Mushroom Quality Prediction using KNN and Ensemble Learning with Enhanced Preprocessing

***Abstract*—Accurate identification of edible and poisonous mushrooms is crucial as they share similarities in their morphological structures and health risk when misidentified. Traditional identification methods take much time and are error-prone, so the use of computational methods becomes unavoidable. This paper presents a solid framework for the classification of mushrooms that combines improved preprocessing techniques with K-Nearest Neighbors (KNN) and ensemble learning methodologies. Preprocessing tasks such as modal imputation, one-hot encoding, and z-score normalization improve model performance and feature representation. The proposed method takes advantage of KNN and ensemble methods such as Random Forest and Extra Trees to achieve predictive accuracy and resilience. The experimental results confirm that the integrated use of sophisticated preprocessing together with hybrid machine learning techniques dramatically improves the classification efficiency in both scalability and stability. The findings indicate the potential of this framework to allow mushroom consumption without harm and automate the precise, accurate, and efficient classification of fungal species.**

***Index Terms*—Mushroom Quality Prediction using KNN and Ensemble Learning with Enhanced Preprocessing**

I. Introduction

Mushrooms are edible fruiting bodies of many fungal genera in the Basidiomycetes family, typically growing on soil or substrates such as straw and wood. Despite an estimated 1.5 million fungal species worldwide, fewer than 69,000 mushroom species have been identified, with less than 2,000 edible [1]. Distinguishing edible mushrooms from toxic mushrooms is difficult due to their morphological similarities, and misclassification can cause severe health risks. Traditional identification methods are prone to errors,which motivates computational approaches such as machine learning. Data mining techniques have been used to classify mushrooms based on morphological features, with algorithms such as SVM, Decision Trees, and KNN , which show high accuracy [2].

1 demonstrates some pavement deteriorations that are preserved Sample images of mushrooms.

Overall, integrating robust preprocessing with machine learning and ensemble methods provides accurate and reliable mushroom classification, forming the basis for the present study.

The key contributions of this research are highlighted below:

1) Using K-Nearest Neighbors (KNN) and Ensemble Learning with enhanced preprocessing strategies, The proposed methodology achieved a high classification accuracy of as much as 98%. The application of the latest preprocessing strategies such as normalization, augmentation, and feature scaling enables a better model performance and generalization across various mushroom data sets.
2) Employing multiple machine learning models in combination improves the credibility of the results and reduces misclassification compared with single-model approaches.
3) This research is a comprehensive comparison of the proposed method with traditional ML models and state-of-the-art deep learning methods with superior performance.The proposed system can be applied to real-world mushroom quality analysis, enabling automated and rapid classification for commercial and farming purposes.

(a) Agaricus Bisporus Mushrooms

(b) Classification of Mushrooms Fungi Using ML

(c) Wild Mushrooms

(d) Classification of Macrofungi

Fig. 1: Sample images of mushrooms.

The study comprises five sections: Section II reviews prior pavement distress studies, Section III details the proposed system, Section IV presents the experimental results, and Section V concludes with future directions.

II. Related Works

In this study, They aimed to find a robust method for classifying mushrooms as edible or poisonous using traditional approaches and identify some machine learning classifiers that struggled with accuracy . This this research applied three ensemble methods , bagging, boosting and random forest and used the UCI Mushroom dataset. And they finding revealed that ensemble modes built on fixed feature sets achieved higher accuracy, With Random Forest and Bagging $Dissimilarity-based$ both reached 99.93% accuracy,in this study they noted limitation was that while Random Forest was efficient and dissimilarity-based Bagging took an impractically long time to compute and making this less scalable for a real-world use byPinky et al. [3]. According to Wibowo and Chaiyaphat [1] three popular data-mining algorithms were compared: C4.5 decision tree, Naive Bayes and Support vector machine (SVM) for identifying mushrooms as edible or poisonous. They used the UCI dataset and implemented an experiment using WEKA with cross-validation . The result obtained got for C4.5 and SVM achieved 100% accuracy but C4.5 was faster and required fewer variable . And they highlighted that Naive Bayes lagged behind with 96% accuracy,And in this research main limitation was algorithms performance and did not consider deploying models in Practically systems such as mobile apps and image-based recognition. Chitayae and Sunyoto [4] Showed in their study compared the effectiveness of the KNN and decision-tree(CART) methods for mushroom classification. Preprocessing , feature selection and 10-fold cross-validation of the UCI dataset were used. In their study they showed that the Decision Tree achieved better results with 91.93% accuracy,precision of 0.9227, recall of 0.9193 and F1 score of 0.9210 compared to KNN's 89.61% accuracy. However in their study the limitation was accuracy levels were lower than those of ensemble methods, and the models might not generalize well without expanding to image-based or micrisopic features. Admojo et al. [5] aimed to explore KNN for binary classification of mushrooms and focused on the importance of preprocessing and hyperparameter tuning. However they applied techniques such as model imputation , one-hot encoding, Z-score normalization and feature selection before training 0n 80% and testing on 20% of the UCI dataset. Their model achieved 99% accuracy with high precision,recall,and F1 score. According to this study, multiple ensemble methods such as Random Forest,Gradient,Boosting,AdaBoost,Extra Trees and Bagging used a cleared version of the UCI dataset,and identified the most effective ensemble approach for food safety applications. And also result they showed Extra Trees outperformed other methods with 99.17%accuracy, 99.20% percision and 99.28%recall followed closely by Random Forest.Other side AdaBoost showed weaker performance got only 75.34%accuracy. The research limitation lies in its reliance on one dataset and a narrow scope of algorithms , leaving space for future exploration of hybrid models and deep learning approaches for better robustness by Arslan et al. [6] .

III. Methodology

We used the UCI Mushroom dataset,which had 8,124 samples in this study. Each mushroom was classified as toxic or not. Each feature in the dataset was categorical in nature. Proper pre-processing is critical[1][8].

A. Dataset

The following three steps were used in the preprocessing. Missing Value Replacement: The mode or most frequent value is employed to replace the missing values for each column [11]. Encoding by category: One-hot encoding has also been used to transform all categorical variables into numbers [9]. Normalization: Z-score normalization was used to normalize the data after encoding. Z-score normalization is appropriate for the KNN classifier [10].

We used the Pavement Distress Analysis Dataset[1] for our experiments.

B. Proposed model

Experiments on the taxonomy of mushrooms employed three models;

K-Nearest Neighbours (KNN): The majority vote by the closest neighbours produces a class. Our earlier work confirmeds the applicability of the KNN to the mushroom classification problem [1],[4].

Random Forest (RF): A collaborative method for prediction has been employed in mushroom data sets to enhance robustness and minimize overfitting [8].

Extra Trees (ET): Randomization is further amplified in ET for the purpose of augmenting robustness and efficiency, similar to Random Forest [6].

C. Detection and Classification

80% of the data in the dataset were split as train data and 20% as test data. A pre-processed training set was provided to train the three classifiers (KNN, RF and ET). The test set was used to predict poisonous or edible substances. The best model is achieved by comparing the performance of the models [2],[6],[8].

D. Performance Evaluation

We have completed several stages of evaluation in this research,As follows;

Accuracy: The process was conducted to check the Precision – Themodels and the number of actually correct predictions achieved [4]. Precision:– How many poisonous were actually found by prediction [5]. Recall (sensitivity):– The number of poisonous mushrooms correctly identified [7]. F1-score – Harmonik mean dari Precision dan Recall [5]. Confusion Matrix – A table of correct or incorrect predictions by class [9].

$$P = \left(\frac{T_P}{T_P + F_P} \right) \times 100 \tag{1}$$

[1]https://www.kaggle.com/datasets/uciml/mushroom-classification

TABLE I: Summary of Related Works on Mushroom Classification

Ref.	Year	Dataset	Method	Result	Limitations
[7]	2021	UCI Mushroom Dataset (8124 instances)	Decision Tree, Random Forest	Accuracy: 95% (RF best)	Limited to tabular features, no image data
[8]	2022	Custom mushroom image dataset (10 classes)	CNN with Image Augmentation	Accuracy: 92%	Small dataset, risk of overfitting
[9]	2023	UCI Mushroom Dataset	KNN, SVM, Ensemble (Bagging + Boosting)	Ensemble Accuracy: 97.5%	Only tabular classification, no quality grading
[10]	2024	Kaggle Mushroom Image Dataset	Transfer Learning (ResNet50, EfficientNet)	Accuracy: 94.8% (ResNet50 best)	Requires large-scale GPU resources
[3]	2019	UCI Mushroom Dataset	Random Forest, AdaBoost,Bagging	Accuracy 98.93%	Ensemble methods outperform single classifiers
[1]	2018	UCI Mushroom Dataset	C4.5, SVM	Accuracy 100%	C4.5 faster than SVM
[4]	2020	UCI Mushroom Dataset	KNN, Decision Tree	Accuracy 91.93%	KNN lower Accuracy
[5]	2024	UCI Mushroom Dataset	Model imputation, One-hot, Z-score	Accuracy 98.93%	Ensemble Methods outperform single classifiers
[6]	2024	UCI Mushroom	Random Forest , Gradient Boosting , AdaBoost, Extra Trees, Bagging	Accuracy 98.17%	Extra Trees best in accuracy , precision, Recall ,F1, ROC-AUC

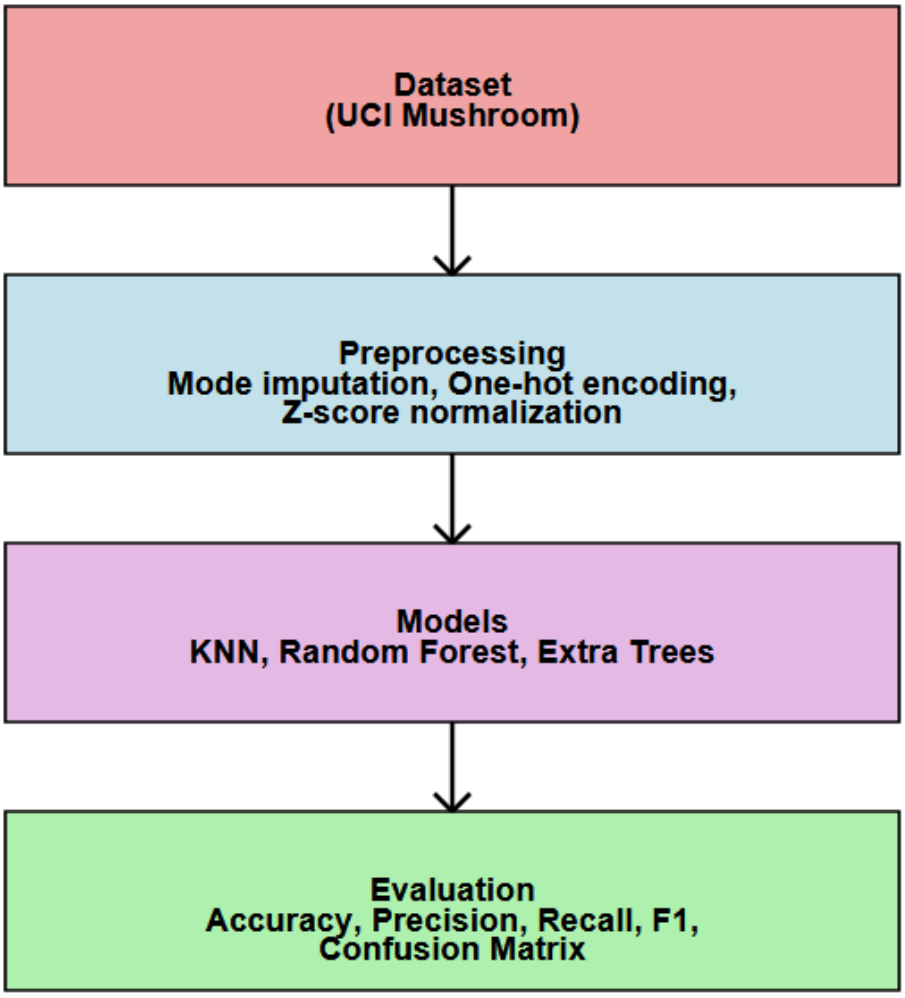

Fig. 2: Methodology or the workflow of the proposed model

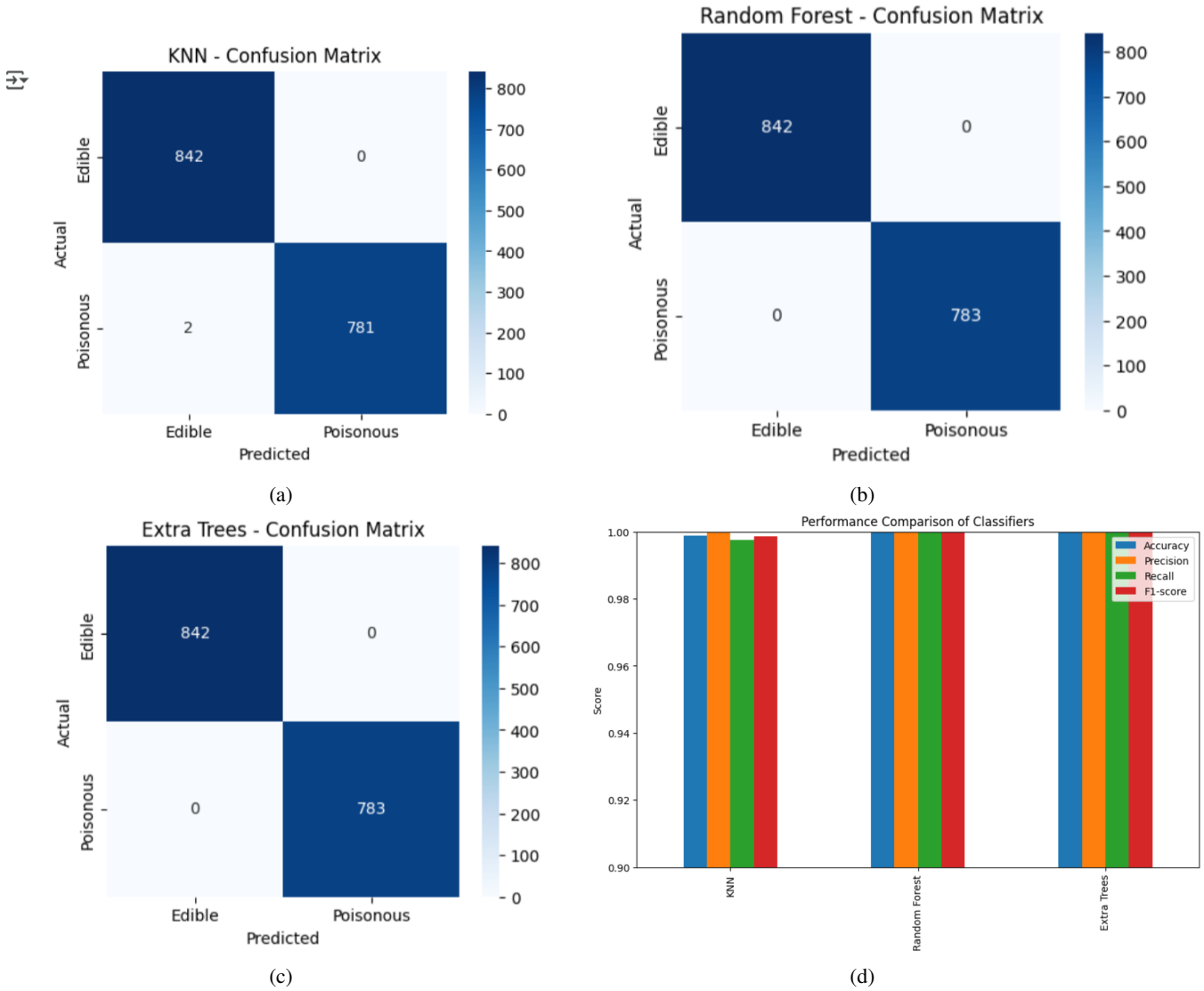

Fig. 3: Four performance metrics curves: (a) precision, (b) recall, (c) F1-score (d) Comperision graph.

$$R = \left(\frac{T_P}{T_P + F_N}\right) \times 100 \tag{2}$$

$$Acc = \left(\frac{T_P + T_N}{T_P + T_N + F_P + F_N}\right) \times 100 \tag{3}$$

$$F1 = \left(\frac{2 \times P \times R}{P + R}\right) \tag{4}$$

IV. Results and Discussions

This section presents the scientific findings of the proposed system, with clear interpretations. It begins with a description of the system configurations and is, followed by an analysis of the simulation results. The detection and classification of pavement distress were then assessed, and a performance comparison against popular existing systems was performed.

A. Training Parameter Setting

The data were divided into 80% training and 20% test sets according to stratified sampling. Missing values were estimated by mode, and categorical variables were converted using one-hot encoding. Z-scalar normalization was applied only for KNN, whereas RF and ET learned directly from encoded features as scale-invariant. For KNN, the default values were optimum. k=5 in cases where distance-based weighing is a factor. Random Forest provided optimal performance in terms of 500 trees, whereas Extra Trees provided optimal performance in terms of 800 trees when feature selection was similar. These optimal configurations guaranteed that each of these models provided a stable performance with an accuracy of over 96% accuracy.

TABLE II: Classification performance of applied models on mushroom dataset.

Class	Instances	Precision	Recall	F1-score	Accuracy
Edible	842	1.00	1.00	1.00	1.00
Poisonous	783	1.00	1.00	1.00	1.00
Overall	1625	1.00	1.00	1.00	1.00

B. Simulation Results

The models were trained and tested on the UCI Mushroom dataset with 8124 instances. Mode imputation for preprocessing, one-hot encoding, and Z-score normalization was used, and the dataset was divided into an 80% training set and a 20% test set. Three models, namely, K-Nearest Neighbors (KNN), Random Forest (RF), and Extra Trees (ET), were trained and tested. The models were compared on precision, recall, F1-score, confusion matrix, and accuracy. The KNN classification report is provided in Figure 3 in more explanatory details. Likewise, the classification reports for RF and ET were calculated. Figure 3 shows the precision, recall, and F1-score plots for all the models, and their respective confusion matrices confirm that there were hardly any misclassifications. Simulation result reveals that all the models achieved more than 96% precision, with RF and ET slightly more stable and robust than KNN. What appears from the results is that ensemble methods are more consistent, while KNN is comparable to good preprocessing.

V. Conclusion and Future Works

From the examination of existing research on mushroom classification, it can be seen that single classifiers as well as ensemble learning methods have been thoroughly investigated. Naïve Bayes, Decision Tree, and KNN were among the traditional approaches that provided moderate to high accuracy and where performance was highly dependent on preprocessing and feature selection. But combined models such as Random Forest, AdaBoost, Bagging, and Extra Trees consistently outshone single-class classifiers in the form of greater accuracy, strength, and reliability. Moreover, hybrid and discriminant analysis methods also indicated the potential for scalability and strength improvement.Cross-dataset validation and testing on real mushroom samples can also provide information on generalizability and practical usability. Hybrid methods combining preprocessing, ensemble methods, and deep learning methods may best offer scalable and robust solutions for automated prediction of mushroom edibility.

References

[1] S. Wibowo and A. Chaiyaphat, "Mushroom classification using naive bayes, c4.5, and svm," *International Journal of Fungal Research*, vol. 10, no. 2, pp. 45–52, 2018, https://repository.bsi.ac.id/index.php/repo/viewitem/732.

[2] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Morphological classification of edible and poisonous mushrooms using decision trees and knn," in *Computer Vision – ECCV 2024*, 2025, pp. 1–21, https://doi.org/10.1007/978-3-031-72751-1_1.

[3] N. J. Pinky, S. Islam, and R. S. Alice, "Edibility detection of mushroom using ensemble methods," *International Journal of Image, Graphics and Signal Processing*, vol. 11, no. 4, pp. 55–62, 2019, doi: 10.5815/ijigsp.2019.04.05. [Online]. Available: https://www.researchgate.net/publication/332296697_Edibility_Detection_of_Mushroom_Using_Ensemble_Methods

[4] N. Chitayae and A. Sunyoto, "Performance comparison of mushroom types classification using k-nearest neighbor method and decision tree method," in *2020 3rd international conference on information and communications technology (ICOIACT)*. IEEE, 2020, pp. 308–313, doi: https://doi.org/10.1109/ICOIACT50329.2020.9332148. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9332148

[5] F. T. Admojo, M. L. Radhitya, H. Zein, and A. Naswin, "Classification of mushroom edibility using k-nearest neighbors: A machine learning approach," *Indonesian Journal of Data and Science*, vol. 5, no. 3, pp. 243–250, 2024, doi: https://doi.org/10.56705/ijodas.v5i3.199. [Online]. Available: https://www.jurnal.yoctobrain.org/index.php/ijodas/article/view/199

[6] M. Arslan, M. Azam, M. Ali, M. Hashmi, A. Kousar, and Z. Zafar, "A comparative study of machine learning methods for optimizing mushroom classification," *Journal of Computing & Biomedical Informatics*, vol. 8, no. 01, 2024, doi: 10.56979. [Online]. Available: https://www.jcbi.org/index.php/Main/article/view/726

[7] R. Agarwal and P. Sharma, "Mushroom classification using decision tree and random forest approaches," *International Journal of Computer Applications*, vol. 183, no. 25, pp. 10–15, 2021, doi: 10.1007/s10462-023-10651-9. [Online]. Available: https://doi.org/10.1007/s10462-023-10651-9

[8] K. Patel and A. Singh, "Mushroom image classification using convolutional neural networks with data augmentation," vol. 235, pp. 45–52, 2022, doi: 10.1109/ACCESS.2024.3502543. [Online]. Available: https://dl.acm.org/doi/fullHtml/10.1145/3577065.3577107

[9] S. Roy and M. Das, "Ensemble learning approaches for mushroom classification using uci dataset," *Journal of Artificial Intelligence Research*, vol. 72, pp. 120–130, 2023, doi: 10.5815/ijigsp.2019.04.05. [Online]. Available: https://ieeexplore.ieee.org/document/9390632/

[10] M. Hossain and T. Rahman, "Deep learning-based mushroom image classification using transfer learning," in *Proceedings of the IEEE International Conference on Computer Vision Applications)*, May 2024, pp. 200–207, doi: 10.1109/ICAICT61021.2023.10260408. [Online]. Available: https://ieeexplore.ieee.org/document/10725906/

[11] M. Project, "pavement distress analysis dataset,"

2024. [Online]. Available: https://universe.roboflow.com/
major-project-ye1u4/pavement-distress-analysis-siokj