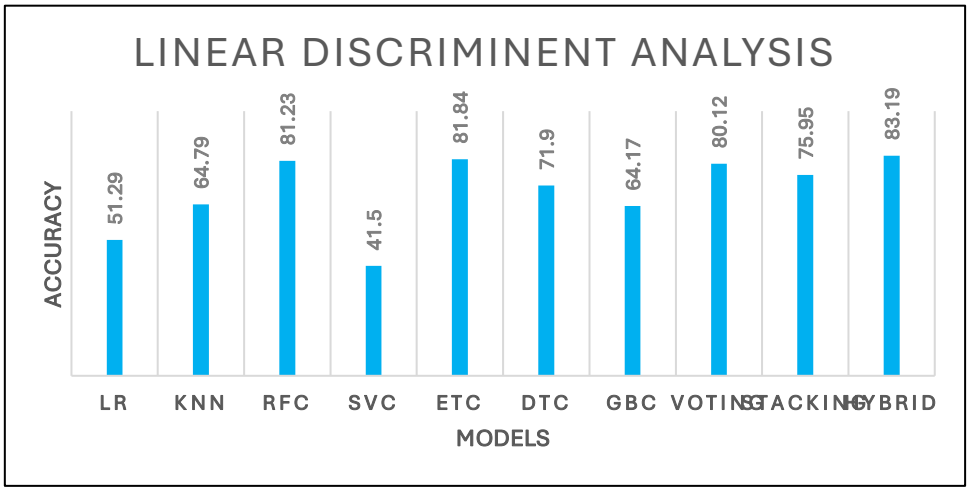


Figure 21. Comparison of Linear Discriminant Analysis

5.5 Comparative Analysis with Existing Studies

Table 8 compares the proposed dimensionality reduction framework with existing machine learning approaches reported in the literature. Previous studies achieved strong performance using traditional machine learning and ensemble models. Devika Pawar (2019) reported a maximum accuracy of 87.33% using Random Forest, while Jaskaran Singh (2021) achieved 89% accuracy using Random Forest. Qingwen Zeng (2023) reported 87.22% accuracy using a Stacking Ensemble approach.

The proposed framework achieved an accuracy of 89.08% using correlation-based feature selection combined with the Hybrid Ensemble model, outperforming most existing approaches. The improved performance can be attributed to the integration of multiple dimensionality reduction strategies and ensemble learning techniques, which effectively reduced feature redundancy and enhanced classification capability.

Overall, the experimental findings demonstrate that dimensionality reduction techniques significantly influence wine quality prediction performance. Among all investigated approaches, correlation-based filter feature selection combined with the proposed Hybrid Ensemble model produced the best overall classification results. The study also confirms that ensemble and hybrid models consistently outperform traditional machine learning classifiers in wine quality prediction tasks due to their ability to capture complex relationships and improve generalization performance.

Table 8. Comparison of the Proposed Dimensionality Reduction Framework with Existing Machine Learning Approaches for Wine Quality Prediction

Author	Year	Models Used	Accuracy (%)
Devika Pawar	2019	Random forest	87.33
		SGD	
		SVC	81
		Logistic regression	85
			86
Jaskaran Singh	2021	Random forest	89
Qingwen Zeng	2023	XGB	86.5
		Random forest	85.12
		SVM	68.27
		MLP	732

		Stacking	87.22
Proposed Work	**2026**	**Dimensionality Reduction Using Machine Learning**	**89.08**

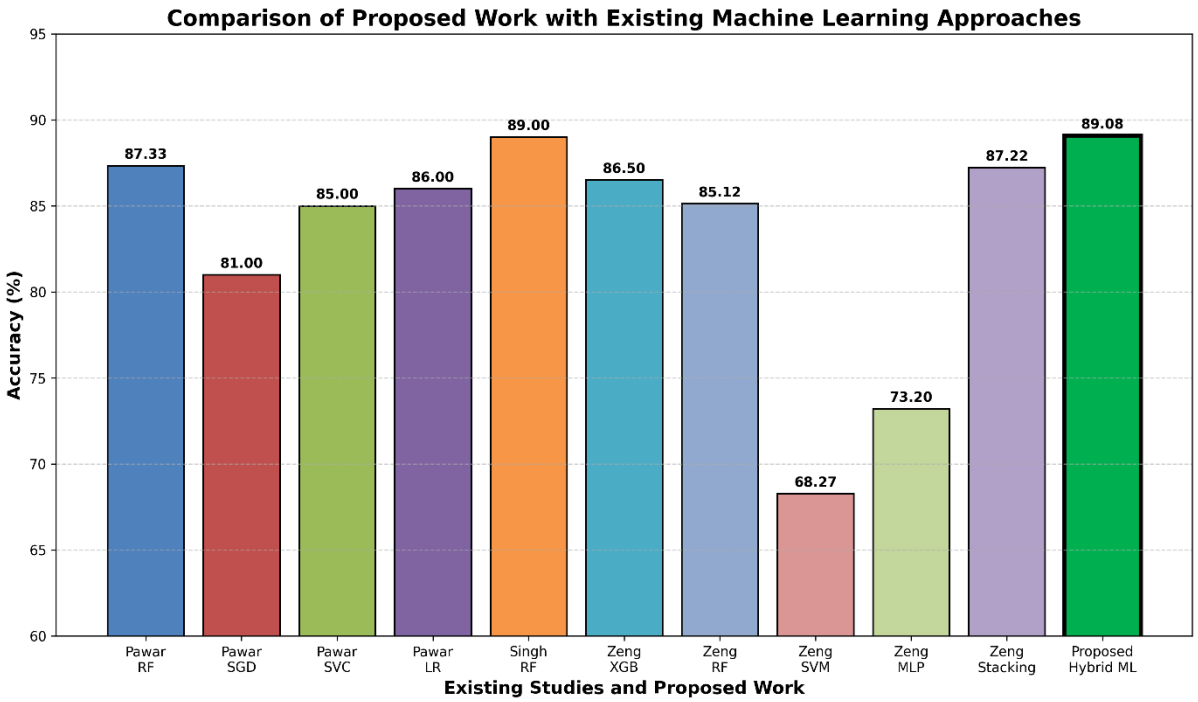


Figure 22. Comparison of the Proposed Work with Existing Machine Learning Approaches for Wine Quality Prediction

Figure 22 presents the comparative accuracy analysis between the proposed dimensionality reduction framework and existing machine learning approaches for wine quality prediction. The results show that the proposed Hybrid Ensemble model achieved the highest accuracy of 89.08%, outperforming most previously reported models such as Random Forest, XGB, SVM, and Stacking approaches. The figure clearly demonstrates the effectiveness of combining dimensionality reduction techniques with ensemble learning methods to improve wine quality prediction performance.

Discussion

This study found that ensemble tree-based models, particularly Gradient Boosting and Random Forest, performed best in predicting wine quality, outperforming the SVM model. These results are consistent with prior findings where ensemble models demonstrated superior accuracy due

to their ability to capture nonlinear feature interactions (Shaw et al., 2020; Dahal et al., 2021). The moderate overall accuracy (~70%) reflects the inherent complexity in modeling sensory-based quality from chemical properties.

SVM's lower performance may stem from class imbalance and overlapping quality labels in the dataset, as previously noted by Jain et al. (2023). While some studies reformulate the task as regression or binary classification to improve performance (Dasgupta & Abraham, 2025), our study retained the full quality range for finer resolution, trading off some accuracy for detail.

Feature importance analysis reaffirmed alcohol as the top predictor, followed by sulphates and volatile acidity aligning with earlier studies (Gupta, 2018; Jain et al., 2023; Kumar et al., 2020; Aich et al., 2018). The role of free sulfur dioxide also emerged, echoing Luque and Mazzoleni (2024), who highlighted its nuanced influence on wine characteristics.

Our findings support literature trends that ensemble models outperform SVM on this dataset (Kumar et al., 2020; Shaw et al., 2020), although exceptions exist. For instance, Aich et al. (2018) and Trivedi and Sehrawat (2018) found SVM competitive under certain conditions. Moreover, simplified classifications (e.g., binary or trinary) often yield better accuracy, as Mahima et al. (2020) demonstrated with a Random Forest–KNN hybrid.

Gradient Boosting slightly outperformed Random Forest in our study, a trend noted in recent works (Dahal et al., 2021; Jain et al., 2023). This aligns with the idea that boosting models improve learning on difficult cases. Interestingly, Bhardwaj et al. (2022) achieved 100% accuracy in a niche dataset using synthetic augmentation and AdaBoost, showcasing how controlled conditions can inflate performance.

In conclusion, our results affirm that classical ML models can effectively predict wine quality using physicochemical data, with ensemble methods consistently ranking top. Feature importance remained stable across models, reinforcing alcohol and acidity as key quality drivers.

Conclusion

This research confirms that machine learning, particularly ensemble models like Gradient Boosting and Random Forest, can effectively predict wine quality from chemical data. Gradient Boosting yielded the best accuracy (~70%), while SVM underperformed. Alcohol, sulphates,

and acidity were consistently identified as top features, aligning with established findings (Dahal et al., 2021; Jain et al., 2023).

By comparing multiple algorithms on a widely used dataset and analyzing feature importance, this study contributes a robust comparative assessment to the literature. Retaining multi-class labels enhanced model granularity but posed prediction challenges.

Future research could incorporate sensory data, apply augmentation or transfer learning, or adapt models for real-time quality monitoring. Enhanced interpretability and model generalization across wine types remain valuable goals. Overall, ML presents a viable, objective approach to supplement traditional wine evaluation methods.

Declarations

Funding

No funding was received to support this study.

Ethics and Consent to Participate

Not applicable. This study used a publicly available wine quality dataset and did not involve human participants or animals.

Competing Interests

The authors declare that they have no competing interests.

Reference

Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. (2018, February). *A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques*. In **2018 20th International Conference on Advanced Communication Technology (ICACT)** (pp. 139–143). IEEE. https://doi.org/10.23919/ICACT.2018.8323674

Anguita, D., Boni, A., Ridella, S., Rivieccio, F., & Sterpi, D. (2005). Theoretical and practical model selection methods for support vector classifiers. In L. Wang (Ed.), *Support vector machines: Theory and applications* (Vol. 177, pp. 159–182). Springer. https://doi.org/10.1007/10984697_7

Anowar, F., Sadaoui, S., & Selim, B. (2021). Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE,

ICA, t-SNE). *Computer Science Review, 40*, 100378. https://doi.org/10.1016/j.cosrev.2021.100378

Bhardwaj, P., Tiwari, P., Olejar, K., Parr, W., & Kulasiri, D. (2022). A machine learning application in wine quality prediction. *Machine Learning with Applications, 8*, 100261. https://doi.org/10.1016/j.mlwa.2022.100261

Bhavya, A. G. (2020). A survey on wine dataset classification. *International Journal of Scientific Research & Engineering Trends, 6*(3), May–June.

Blessie, E. C., & Karthikeyan, E. (2012). Sigmis: A feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology, 6*(3), 385–394. https://doi.org/10.1260/1748-3018.6.3.385

Borugadda, P., Lakshmi, R., & Sahoo, S. (2023). Transfer learning VGG16 model for classification of tomato plant leaf diseases: A novel approach for multi-level dimensional reduction. *Pertanika Journal of Science & Technology, 31*(2), 813–841. https://doi.org/10.47836/pjst.31.2.09

Chen, J., Li, R., & Yang, J. (2022). A prediction model for quality of red wine through explainable artificial intelligence. In *Proceedings of the 2022 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control* (pp. 268–272). IEEE. https://doi.org/10.1109/SDPC55702.2022.9916008

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics, 11*(2), 278–289. https://doi.org/10.4236/ojs.2021.112015

Dasgupta, T., & Abraham, A. (2025). Wine quality assessment using machine learning. In V. Sakalauskas, A. Bajaj, A. Abraham, K. R. Madhavi, & P. Manghirmalani Mishra (Eds.), *Bio-Inspired Computing: Proceedings of IBICA 2023* (pp. 260–268). Springer, Cham. https://doi.org/10.1007/978-3-031-78946-5_25

Di, S., & Yang, Y. (2023). *Prediction of red wine quality using one-dimensional convolutional neural networks*. arXiv. https://arxiv.org/abs/2208.14008

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research, 61*, 863–905. https://doi.org/10.1613/jair.1.11192

Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science, 125*, 305–312. https://doi.org/10.1016/j.procs.2017.12.041

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process, 5*(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

International Organisation of Vine and Wine. (2022). *State of the world vitivinicultural sector in 2021*. OIV Publications.

Jain, K., Kaushik, K., Gupta, S. K., Mahajan, S., & Kadry, S. (2023). Machine learning-based predictive modelling for the enhancement of wine quality. *Scientific Reports, 13*, Article 17042. https://doi.org/10.1038/s41598-023-44111-9

Jović, A., Brkić, K., & Bogunović, N. (2015, May). A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1200–1205). IEEE. https://doi.org/10.1109/MIPRO.2015.7160458

Khilari, N., Hadawale, P., Shaikh, H., & Kolase, S. (2022). Analysis of machine learning algorithm to predict wine quality. *International Journal of Scientific Research in Science, Engineering and Technology, 9*(2), 231–236. https://doi.org/10.32628/IJSRSET229235

Koshiga, N., Borugadda, P., & Shaprapawad, S. (2023, April). *Prediction of heart disease based on machine learning algorithms*. In *2023 International Conference on Inventive Computation Technologies (ICICT)* (pp. 713–720). IEEE. https://doi.org/10.1109/ICICT57646.2023.10134422

Kulasiri, D., Somin, S., & Kumara Pathirannahalage, S. (2024). A machine learning pipeline for predicting Pinot Noir wine quality from viticulture data: Development and implementation. *Foods, 13*(19), 3091. https://doi.org/10.3390/foods13193091

Kumar, S., Agrawal, K., & Mandan, N. (2020). Red wine quality prediction using machine learning techniques. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCI48352.2020.9104095

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Lee, J. A., Peluffo-Ordóñez, D. H., & Verleysen, M. (2015). Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing, 169*, 246–261. https://doi.org/10.1016/j.neucom.2014.12.095

Luque, A., Mazzoleni, M., Zamora-Polo, F., Ferramosca, A., Lama, J. R., & Previdi, F. (2023). Determining the importance of physicochemical properties in the perceived quality of wines. *IEEE Access, 11*, 115430–115449. https://doi.org/10.1109/ACCESS.2023.3325676

Mahabub, A. (2020). A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers. *SN Applied Sciences, 2*(4), 525. https://doi.org/10.1007/s42452-020-2326-yPMC+3

Mahima, Gupta, U., Patidar, Y., Agarwal, A., & Singh, K. P. (2020). Wine quality analysis using machine learning algorithms. In D. K. Sharma, V. E. Balas, L. H. Son, R. Sharma, & K. Cengiz (Eds.), *Micro-Electronics and Telecommunication Engineering* (pp. 11–18). Springer, Singapore. https://doi.org/10.1007/978-981-15-2329-8_2

Mohana, R., Sharma, P., & Sharma, A. (2023**).** Ensemble framework for red wine quality prediction. *Food Analytical Methods, 16*(1), 30–44. https://doi.org/10.1007/s12161-022-02367-3

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*, 21. https://doi.org/10.3389/fnbot.2013.00021

Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: An ensemble classifier. In J. Hemanth, X. Fernando, P. Lafata, & Z. Baig (Eds.), *Proceedings of the International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI 2018)* (pp. 758–763). Springer International Publishing. https://doi.org/10.1007/978-3-030-03146-6_86

Pawar, D., Mahajan, A., & Bhoithe, S. (2019). Wine quality prediction using machine learning algorithms. *International Journal of Computer Applications Technology and Research, 8*(9), 385–388. https://doi.org/10.7753/IJCATR0809.1010

Peng, L., Yuan, R., Shen, L., Gao, P., & Zhou, L. (2021). LPI-EnEDT: An ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA–protein interaction data classification. *BioData Mining, 14*, Article 50. https://doi.org/10.1186/s13040-021-00277-4

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys, 28*(1), 71–72. https://doi.org/10.1145/234313.234346

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47*(5), 1712–1717. https://doi.org/10.1109/18.930915

Sánchez-Maroño, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007). Filter methods for feature selection: A comparative study. In H. Yin, P. Tino, E. Corchado, W. Byrne, & X. Yao (Eds.), *Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2007)* (pp. 178–187). Springer. https://doi.org/10.1007/978-3-540-77226-2_19

Shaw, B., Suman, A. K., & Chakraborty, B. (2020). Wine quality analysis using machine learning. In J. Mandal & D. Bhattacharya (Eds.), *Emerging Technology in Modelling and Graphics* (pp. 239–247). Springer Singapore. https://doi.org/10.1007/978-981-13-7403-6_23

Singh, J., & Manchanda, M. (2021). A machine learning based mechanism for wine quality prediction. *Ilkogretim Online, 20*(3), 3246–3250. https://doi.org/10.17051/ilkonline.2021.03.337

Supriatna, D. J. I., Saputra, H., & Hasan, K. (2023). Enhancing the red wine quality classification using ensemble voting classifiers. *Infolitika Journal of Data Science, 1*(2), 42–47. https://doi.org/10.60084/ijds.v1i2.95

Török, D. F. (2023). Machine Learning for Predicting Wine Quality and its Key Determinants Based on Physicochemical Properties. *Sage Science Review of Applied Machine Learning*, *6*(11), 1–21

Trivedi, A., & Sehrawat, R. (2018). Wine quality detection through machine learning algorithms. In *Proceedings of the 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)* (pp. 1756–1760). IEEE. https://doi.org/10.1109/ICRIEECE44171.2018.9009111

Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. In J. Cabestany, A. Prieto, & D. F. Sandoval (Eds.), *Computational Intelligence and Bioinspired Systems: Proceedings of the 8th International Work-Conference on Artificial Neural Networks (IWANN 2005)* (pp. 758–770). Springer. https://doi.org/10.1007/11494669_93

Ververidis, D., & Kotropoulos, C. (2005, September). Sequential forward feature selection with low computational cost. In *Proceedings of the 13th European Signal Processing Conference (EUSIPCO)* (pp. 1–4). IEEE.

Wang, J., Liu, C., Li, L., Li, W., Yao, L., Li, H., & Zhang, H. (2020). A stacking-based model for non-invasive detection of coronary heart disease. *IEEE Access, 8*, 37124–37133. https://doi.org/10.1109/ACCESS.2020.2975377

Yavas, C. E., Kim, J., Chen, L., Kadlec, C., & Ji, Y. (2025). Exploring predictive modeling for food quality enhancement: A case study on wine. *Big Data and Cognitive Computing, 9*(3), 55. https://doi.org/10.3390/bdcc9030055

Yu, H., & Yang, J. (2001). A direct LDA algorithm for high-dimensional data—with application to face recognition. *Pattern Recognition, 34*(10), 2067–2070. https://doi.org/10.1016/S0031-3203(00)00162-X

Zeng, Q. (2023). Prediction of wine quality using ensemble learning approach of machine learning. In G. Vilas Bhau, Y. Shvets, & H. Mallick (Eds.), *Proceedings of the 2022 International Conference on Mathematical Statistics and Economic Analysis (MSEA 2022)* (pp. 770–774). Atlantis Press. https://doi.org/10.2991/978-94-6463-042-8_110

Zhang, H. (2024). Research of machine learning and feature selection in wine quality prediction. In *Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (IAMPA 2024)* (pp. 5–12). SciTePress. https://doi.org/10.5220/0012982300004601

Zhang, H., Berg, A. C., Maire, M., & Malik, J. (2006, June). SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (Vol. 2, pp. 2126–2136). IEEE. https://doi.org/10.1109/CVPR.2006.301

Zhang, H., Chen, L., & Liu, X. (2021). Comparison of PCA and LDA dimensionality reduction algorithms based on wine dataset. In *Proceedings of the 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications* (pp. 234–238). IEEE.

Zhou, J. Y., Song, L. W., Yuan, R., Lu, X. P., & Wang, G. Q. (2021). Prediction of hepatic inflammation in chronic hepatitis B patients with a random forest-backward feature elimination algorithm. *World Journal of Gastroenterology, 27*(21), 2910–2920. https://doi.org/10.3748/wjg.v27.i21.2910