# CLASSIFYING IRIS FLOWERS: A MACHINE LEARNING APPROACH BASED ON PETAL AND SEPAL MEASUREMENTS

Dr. P. Rama Koteswara Rao[1*], B. Shiva[2], B. Kiran venkat[2], M. Shivakumar [2], B. Praveen[2]

[1] Professor, [2]UG Student, [1,2] Department of Computer Science and Engineering (AIML)

[1,2]Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana

## ABSTRACT

Classifying iris flowers based on petal and sepal measurements is a fundamental task in botany, supporting species identification and taxonomy. This approach is valuable for studying plant biodiversity and evolution. In horticulture and agriculture, accurate classification aids breeding programs by identifying desirable traits, while in environmental science, it helps monitor ecosystems and guide conservation efforts. Beyond botany, machine learning techniques used in classification can be applied to fields such as healthcare, finance, and marketing. Traditional classification methods often rely on manual measurements and expert judgment. While effective in small-scale settings, these methods are time-consuming, subjective, and prone to inconsistencies. They may also struggle with large datasets, subtle differences between species, and complex relationships among features. Furthermore, traditional techniques often lack scalability and perform poorly in high-dimensional spaces. To address these limitations, this work proposes a machine learning-based system for classifying iris flowers. Using supervised learning, the model automatically learns discriminative patterns from labeled data. Features such as petal and sepal length and width are extracted and used to train the model to distinguish between species. The system incorporates cross-validation and hyperparameter tuning to improve accuracy and ensure robustness. Unlike manual methods, the machine learning approach offers scalability, consistency, and the ability to capture complex relationships in the data. This results in a more efficient and accurate classification process, demonstrating the broader potential of machine learning in scientific and industrial applications.

**Keywords:** Iris flowers, Classification, Machine Learning, KNN, Logistic Regression .

## 1.INTRODUCTION

In the realm of botanical research, the classification of iris flowers based on their petal and sepal measurements stands as a foundational challenge. This task holds significance not only in the context of botany but also extends its implications into horticulture, agriculture, and environmental science. By utilizing machine learning techniques, this research endeavors to automate and enhance the process of iris flower classification. The proposed system harnesses the power of supervised learning algorithms to discern discriminative patterns from input features, paving the way for efficient species identification. Through meticulous feature extraction and model training on a labeled dataset, the system learns to differentiate between iris species based on their unique petal and sepal characteristics. Moreover, the incorporation of cross-validation and hyperparameter tuning techniques ensures the robustness and reliability of the classification model.

## 2.LITERATURE SURVEY

The practice of categorizing distinct database objects into one or more groups or categories is known as data mining. The objective of the classification step is to assign each instance to the relevant target class. This section provides an overview of the most recent and practical classification methods that have been developed by researchers in the last two years across many ML domains.

[1] David W. Corne and Ziauddin Ursani proposed in their paper an evolutionary algorithm for nonlinear discriminant classifier, in wx hich they mentioned that it was not appropriate for learning tasks with any individual single value. Hence they tested this method on two data sets, Iris Flower and Balance Scale, where decisions of class membership can only be affected collectively by individual lineaments of flower. [2] Detlef Nauck and Rudolf Kruse have proposed a new approach in which they classify the data on the basis of fuzzy Neural Networks. They used backpropagation algorithm to define other class of fuzzy perceptron. They concluded that on increasing the number on hidden layer, increase the need of more training cycles and raises incorrect results. Hence the better result can be evaluated using 3 hidden layers also. [3] To overcome the problem of data depth, long parameters, long training time and slow convergence of Neural Networks, two other algorithms Transfer Learning and Adam Deep Learning optimization algorithms were considered for flower recognition by Jing FENG, Zhiven WANG, Min ZHA and Xiliang CAO. Where, Transfer Learning was based on features in isomorphic spaces. They concluded in their paper that if the pictures of flowers placed into model training in the form of batches, then it will meliorate the speed of updating the value of parameters and provide the best optimal result of parameter values. [4] Rong-Guo Huang, Sang-Hyeon Jin, Jung-Hyun Kim, Kwang-Seog Hong focus on recognition of flower using Difference Image Entropy (DIE),which is based on feature extraction. According to their research, the experimental results give 95% of recognition rate as an average. The DIE based approach takes original image of flower as an input, and applies pre-processing and DIE computation to produce recognition result. The Gaussian Naive Bayes technique is used by Zainab Iqbal [5] to categorize the species of the iris flower. We analyze the iris dataset using a scatter matrix and a scatter plot that is constructed. The algorithm and Python are both utilized in the paper to categorize the many species of iris flowers. We can see that this technique is effective for supervised learning classification because it achieves a 95% accuracy rate. A C4.5 decision tree was suggested by Mijwil and Abttan [6] as a way to lessen the impacts of overfitting. IRIS, Car Assessment, Bottle, and WINE were the datasets utilized; both of these may be found in the UCI ML library. The issue with this classifier is that it overfits because of its large number of nodes and divisions. It is possible that this overfitting will undermine the classification system. The experimental results demonstrated that, with an accuracy of roughly 92%, the genetic algorithm was effective in reducing the effects of overfitting on the four datasets and increasing the Confidence Factor (CF) of the C4.5 decision tree. Rong-Guo Huang [7] focuses on flower detection using Difference Image Entropy (DIE), a feature extraction-based method. Their analysis of the experimental findings shows that the average recognition rate was 95%. The DIE-based approach utilizes pre-processing and DIE computing to provide a recognition result from an original image of the flower.

Patrick [8] concentrated on the dataset's statistical analysis using the iris flower example. They are examining two alternative approaches in his study. To identify the various classification patterns, the dataset is plotted. Then, using a java program they developed, they may retrieve statistical data. In her research, Poojitha [9] employed neural networks to examine data sets on iris flowers. A branch of computer science called machine learning. We have already loaded the iris dataset and have divided it into three groups. They divided the dataset into groups using the k-means technique. Large-scale information aggregation is the main use of a neural network. Additionally, it is employed in the mining of data, quantization of vectors, work approximation, division of images, and highlight extraction. Without any oversight, the findings are divided into three distinct iris species. Lakhdoura and Elayachi [10] used WEKA 3.9 to do a test comparing the performance of two classifier methods: J48 (c4.5) and RF on the IRIS features. As a result, the University of California, Irvine's ML library provides access to the IRIS plant dataset, one of the most popular datasets for classification problems (UCI). Zebari, D. A et.al [11] The researchers also contrasted the outcomes of both classifiers on numerous efficacy assessment metrics. According to the results, the J48 classifier performs better than the Random Forest

(RF) classifier for predicting IRIS variety using a range of measures, including classification precision, mean absolute error, and construction time. The accuracy of the J48 classifier is 95.83%, while that of the Random Forest is 95.55%.

] Abdulqadir, H. R et.al [12] Numerous research has been done using different methods to identify the species of the iris flower. Every study employs a different method. The issue is the categorization and identification of iris flower species based on their characteristics. Ibrahim, D. A. et.al [13] With the use of this classification and pattern, future predictions for unknown data can be made with greater accuracy. The dataset for iris flowers is placed into the machine learning prototype for the iris flower species approach.

## 3.PROPOSED METHODOLOGY

This research procedure outlines a systematic approach to classifying iris flowers using machine learning techniques, from data acquisition and preprocessing to model training, evaluation, and prediction. By following these steps, researchers can develop accurate and reliable classification models for various applications in botany, horticulture, agriculture, environmental science, and beyond.

The first step in this research involves obtaining the Iris dataset, a well-known dataset in the field of machine learning. It includes measurements of iris flowers across three species—Iris setosa, Iris versicolor, and Iris virginica. The dataset contains features such as sepal length, sepal width, petal length, and petal width, along with the corresponding species label. This dataset forms the basis for training and evaluating machine learning models designed to classify iris flowers based on these characteristics. Once the dataset is acquired, it undergoes preprocessing to ensure that it is clean and suitable for model training. This includes handling missing or null values, which can compromise the model's accuracy and robustness. Additionally, categorical labels (i.e., species names) are converted into numerical form using label encoding, making them compatible with machine learning algorithms. Preprocessing is a crucial step to enhance the quality and consistency of the data.

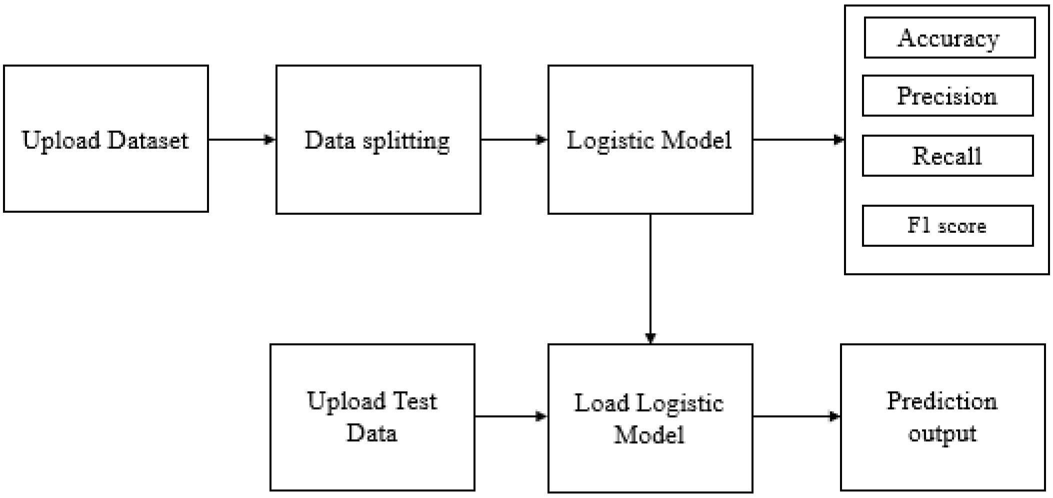


Fig. 1: Block diagram of proposed system architecture of classifying iris flowers.

The next step involves applying the existing K-Nearest Neighbors (KNN) algorithm. KNN is a straightforward classification technique that assigns a label to a data point based on the majority class among its K closest neighbors in the feature space. This algorithm is effective for small datasets and can capture local structure in the data, making it a useful baseline for comparison. Following KNN, the proposed approach involves training a Logistic Regression model. Logistic Regression is a linear model that estimates the probability of a data point belonging to a particular class using the logistic function. It is extended to handle multi-class problems, such as iris flower classification, by applying techniques like one-vs-rest. This method is computationally efficient and offers interpretable results.

After both models are trained, their performance is evaluated and compared using standard metrics such as accuracy, precision, recall, and F1-score. This comparison highlights the strengths and limitations of each model and helps determine which performs better in classifying the iris dataset. Analyzing these results is key to selecting the most effective approach. Finally, the trained Logistic Regression model is used to predict the species of iris flowers in a separate test dataset. By feeding unseen data into the model, predictions are generated and compared with actual labels to assess accuracy. This final step evaluates how well the model generalizes to new data, providing insights into its real-world applicability.

### 3.1 KNN

The K-Nearest Neighbor (K-NN) algorithm is one of the simplest and most intuitive machine learning techniques, primarily used for classification tasks, though it can also be applied to regression problems. As a supervised learning algorithm, K-NN classifies new data points based on their similarity to existing labeled data. It operates under the assumption that similar data points exist in close proximity in the feature space. Notably, K-NN is a non-parametric and lazy learning algorithm, meaning it does not make assumptions about the data distribution and does not learn a model during the training phase.

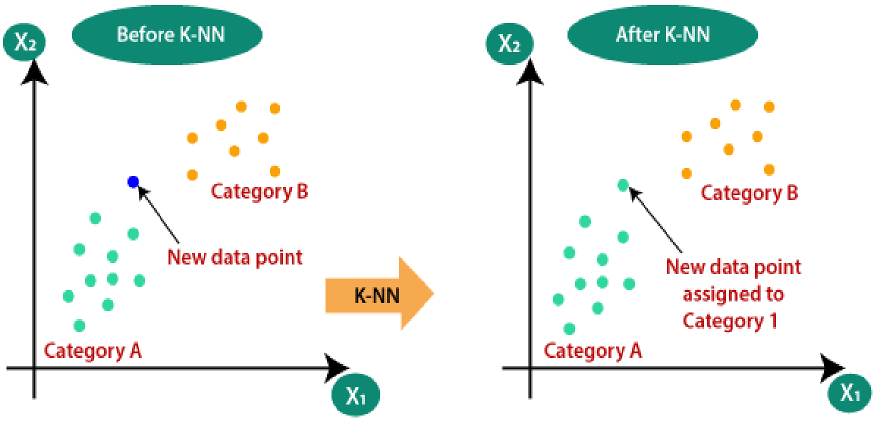


Fig. 2: Working of KNN algorithm.

Instead, it simply stores the dataset and defers the computation until prediction time, where it determines the class of a new data point by evaluating its K closest neighbors, typically using Euclidean distance as the similarity metric. The algorithm works by first selecting the number of neighbors (K), calculating the distance between the new point and all existing data points, identifying the K closest ones, counting how many of those neighbors belong to each class, and finally assigning the class with the majority vote. For example, if three out of five nearest neighbors of a new image resemble cats and two resemble dogs, the algorithm will classify the image as a cat. Choosing the right value of K is essential for optimal performance; a small K (e.g., 1 or 2) may be sensitive to noise and outliers, while a very large K may oversimplify the classification and cause difficulties. Although there is no fixed rule for selecting K, a common practice is to test several values and choose the one that yields the best accuracy, with K=5 being a commonly used starting point.

### 3.2 Logistic Regression

Logistic regression is a supervised learning algorithm used mainly for binary classification problems, where it predicts the probability of an outcome belonging to one of two classes (e.g., 0 or 1) using the sigmoid or logistic function. This function maps input values to a range between 0 and 1, and based on a threshold (commonly 0.5), the model classifies the input into Class 0 or Class 1. Unlike linear regression, logistic regression outputs probabilities, not continuous values. It can also handle multiclass problems through multinomial or ordinal logistic regression. The model assumes a linear relationship between the independent variables and the log-odds of the outcome, and requires that data points are independent, without outliers, and that the sample size is large. Key concepts include independent variables (predictors), the dependent variable (target), odds (ratio of success to failure), log-odds (logarithm of odds), coefficients (which show the influence of predictors), and the intercept. Model parameters are estimated using maximum likelihood estimation to find the best-fitting logistic curve. Overall, logistic regression is a powerful, interpretable, and widely used method for classification tasks.

## 4.RESULTS AND DISCUSSION

Implementing a machine learning approach to classify iris flowers based on petal and sepal measurements is essential for various applications in botany, horticulture, agriculture, and environmental science. This method involves utilizing a dataset containing measurements of iris flowers' petal length, petal width, sepal length, and sepal width, along with their corresponding species labels. The dataset is typically divided into training and testing sets to train and evaluate machine learning models.

### 4.1 Dataset description

The Iris dataset is a popular and commonly used dataset in machine learning and scientific research. It contains information about 150 iris flowers, with each flower described by its sepal length, sepal width, petal length, petal width (all in centimeters), and the species it belongs to. The three species in the dataset are Iris-setosa, Iris-versicolor, and Iris-virginica. The sepal is the outer part of the flower, and the petal is the inner part—both can vary in size and shape depending on the species. These measurements are used as features to help predict the species of each flower. This dataset is useful for building and testing classification models that can identify the species based on these physical characteristics. Researchers use it to find patterns in the data, develop predictive models, and compare the performance of different machine learning algorithms. It also helps scientists study the differences and similarities between iris species, giving insight into their evolution and natural environment. Because of its simple structure and clear categories, the Iris dataset is a valuable tool in both education and research in fields like botany, ecology, and artificial intelligence.

### 4.2 Results description

Fig. 3 shows confusion matrix for a K-nearest neighbors (KNN) classifier applied to the Iris flower dataset. The Iris dataset is a classic dataset used in machine learning that contains 150 samples from three species of iris flowers: Iris-setosa, Iris-versicolor and Iris-virginica. Each flower is described by four features: sepal length, sepal width, petal length, and petal width. The confusion matrix shows the number of correctly and incorrectly classified flower samples by the KNN model. Here's a breakdown of the information in the confusion matrix: Rows represent the actual iris species (True Class). Columns represent the iris species predicted by the KNN model (Predicted Class).

Fig. 4 shows a confusion matrix, which is a 3x3 table used to evaluate the performance of a logistic regression model in classifying Iris flowers into three species: Iris-setosa, Iris-versicolor, and Iris-virginica. The rows represent the actual flower species, while the columns represent the species predicted by the model. The diagonal elements indicate correct predictions—for instance, the model correctly predicted all 10 Iris-setosa flowers, 9 Iris-versicolor flowers, and 6 Iris-virginica flowers. The off-diagonal elements show misclassifications, such as one Iris-versicolor flower incorrectly predicted as Iris-virginica and four Iris-virginica flowers wrongly classified as Iris-versicolor. Overall, the confusion matrix indicates that the logistic regression model is performing well, with most flowers accurately classified and only a few misclassifications, especially among the Iris-versicolor and Iris-virginica classes.

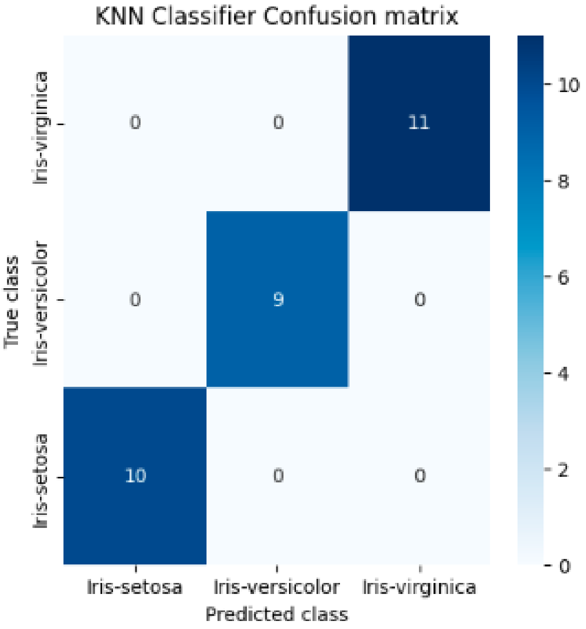


Fig. 3: Confusion matrix obtained using KNN Classifier.

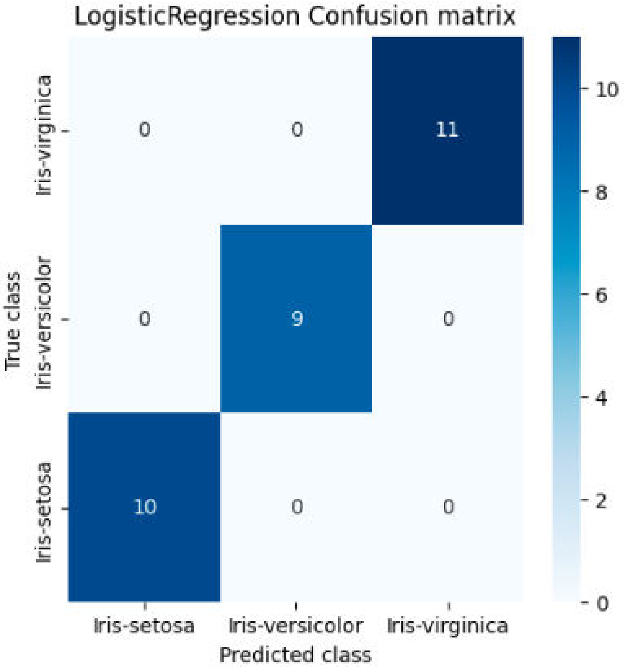
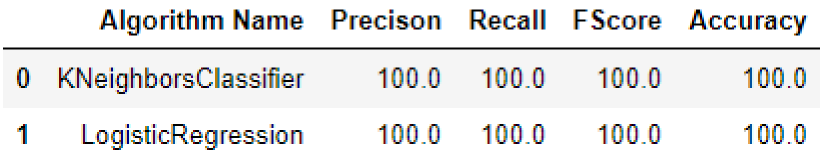


Fig. 4: Confusion matrix obtained using Logistic Regression.



| | Algorithm Name | Precison | Recall | FScore | Accuracy |
|---|---|---|---|---|---|
| 0 | KNeighborsClassifier | 100.0 | 100.0 | 100.0 | 100.0 |
| 1 | LogisticRegression | 100.0 | 100.0 | 100.0 | 100.0 |

Fig. 5: Performance Comparison of algorithms.

Fig. 5 shows that both Logistic Regression and KNeighbors Classifier achieved 100% accuracy across all evaluation metrics, including precision, recall, F1-score, and overall accuracy, indicating that both models performed equally well on the Iris dataset. However, Logistic Regression is chosen as the proposed algorithm due to certain limitations of KNN. KNN can become computationally expensive when dealing with large datasets because it requires comparing each new data point to every point in the training set, which slows down prediction time. In contrast, Logistic Regression is more efficient as it only requires applying a learned equation to make predictions. Additionally, while KNN can handle non-linear relationships, it is more sensitive to noisy data and irrelevant features. Logistic Regression, though best suited for linearly separable data, is more straightforward to interpret and scales better, making it a more practical choice for this specific task.

## 5.CONCLUSION

In conclusion, using machine learning to classify iris flowers based on petal and sepal measurements offers a faster, more accurate alternative to traditional methods. This approach helps in identifying species more easily in botany and is also useful in horticulture, agriculture, and environmental science. By using supervised learning, the system can learn patterns on its own, reducing the need for manual work and expert input. It also handles large datasets well and can recognize even small differences between species. With tools like cross-validation and hyperparameter tuning, the model becomes more reliable and accurate. This is especially helpful in areas like conservation, where knowing the correct species is important. Also, the success of these machine learning methods shows they can be used in many other fields like healthcare, finance, and marketing. Overall, this method improves both speed and accuracy in classifying iris flowers and shows the wide potential of machine learning in many areas.

## REFERENCES

[1]. Ziauddin Ursani and David W. Corne , “A Novel Nonlinear Discriminant Classifier Trained by an Evolutionary Algorithm” DOI: 10.1145/3195106.3195132

[2]. Detlef Nauck and Rudolf Kruse, “NEFCLASS-A Neuro-Fuzzzy approach for the classification of data” DOI: 10.1145/315891.316068

[3] Jing FENG, Zhiwen WANG, Min ZHA and Xinliang CAO, “Flower Recognition Based on Transfer Learning and Adam Deep Learning Optimization Algorithm”. DOI: 10.1145/3366194.3366301

[4] Roung– Guo Huang, Sang-Hyeon Jin, Jung –Hyun Kim and KwangSeck Hong, “Flower Image Recognition Using Difference Image Entropy”. DOI: 10.1145/1821748.1821868

[5] Shilpi Jain, V Poojitha, “By Using Neural Network Clustering tool in MATLAB Collecting the IRIS Flower”, Proc. IEEE , vol. 109, 2020.

[6] M. M. Mijwil and R. A. Abttan, “Utilizing the Genetic Algorithm to Pruning the C4. 5 Decision Tree Algorithm,” Asian J. Appl. Sci. ISSN 2321– 0893, vol. 9, no. 1, 2021.

[7] Roung– Guo Huang, Sang-Hyeon Jin, Jung –Hyun Kim and Kwang- Seck Hong, "Flower Image Recognition Using Difference Image Entropy". DOI: 10.1145/1821748.1821868 Academic Journal of Nawroz University (AJNU), Vol.11, No.4, 2022

475

[8] K R Rathy, Arya Vaishali, "Classification of Dataset using Efficient Neural Fuzzy Approach", vol. 099, August 2019.

[9] D. Decoste, E. Mjolsness. 2001. "State of the art and future prospects by using Machine Learning", vol. 320, 2013.

[10] Y. Lakhdoura and R. Elayachi, "Comparative Analysis of Random Forest and J48 Classifiers for 'IRIS' Variety Prediction," Glob. J. Comput. Sci. Technol., 2020

[11] Zebari, D. A., Abrahim, A. R., Ibrahim, D. A., Othman, G. M., & Ahmed, F. Y. (2021). Analysis of Dense Descriptors in 3D Face Recognition. In *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)* (pp. 171-176). IEEE.

[12] Abdulqadir, H. R., Abdulazeez, A. M., & Zebari, D. A. (2021). Data mining classification techniques for diabetes prediction. Qubahan Academic Journal, 1(2), 125-133.

[13] Ibrahim, D. A., Zebari, D. A., Ahmed, F. Y., & Zeebaree, D. Q. (2021, November). Facial Expression Recognition Using Aggregated Handcrafted Descriptors based Appearance Method. In 2021 IEEE 11th International Conference on System Engineering and Technology (ICSET) (pp. 177-182). IEEE.