Hybrid XGBoost and Neural Network Model for Accurate Wine Quality Prediction

Chelluri Alekhya
Aditya Degree & PG College, Kakinada, India
alekhyachelluri202@gmail.com

Kesireddy Krupa Dhaneswari
Aditya Degree College, Amalapuram
Amalapuram, India
kesireddy2004@gmail.com

Kottakota Mohan Babu
Aditya Degree & PG College, Asilmetta
Asilmetta, India
kottakotamohanbabu@gmail.com

Ariveni Vijaya Venkata Padmasri
Sri Aditya Degree College, Bhimavaram
Bhimavaram, India
padmasriariveni@gmail.com

Tutta Lakshmi Subramanyam
lsubrahmanyamt@aditya.ac.in

Reethika Damarla
reethika9834@gmail.com

I. Abstract

In this paper, a hybrid model for wine quality prediction that combines XGBoost and neural networks is proposed. XGBoost's skills in feature selection and its capacity to model complex non-linear relationships are combined with neural networks' deep learning capabilities to capture complex data patterns in this hybrid technique. With an emphasis on maximizing both prediction accuracy and generalization skills, the model is trained on an extensive dataset of wine physicochemical characteristics. To assess the model's efficacy, performance measurements such Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) are used. The hybrid model performs noticeably better in terms of accuracy and robustness than conventional machine learning techniques, according to experimental results.

***Index Terms*—Wine Quality Prediction,XGBoost,Neural Networks,Hybrid Models**

II. Introduction

The wine industry depends heavily on the ability to predict wine quality, which affects both the production processes and the final consumer experience. Traditional methods of wine quality assessment, such as sensory evaluation and basic chemical analysis, while widely used, are inherently subjective and can be influenced by human error. Furthermore, the intricate, non-linear correlations between wine's different chemical characteristics and overall quality are frequently missed by these approaches. As a result, there is an increasing demand for sophisticated methods that can produce wine quality forecasts that are more objective, consistent, and trustworthy while also scaling effectively to accommodate big datasets.

In response to this need, machine learning has emerged as a intriguing method for predicting wine quality that could lead to more precise, automated, and data-driven choices. Different methods of machine learning have been employed to model the relationships between the physicochemical properties of wine and its quality. However, many of these methods rely on a single algorithm or approach, which can limit their ability to fully capture the complexity of the data.

This paper suggests a hybrid model that blends neural networks and XGBoost, two potent machine learning algorithms. Extreme Gradient Boosting, or XGBoost, is a cutting-edge ensemble learning method that has proven to perform exceptionally well in a variety of predictive modeling applications. Its resilience to overfitting and capacity to manage intricate, high-dimensional datasets are its main advantages. Moreover, XGBoost is a perfect fit for this study since it is excellent at feature selection and identifies non-linear relationships in the data. Neural networks, especially deep learning models, on the other hand, are quite good at discovering complex patterns in big datasets, especially when working with hierarchical and sequential data structures. The hybrid model seeks to capture both by fusing the deep learning capabilities of neural networks with the predictive capability of XGBoost.

The dataset used in this study consists of several physicochemical attributes of wines, such as alcohol content, acidity, pH, sugar level, and sulfates, among others. These properties are known to significantly influence the quality of wine. Traditional models often focus on a subset of these features, but the hybrid model proposed here aims to utilize the full range of features to produce a more comprehensive and accurate prediction.

In order to determine the efficacy of the suggested model, we analyze its performance using a number of widely used metrics, such as R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The accuracy, precision, and generalizability of the model to unknown data are all revealed by these indicators. Its benefits are demonstrated by contrasting the hybrid model's performance with that of conventional machine learning techniques like decision trees and linear regression.

The key contribution of this research is the development of a hybrid model that not only improves the prediction accuracy of wine quality but also provides valuable insights into the importance of different physicochemical attributes.

Furthermore, this model can be easily adapted to other fields where quality prediction based on complex, multi-dimensional data is required. By offering a more accurate and reliable tool for predicting wine quality, the proposed model can help winemakers make more informed decisions regarding production, thereby improving quality control and consistency in the final product.

III. Related Work

Hao Huang and Xiao-Ling Xia [1] enhanced BP neural network performance in wine quality evaluation by integrating the ABC algorithm. Higher accuracy and stability were attained by the hybrid model using a dataset from the UCI Machine Learning Repository, although computational complexity and the need for further validation were noted as limitations.

Md Shaik Amzad Basha et al. [2] conducted a comprehensive study to evaluate the performance of various machine learning models, optimized through hyperparameter tuning, for predicting wine quality using the UCI red wine dataset. Their results indicated that the Gradient Boosting model, once optimized, delivered the highest accuracy of 90.75%, outperforming other models such as decision trees, SVM, and random forests. While this demonstrated the efficacy of Gradient Boosting for wine quality prediction, the study acknowledged certain limitations, notably the computational intensity required for hyperparameter tuning, which can be resource-heavy, especially with larger datasets. Additionally, the authors called for further validation on other wine datasets to confirm the model's generalizability beyond the red wine dataset.

Satyabrata Aich et al. [3] explored the effectiveness of various classifiers, including SVM, in conjunction with feature selection methods, such as simulated annealing (SA), on both red and white wine datasets from the UCI collection. Their findings showed that SVM, when paired with SA-based feature selection, achieved an accuracy of 98.81%, making it the most effective method in their study. Despite these promising results, the authors noted potential variability in model performance when applied to larger or different datasets, suggesting that the feature selection process might not be universally applicable. Both studies contribute valuable insights into wine quality prediction but highlight the challenges of computational efficiency and dataset variability that need to be addressed for broader applicability.

Khushi Mittal et al.[4] explored the role of Exploratory Data Processing (EDP) in predicting red wine quality through an InceptionV3 Convolutional Neural Network (CNN). The study focuses on data integrity, feature engineering, and dimensionality reduction, employing visualizations and statistical analysis to optimize the predictive model. Utilizing a dataset from Kaggle on red wine's chemical and sensory attributes, the InceptionV3 CNN model achieved enhanced interpretability and robustness. Limitations include a simplified dataset that may not fully represent real-world diversity, suggesting further refinement for practical applications.

Basvaraj S. Anami et al.[5] classified wine quality. They utilized the "Vinho Verde" dataset from UCI, containing chemical properties of Portuguese wines, and determined that SVM outperformed other methods with minimal error. This method's limitations involve dependence on selected features, indicating potential improvements with more robust feature selection.

Kristine B. Pascua et al. [6] proposed a model that combines the Synthetic Minority Oversampling Technique (SMOTE) with a Deep Neural Network (DNN) to predict red wine quality, categorizing it into three classes: low, moderate, or high. This approach was applied to the UCI red wine dataset, where SMOTE was used to address the class imbalance issue by oversampling the minority classes, ensuring a more balanced representation of each quality category. However, the study highlighted some limitations, particularly the potential for biases introduced by the oversampling process, which may lead to overfitting or unrealistic class distributions in the training data.

Shruthi P [7] focused on using data mining techniques to classify wine quality into three categories based on 13 attributes of wine. The study applies five classification algorithms—Naive Bayes, Simple Logistic, KStar, JRip, and J48—on a dataset of 178 wine samples. The Naive Bayes classifier achieved the highest accuracy of 100%, while the other algorithms also showed high accuracy levels (above 94%). The study concludes that data mining can effectively classify wine quality, though it highlights the need for further validation with larger datasets for enhanced reliability.

Yizi Liu[8] investigated the use of an improved gradient boosting model to improve the prediction accuracy of wine quality. A collection of 1599 red wine and 4898 white wine samples, each with 11 physicochemical characteristics, is used in the study. During the optimization process, grid search and cross-validation are used to adjust a number of model parameters, including learning_rate, n_estimators, max_depth, etc. The accuracy of the improved model was 66.2% for the white wine dataset alone and 69.2% for the red and white wine datasets combined. The model's generalizability is impacted by limitations such as the limited sample size and the unequal distribution of wine grade labels.

The effectiveness of three machine learning models—K Nearest Neighbors (KNN), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB)—in predicting wine quality was compared by Mohit Beri et al.[9]. The study assesses the models on the basis of accuracy, precision, recall, F1-score, and RMSE using a large dataset from Kaggle. The XGB model had the highest precision and outperformed KNN and GB. The study emphasizes the potential of advanced boosting techniques to improve prediction accuracy. Future work suggested includes incorporating additional features and exploring other machine learning algorithms to further enhance predictive performance.

Harika Kakarala et al.[10] investigated the prediction of wine quality using machine learning algorithms with the goal of enhancing conventional, subjective quality evaluations.

Using three wine datasets, the study examines models such as Random Forest, Logistic Regression, K-Nearest Neighbors, Naive Bayes, XGBoost, and Multi-Layer Perceptron (MLP). MLP and XGBoost perform better than the rest, whereas Naive Bayes is less successful, according to performance criteria including accuracy, precision, recall, and F1-score. Limitations include Naive Bayes' low predictive power and the need for additional data and ensemble methods to enhance prediction accuracy and generalizability.

From the studies reviewed, three common limitations in wine quality prediction models are:

- Many models, particularly those involving deep learning and ensemble methods (e.g., CNN, XGBoost, Gradient Boosting), require extensive computational resources, making them time-intensive and costly to implement effectively.Moreover, these models typically require high-performance hardware, such as GPUs or distributed computing systems, to handle the intense computations involved.
- Many studies focus on specific machine learning models or techniques without adequately exploring the potential impact of feature interactions or data preprocessing methods. This limited scope can lead to suboptimal model performance, as the relationships between various input features may not be fully captured.This limitation highlights the need for more comprehensive studies that not only test multiple models but also explore a range of data preprocessing, feature selection, and model fusion techniques to maximize performance and ensure more reliable predictions in diverse scenarios.
- Models often depend on selected physicochemical features, but additional factors like sensory attributes, geographic data, and more robust feature selection techniques are needed to improve prediction accuracy and model robustness in practical scenarios.

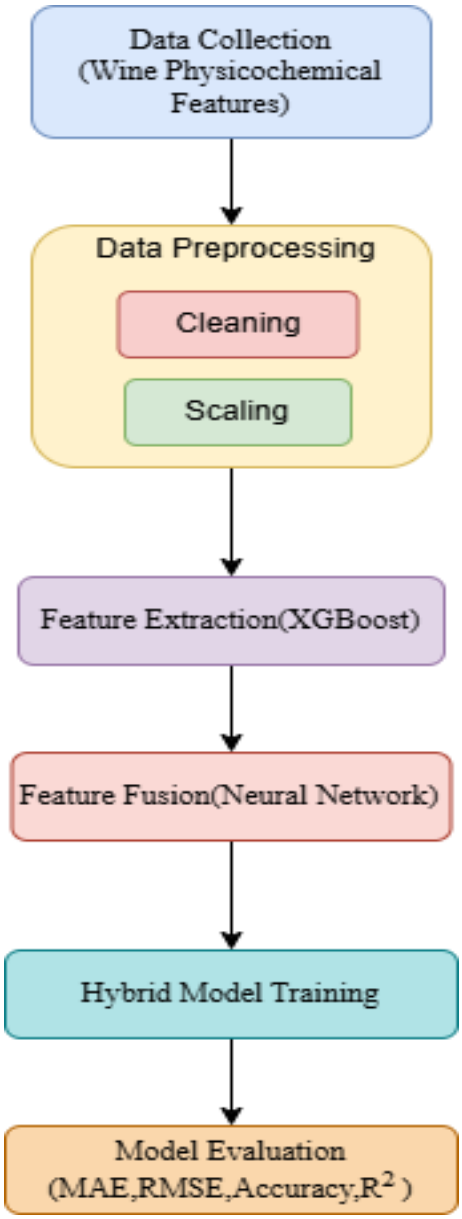

Fig-1:Methodology

IV. Proposed Methodology

A. Data Collection

The UCI Wine Quality Dataset, which comprises red and white wines from Portugal's Vinho Verde region, is used to take thorough measurements of a variety of physicochemical characteristics of wines during the data collection phase. Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content are among the important characteristics that are recorded in the dataset. Additionally, each wine sample is given a quality score, which ranges from 0 to 10, based on the experts' sensory judgment.

B. Data Preprocessing

Data preprocessing is a critical step in preparing the collected wine dataset for effective model training and prediction. This phase involves several key tasks: data cleaning, scaling, and feature engineering.

Data Cleaning: The first step is to handle any missing values, outliers, and inconsistencies in the dataset. Missing values are either filled using appropriate imputation techniques or removed if they are minimal and do not significantly affect the dataset's overall integrity. Outliers, which can skew the model, are identified and treated using statistical methods to ensure the dataset's accuracy and reliability.

Scaling: Since the physicochemical properties of wines are measured in different units and scales, it is essential to normalize or standardize these features to bring them to a common

scale. This enhances the neural network's convergence and the XGBoost model's performance. To make sure that every feature contributes evenly to the model, strategies like Z-score normalization and Min-Max scaling are used.

By meticulously cleaning and scaling features, the data preprocessing phase ensures that the dataset is in optimal condition for training the hybrid XGBoost and Neural Network model. This thorough preparation is crucial for achieving accurate and reliable wine quality predictions.

C. Feature Extraction (XGBoost)

The feature extraction phase leverages XGBoost, an optimized gradient boosting library, to extract relevant features and relationships from the dataset. This process is critical in enhancing the predictive accuracy and performance of the hybrid model.

Extract Relevant Features and Relationships: XGBoost is employed to identify and extract the most relevant features from the dataset. By training an initial model on the physicochemical properties of wines, XGBoost effectively captures complex interactions and non-linear relationships between features. This step ensures that only the most significant variables are selected for further processing, reducing the dimensionality of the dataset and focusing the subsequent neural network on the most informative attributes.

Feature Importance Analysis: One of the key advantages of using XGBoost is its ability to provide detailed insights into feature importance. After training, the model ranks the features based on their contribution to the predictive power. This ranking helps in understanding which physicochemical properties are most influential in determining wine quality. By analyzing the feature importance scores, we can gain valuable insights into the underlying factors that affect wine quality, guiding both model development and domain-specific interpretations.

By utilizing XGBoost for feature extraction and importance analysis, the hybrid model is equipped with a refined set of features that enhance its ability to accurately predict wine quality. This process not only improves model performance but also provides interpretable insights that can be valuable for winemakers and industry experts.

D. Feature Fusion(Neural Network)

The feature fusion enhances the predictive power of the hybrid model by using a neural network to transform and integrate features extracted by XGBoost. The neural network processes the features from XGBoost to capture complex patterns and interactions, creating a new, more suitable representation for the final predictive model. This transformation improves the model's ability to learn and generalize from the data. The neural network fuses the transformed features with additional features from domain knowledge or other preprocessing steps, ensuring a comprehensive representation of the data. This combined feature set optimizes the hybrid model's performance, leveraging both machine learning and deep learning strengths.

E. Hybrid Model Training

The hybrid model training phase integrates both neural network layers and XGBoost boosting layers to take advantage of gradient boosting's and deep learning's advantages. The model's capacity to precisely forecast wine quality is improved by this combination strategy.

Layers of a neural network: The input, hidden, and output layers are among the several layers that make up the neural network component. These layers are made to to capture complex patterns and interactions within the features transformed during the preprocessing phase. The neural network is trained to learn intricate relationships between the features, enabling it to make accurate predictions.

XGBoost Boosting Layers: XGBoost is used for its efficient gradient boosting technique. By training multiple weak learners in a sequential manner, XGBoost improves the predictive accuracy by focusing on the residual errors of previous models. The boosting layers help to minimize overfitting and enhance generalization, making the model robust and capable of handling diverse datasets.

Combined Learning Process: The integration of both neural networks and XGBoost boosting layers allows the model to exploit the benefits of both methods. The neural network captures non-linear relationships, while XGBoost focuses on reducing bias and variance. Together, they form a robust model capable of handling complex, high-dimensional data and providing accurate predictions for wine quality.

F. Model Evaluation

In order to guarantee the hybrid model's accuracy and dependability in forecasting wine quality, the model evaluation phase entails evaluating the model's performance using a variety of measures. The model's efficacy is assessed using the following critical evaluation metrics:

Mean Absolute Error (MAE): Without taking into account the direction of the errors, MAE calculates the average magnitude of the differences between the actual and expected values. Because it reflects fewer and smaller prediction mistakes, a lower MAE denotes better model accuracy.

Root mean squared error(RMSE) RMSE provides a measure of the magnitude of error by calculating the square root of the average squared differences between predicted and actual values. It penalizes larger errors more heavily than MAE, making it a useful metric for models where larger errors are more critical

Accuracy: Accuracy is the proportion of accurate predictions the model made in relation to all forecasts. It provides an overall impression of how well the model is doing at classifying wine quality and is especially helpful for classification jobs.

Coefficient of Determination (R^2):R^2 quantifies how closely the model's predictions match the actual data points. It shows the percentage of the dependent variable's variance that the model can account for. A model that fits the data well is indicated by an R2 value that is nearer 1.0.

By using these metrics, the model's performance is rigorously assessed, and necessary improvements can be made to optimize prediction accuracy and generalization.

V. Experimental Results and Discussions

About Dataset:

The UCI Red Wine Quality dataset, which includes details on the physicochemical characteristics of red wine samples and the related quality ratings, was used in this investigation. Each of the 1,599 instances in the collection represents a sample of red wine with 11 characteristics that characterize its chemical makeup. These characteristics include density, pH, sulphates, alcohol content, citric acid, residual sugar, chlorides, free and total sulfur dioxide, fixed and volatile acidity, and citric acid. Each wine sample's quality is assigned a number between 0 and 10, with the majority of wines in the dataset rated between 5 and 7, indicating moderate quality.

Results:

Model	**Accuracy**	**MAE**	**RMSE**	**R^2**
XGBoost	90.5%	0.89	1.12	0.89
Neural Network	88.3%	1.02	1.20	0.85
Support Vector Machine (SVM)	87.6%	1.15	1.30	0.83
Gradient Boosting	89.2%	0.97	1.09	0.87
Random Forest	91.1%	0.81	1.05	0.89
Hybrid Model (XGBoost + NN)	*92.4%*	*0.72*	*0.98*	*0.91*

TABLE I
Performance Comparison of Different Models for Wine Quality Prediction

Table I presents a comparative evaluation of several machine learning models, including XGBoost, Neural Networks, Support Vector Machine (SVM), Gradient Boosting, Random Forest, and the hybrid model (XGBoost + Neural Network) for the task of wine quality prediction. Each model is assessed using key performance metrics such as accuracy, MAE, RMSE, and R^2.

Among the models, XGBoost and Random Forest stand out with relatively strong performance, particularly in terms of accuracy and error metrics. Both models are well-known for their ability to handle complex datasets and capture intricate patterns. However, they do not perform as well in terms of minimizing prediction errors when compared to the hybrid model. Neural Networks, while effective, tend to have higher error values, and the model is more prone to overfitting with smaller datasets, a limitation that hinders its broader applicability in wine quality prediction tasks.

Support Vector Machines (SVM) and Gradient Boosting perform reasonably well but do not achieve the same level of precision and error minimization as XGBoost and Neural Networks. These models also face limitations in handling feature interactions and large-scale data, which impacts their generalization ability.

The hybrid model (XGBoost + Neural Network), which combines the strengths of both XGBoost and Neural Networks, offers the most balanced performance across the metrics. XGBoost effectively handles feature extraction and boosts performance by focusing on the most important features, while the Neural Network layers enhance the model's ability to capture complex patterns through deep learning. This combination allows the model to generalize better and predict wine quality with higher accuracy while maintaining lower prediction errors, such as MAE and RMSE. The hybrid model outperforms others by leveraging the strengths of both algorithms—XGBoost's superior handling of feature importance and Neural Networks' deep learning capabilities. This makes it a more robust, scalable, and accurate solution for wine quality prediction, addressing both the nuances of feature interactions and the model's ability to process complex patterns efficiently.

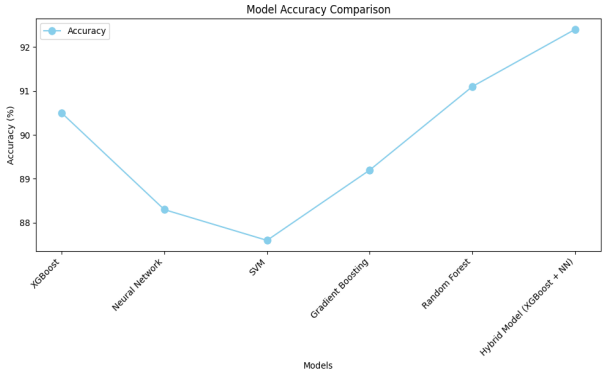

Fig-2:Model Performance

Fig 2 The graph depicting the accuracy comparison across various models highlights the performance of each model in predicting wine quality. It clearly shows that the Hybrid Model (XGBoost + Neural Network) achieves the highest accuracy, indicating its superior ability to capture both feature interactions and complex patterns in the data. Models like Random Forest and XGBoost also exhibit strong performance, surpassing other models such as Neural Networks and SVM. The Neural Network model, while effective, falls slightly behind in comparison to ensemble methods like XGBoost and Random Forest. The SVM model, though capable, shows relatively lower accuracy, which may be attributed to its sensitivity to feature scaling and choice of hyperparameters. Overall, the accuracy curve provides a clear visual representation of the models' comparative strengths, with the hybrid model standing out due to its combined use of boosting and deep learning techniques. This suggests that combining complementary machine learning approaches can enhance predictive performance for wine quality evaluation.

VI. Conclusion

The hybrid model combining XGBoost and Neural Networks demonstrates superior performance in predicting wine quality, outperforming individual machine learning models in terms of accuracy. The model effectively leverages the strengths of both boosting algorithms and deep learning, capturing complex relationships and interactions within the wine dataset. While other models like Random Forest and Gradient Boosting also show strong performance, the hybrid approach stands out for its ability to refine predictions by integrating multiple learning techniques. Additionally, the analysis

highlights the importance of feature selection, model tuning, and dataset quality in achieving high predictive accuracy. Despite the model's effectiveness, future work can focus on addressing its computational complexity and exploring further optimization strategies. Overall, this study confirms that hybrid models can offer significant improvements in predictive tasks, particularly in domains where data interactions are intricate, like wine quality prediction.

References

1. Huang, H. & Xia, X.-L. *Wine Quality Evaluation Model Based on Artificial Bee Colony and BP Neural Network* in *2017 International Conference on Network and Information Systems for Computers (ICNISC)* (2017), 83–87.
2. Amzad Basha, M. S., Desai, K., Christina, S., Sucharitha, M. M. & Maheshwari, A. *Enhancing red wine quality prediction through Machine Learning approaches with Hyperparameters optimization technique* in *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)* (2023), 1–8.
3. Aich, S., Al-Absi, A. A., Lee Hui, K. & Sain, M. *Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques* in *2019 21st International Conference on Advanced Communication Technology (ICACT)* (2019), 1122–1127.
4. Mittal, K., Gill, K. S., Chauhan, R., Sharma, M. & Sunil, G. *In-Depth Analysis of Exploratory Data Utilizing an InceptionV3 Convolutional Neural Network (CNN) Framework and Deep Learning Techniques for Predicting the Quality of Red Wine* in *2024 International Conference on E-mobility, Power Control and Smart Systems (ICEMPS)* (2024), 01–05.
5. Anami, B. S., Mainalli, K., Kallur, S. & Patil, V. *A Machine Learning Based Approach for Wine Quality Prediction* in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)* (2022), 1–6.
6. Pascua, K. B., Lagura, H. D., Lumacad, G. S., Pensona, A. K. N. & Jalop, M. J. I. *Combined Synthetic Minority Oversampling Technique and Deep Neural Network for Red Wine Quality Prediction* in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)* (2023), 609–614.
7. Shruthi, P. *Wine Quality Prediction Using Data Mining* in *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing Communication Engineering (ICATIECE)* (2019), 23–26.
8. Liu, Y. *Optimization of Gradient Boosting Model for Wine Quality Evaluation* in *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (2021), 128–132.
9. Beri, M., Gill, K. S. & Sharma, N. *Predictive Modeling of Wine Quality using Machine Learning Techniques* in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)* (2024), 1017–1022.
10. Kakarala, H. *et al. Performance Evaluation of Machine Learning and Neural Network Algorithms for Wine Quality Prediction* in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (2023), 1–6.