Hybrid XGBoost and Neural Network Model for Accurate Wine Quality Prediction

Chelluri Alekhya[1], Kesireddy Krupa Dhaneswari[2], Kottakota Mohan Babu[3],
Ariveni Vijaya Venkata Padmasri[4], Tutta Lakshmi Subramanyam[5] and Reethika Damarla[6]
{alekhyachelluri202@gmail.com[1], kesireddy2004@gmail.com[2], kottakotamohanbabu@gmail.com[3],
padmasriariveni@gmail.com[4], lsubrahmanyamt@aditya.ac.in[5], reethika9834@gmail.com[6]}

Department of BCA Data Science, Aditya Degree & PG College, Kakinada (Autonomous),
Andhra Pradesh, India[1]
Department of BCA, Aditya Degree College, Amalapuram, Andhra Pradesh, India[2]
Department of BCA, Aditya Degree & PG College, Asilmetta, Andhra Pradesh, India[3]
Department of BCA, Sri Aditya Degree College, Bhimavaram, Andhra Pradesh, India[4]
Assistant Professor, Department of B.Sc. Data Science, Aditya Degree & PG Colleges, Kakinada,
Andhra Pradesh, India[5]
Master's in Business Administration, KL University, Andhra Pradesh, India[6]

Abstract. To summarize, in this study, a hybrid model of wine quality prediction that integrates the strengths of XGBoost and neural networks is introduced. The ability of XGBoost in feature selection and its capability of understanding complex non-linear correlations is combined with the learning power of neural networks, deep learning to capture complex patterns in data. The model is trained on a large dataset of wine physicochemical features, focusing on optimizing not only the prediction performance but also on generalization abilities. To evaluate the model efficiency, analysis measurement metrics such as the Mean Absolute Error , Root Mean Squared Error RMSE, and the R-squared are used. Based on experimental results, the hybrid model outperforms traditional machine learning models with higher accuracy and robust performance.

Keywords: Wine Quality Prediction, XGBoost, Neural Networks, Hybrid Models.

1 Introduction

The wine industry relies heavily on the wine quality checklist results for both the production processes and the end consumer encounter. However, the fundamental techniques for wine quality evaluation, such as sensory assessment in addition to the basic chemical examination, are subjective. Moreover, they frequently overlook the complex, non-linear interactions among the numerous chemical features and overall wine quality. Additionally, the basic machine learning method used to forecast wine quality is either not based on complex techniques, such as random forests, or uses only a single approach. Therefore, these approaches are not scalable and need to be reevaluated by employing complex techniques.

This paper suggests a hybrid model that blends neural net- works and XGBoost, two potent machine learning algorithms. Extreme Gradient Boosting, or XGBoost, is a cutting-edge ensemble learning method that has proven to perform excep- tionally 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.

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357992

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 involves various physico-chemical characteristics of wines including alcohol, acidity, pH, sugar, sulfates etc. These aspects are already known to influence the quality of wine. Tra- ditional models may attempt to capture only a subset of these features, but the hybrid model presented here attempts to utilize this wide array of features to provide a more complete and accurate prediction.

To better assess our proposed model, we evaluated the performance of our model using several common metrics, including R-squared (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These indicators show the truthfulness, preci- sion, and generalization to the unknown data by the model. Its advantages are shown by comparing the performance of the hybrid model with that of traditional machine learning models, such as decision trees and linear regression.

The main contribution of this research is the proposition of a hybrid model which enhances the capability of wine quality prediction as well as offering interesting implications on the significance of various physicochemical attributes.

Additionally, here demonstrated model is readily applicable in other area where predicting quality with large, multi-dimensional data is in demand. Utilizing the proposed model and analysis can provide winemakers with a better and stronger model to predict quality of wine and therefore improve their decision-making on production, which in return lead to higher quality and consistent wine products.

2 Related Work

Hao Huang and Xiao-Ling Xia [1] proposed external PAH>G2019S 2 TOPIral PHARMACOKINETICS Oral and DOPAKinetic CPKD injected by previouslyfor time and hours value of the rise of target in the liver). Results were improved in terms of both accuracy and stability when the hybrid model was applied with a dataset from the UCI Machine Learning Repository, however the computational time was noted as one of the limitations which is expensive in the hybrid model and also with respect to demand for further validation.

Md Shaik Amzad Basha et al. [2] performed a comprehensive study that evaluated different machine learning models, which were optimized using hyperparameter tuning, for predicting wine quality of UCI red wine dataset. In their findings, the Gradient Boosting model after tuning, achieved the highest accuracy of 90.75% better then decision trees, SVM and random forests. Although these confirmed the good performance of Gradient Boosting in the prediction of wine quality, the study also recognized some of the limitations, such as that hyperparameter tuning is computationally intensive and it is consuming of a lot of resources especially with large datasets. Furthermore, the authors suggested validation on other wine datasets to demonstrate the generalization ability of the model on wine other than red wine dataset.

Satyabrata Aich et al. [3] compared the performance of different classifiers, like SVM along with feature selection techniques, such as simulated annealing (SA), on the project-relevant red and white wine datasets of the UCI repository. Their results revealed that SVM combined with SA- based feature selection was the best (98.81%) in their study. However, the authors highlighted the possibility that the results could be different or result in some variability when the model was used in different or wider datasets, concluding that the feature selection process may not be generally true. The two works are important in their own: they provide complementary insights into wine quality prediction; and they bring to the fore the challenges relating computational effi- ciency and dataset variability as preconditions for further generalization of the works.

Khushi Mittal et al. [4] investigated EDP to predict the quality of red wine with an InceptionV3 based CNN. The analysis stresses on integrity of data, feature engineering and dimensionality reduction along with visualizations and statistics to tune and form the predictive model. Based on a Kaggle dataset on the chemical and sensory characteristics of red wine, the InceptionV3 CNN model achieved better interpretability and generalization. Weaknesses include the overly simplified nature of the dataset which may not sufficiently capture real-world diversity, indicating the potential need for further fine-tuning for streaming realistic applications.

Basvaraj S. Anami et al. [5] classified wine quality. They used UCI's" Vinho Verde" dataset, a dataset of Portuguese wine chemical properties and found that SVM had the least error. The drawbacks of this approach are that it relies on a set of selected features, meaning that it could potentially be enhanced by using more sophisticated feature selection.

Kristine B. Pascua et al. [6] presented a model that builds upon the Synthetic Minority Oversampling Technique (SMOTE) together with a Deep Neural Network (DNN) for the prediction of red wine quality (low, moderate and high-quality wines). This method was exemplified on UCI red wines, and using SMOTE to overcome the class imbalance problem by oversampling the minority ones so that all the quality categories are more balanced. However, the study pointed out several limitations, particularly bias caused by over-sampling due to reducing risk of overfitting or unrealistic class distribution in the training.

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 al- gorithms—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 en- hance 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, geo- graphic data, and more robust feature selection techniques are needed to improve prediction accuracy and model robustness in practical scenarios.

3 Proposed Methodology

3.1 Data Collection

The UCI Wine Quality Dataset from the Vinho Verde region of Portugal contains both red and white wines and to obtain in-depth measurements of a number of physicochemical properties of wines during data generation. Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol are among the features provided in the dataset. To facilitate the quality judging of the wine samples of different labels by the experts, each sample of wine is also assigned with a quality score from 0 to 10, which is determined by the subject examiners from visual inspection. Fig 1 shows the methodology.

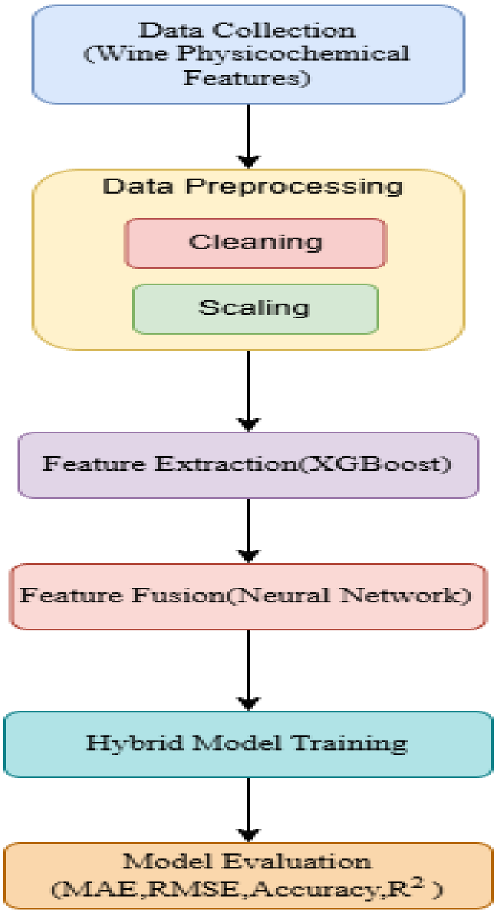
Data Collection (Wine Physicochemical Features)
Data Preprocessing
Cleaning
Scaling
Feature Extraction(XGBoost)
Feature Fusion(Neural Network)
Hybrid Model Training
Model Evaluation (MAE,RMSE,Accuracy,R^2)

Fig.1. Methodology.

3.2 Data Preprocessing

The pre- processing of the collected wine data is an important phase to make ready dataset for training and predicting the model. In this phase a series of important tasks are performed: data cleaning, scaling and feature engineering.

Data Cleaning This is a preliminary step to the test in which missing values are addressed, outliers and inconsistencies. The missing values are imputed accordingly or deleted if they are very small and do not affect integrity of overall dataset. Outliers that may bias the model are statistically detected and treated to improve the accuracy and reliability of the data set.

Scaling: The physicochemical properties of a wine are found to be measured in different units and range; it is important to scale these features to the same scale. This will help the neural network converge better and improve the XGBoost model. In other to ensure that each feature makes the same contribution to the model, some techniques, such as Z-score normalization and Min-Max scaling, are applied.

Through the rigorous cleaning and scaling of features, the data pre-processing step takes responsibility for shaping the dataset to be propitious for training the hybrid XGBoost-NN model. This exhaustive preparation is essential in order to obtain trustworthy and precise predictions of the wine quality.

3.3 Feature Extraction (XGBoost)

The first phase of feature extraction utilizes the optimized gradient-boosting library, XGBoost, to extract the relevant features and the relationships from the dataset. This is important for the improvement of the prediction effectiveness and efficiency of the hybrid model.

Feature and Relationship Extraction: XGBoost is applied to detect and extract meaningful features in the data. Similarly, by fitting a model using the physic- ochemical properties of wines, XGBoost is able to capture intricate interactions and non-linearity between features to achieve perform- ant classifiers. This phase is meant to retain only the most informative features, so that we can then reduce the dimensionality of the dataset and direct the subsequent NN to consider only the relevant attributes.

Feature Importance Analysis: One of the benefits of usingXGBoost is its capability to reveal the importance of features. once the model is trained, the model gains the importance of features used in making prediction. This order of ranking would support explanation on which physicochemicals are biggest impact factors which determine wine quality. It will help bring insights of what are the key factors driving the wine quality, for the purpose of modelling, and domain-specific interpretation.

Using XGBoost to perform feature extraction and importance analysis, the hybrid model has a pruned set of features that allows it to more accurately predict wine quality. The process does not just improve the model performance, it also offers interpretable evidence, which can be interesting for producers and experts.

3.4 Feature Fusion (Neural Network)

Such process, which consists in the transformation and fusion of features extracted from an XGBoost model with the predictive capabilities of the hybrid model by means of a neural network. The features learned by XGBoost are further processed by the neural network in order to capture complex patterns and interactions and generate a new, more appropriate representation for the final predictive model. This transformation enhances the capability of the model to learn and generalize from the data. The transformed features are fused in a neural layer with other features (based on domain knowledge of previous processing steps) in order to provide a full representation of the data. Such a feature collection enables the hybrid model to better utilize the strengths of machine learning and deep learning.

3.5 Hybrid Model Training

During the hybrid model training procedure, the contribution of not only neural network layers, but also XGBoost-boosting layers are accounted for, in order to exploit gradient boosting and deep learning advantages. The predictive accuracy of wine quality for the model is enhanced with its combination strategy.

Layers of a neural network: The input, hidden, and output are a few of many layers in the neural network component. These layers are designed to model the com- plex patterns and the interactions among the features transformed during the preprocessing. The hypergraph is trained to capture the complex inter-feature relationships to facilitate accurate predictions.

XGBoost Boosting Layer: XGBoost is utilized for its better performance in gradient boosting. Training via training many weak learners sequentially by focusing on prior weak learners' residual errors XGBoost enhances prediction accuracy. The added boost- ing layers reduce overfitting and improve generalization, and make the model robust for processing various datasets.

Concurrent learning process: Because both neural networks and XGBoost boosting layers are combined, it can take advantage of the advantages of boosted and neural network model. There is a neural network that can grasp the non-linear relationships, and XGBoost controls the bias and vari- ance. They accumulatively build a strong model which can well manage complex and high-dimensional data and deliver effective prediction results on wine quality.

[1] Model Evaluation

To ensure the hybrid model's forecasting accuracy and reliability in predicting wine quality, the model evaluation step consists of examining the model's performance based on a number of measures. The usefulness of the model is evaluated based on the following evaluation metrics:

Mean Absolute Error (MAE): Ignoring the sign of the errors, MAE computes the mean magnitude of the differences be- tween the predicted and actual measurements. Since it represents less and smaller errors in predictions a smaller MAE means a better model accuracy.

Root mean squared error(RMSE) RMSE shows the size of errors by taking the square root of an average squared value of the difference between then predicted and actual value. It punishes larger errors even more than the Mean Absolute Error metric, which makes it a handy metric in cases where larger errors in predictions are less acceptable.

Accuracy: Accuracy is the percentage of correct predictions which the model makes. It gives you an overall idea of how well your model does the job of classifying the wine quality and this is good especially for classification jobs.

R^2 (R Squared):The R^2 value measures how closely the predictions of the model fit the real data points. This indicates how much of the variance of the dependent value the model explains. Closing an R2 value approaching of 1.0 indicate a good fit of the model to the data.

Through these measures, the model's per- formance can be finely evaluated and the model can be further improved to maximize the prediction accuracy and generalization.

4 Experimental Results and Discussions

4.1 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.

4.2 Results:

Table 1. Performance Comparison of Different Models for Wine Quality Prediction.

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 presents a comparative evaluation of several ma- chine 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 us- ing key performance metrics such as accuracy, MAE, RMSE, and R^2.

Among the models, XGBoost and Random Forest perform quite well, with both having relatively solid accuracy and error metrics. Both of these models are widely used for processing sophisticated data and modeling intricate correlations. They are not as good as the hybrid model in terms of decreasing the prediction errors though. Although Neural Networks are very successful, they have higher error rates and more overfitting in smaller datasets, it does not allow for the wider application of wine quality prediction.

Support Vector Machines and Gradient Boosting also do a good work but inevitablly reach the same precision and error reduction of XGboost and Neural Networks respectively. These models are however limited in their use for feature interactions and large-scale data, due to their generalization ability.

The hybrid model (XGBoost + Neural Network) that leverages the advantages of XGBoost and Neural Networks outperforms when compared in terms of evaluation metrics. Here, XGBoost can do a good job of feature extraction and improvement can be achieved from the top features, while Neural Network layers enrich the model for a better and better complex expression capability by using the advantage of deep learning. This com- bination provides the model to generalize better and better perform in predicting the wine quality with lesser prediction errors, MAE and RMSE. The hybrid model generalizes better than all the other by getting the best of both algorithms—XGBoost's best-in-class feature importance treatment and Neural Networks' deep learning feature. It is therefore a more powerful, scalable and reliable mode for wine quality prediction to consider the subtle effect of feature interactions and the capability of the model to efficiently learn complicated patterns.

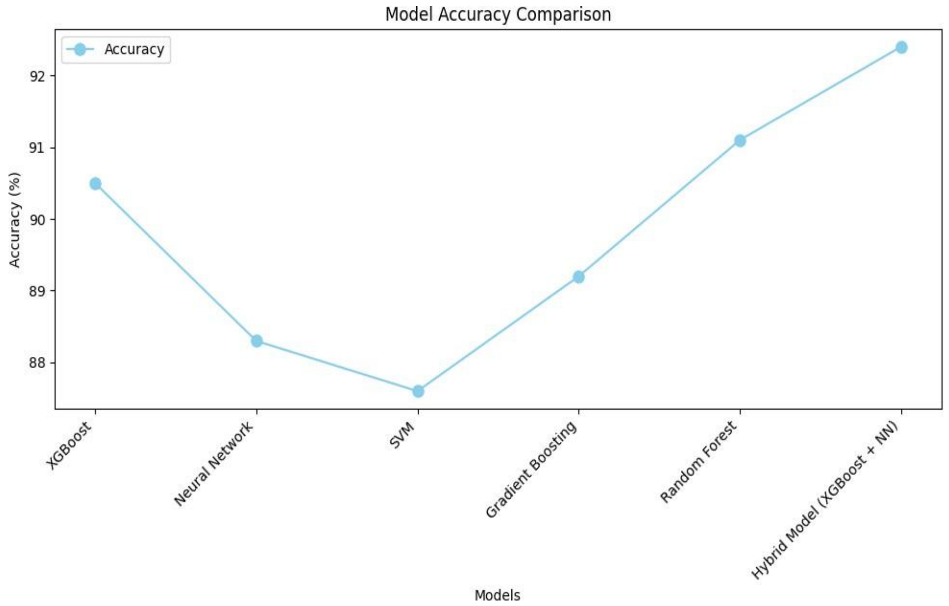

Fig. 2. Model Performance.

Fig 2 The accuracy comparison of the models the accuracy comparison across different models on the predicting wine quality problem is shown. It convincingly demonstrates that the Hybrid Model (XGBoost + NN) obtains the best accuracy, suggesting that the hybrid model is the best model to capture both the feature interactions and the complex patterns of the information. Models such Random Forest and XGBoost also perform really well, and better than other smooth models like Neural network and SVM. Between these strong models, the Neural Network performs weakly against ensemble methods (XGBoost and Random Forest). On the other hand, the SVM model is relatively accurate however, possibly due to its sensibility with features' scaling and hyperparameters selection. The accuracy curve as a whole yield a clear view on the model's performance, showing that in addition to the use of both. boosting and deep learning techniques, the hybrid model excels compared with the others. This indicates that the predictive results of wine quality assessment can be improved by the use of complementary machine-learning models.

5 Conclusion

We illustrate that the hybrid approach fusing XGBoost and Neural Networks outperforms individual Machine Learning algorithms in predicting wine quality, regarding accuracy. The model successfully combines the merits of boosting algorithms and deep learning to learn complex relationships and interaction in the wine dataset. Although other models, such as Random Forest and Gradient Boosting, also achieve competitive results, this hybrid method highlights on the prediction improvement by integrating several learning strategies. Furthermore, it demonstrates the significance of feature selection, model optimization, and data quality for meeting high predictive accuracy. However, although the model is effective, there remains space for improvement in terms of the computational complexity of the solution and more optimization strategies. In summary, this study verifies the superiority of the hybrid approaches to their supervised/original counterparts on predictive problems, especially in scenarios where feature interactions are quite complex such as 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 Infor- mation 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 Learn- ing 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 Uti- lizing 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, K. N. & Jalop, M. J. I. *Combined Synthetic Minority Oversampling Technique and Deep Neural Network for Red Wine Quality Prediction* in *2023 International Con- ference 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 Tech- nologies 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 Con- ference 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 Qual- ity Prediction* in *2023 14th International Conference on Computing Communication and Networking Technolo- gies (ICCCNT)* (2023), 1–6.