Prediction Model Based on Iris Dataset Using Machine Learning Models: A Review

Arpan G Singh[1], Madhuvan Dixit[2]
[1]Research Scholar, [2]Head and Professor
Computer Department of IT, NIIST, Bhopal, Madhya Pradesh, India

Abstract: The development of a prediction model based on the well-known Iris dataset using various machine learning algorithms. The primary objective is to accurately classify the species of iris flowers—Setosa, Versicolor, and Virginica—based on four key features: sepal length, sepal width, petal length, and petal width. To achieve this, several machine learning models, including logistic regression, decision tree, k-nearest neighbors (KNN), support vector machine (SVM), and random forest, were implemented and evaluated. The dataset was preprocessed to ensure data quality and split into training and testing sets for performance validation. Model performance was assessed using accuracy, precision, recall, and F1-score metrics. Among the tested models, the random forest and SVM classifiers demonstrated superior performance, achieving high accuracy in predicting the correct iris species. This study highlights the effectiveness of machine learning techniques in handling classification tasks and provides insights into model selection for similar predictive analytics problems in real-world applications.

Keywords: Machine Learning, Iris Flower, Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, SVM.

How to cite this article: Arpan G Singh, Madhuvan Dixit. (2026). Prediction Model Based on Iris Dataset Using Machine Learning Models: A Review, International Journal of Scientific Modern Research and Technology (IJSMRT), ISSN: 2582-8150, Volume-23, Issue-02, Number-01, April-2026, pp.01-07, URL: https://www.ijsmrt.com/wp-content/uploads/2026/06/IJSMRT-26040201.pdf

IJSMRT-26040201

I. INTRODUCTION

The Iris dataset, one of the most well-known datasets in machine learning, is widely used for classification tasks due to its simplicity and well-structured features. This dataset consists of 150 samples of iris flowers, categorized into three species: Setosa, Versicolor, and Virginica, with four measured attributes—sepal length, sepal width, petal length, and petal width. A prediction model based on this dataset leverages machine learning algorithms to classify the species of an iris flower based on its given attributes. Various classification models, such as Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Artificial Neural Networks (ANN), can be applied to train and evaluate the dataset for accurate species prediction. By utilizing techniques such as data preprocessing, feature selection, and hyper parameter tuning, machine learning models can achieve high accuracy in classifying iris species. The implementation of such models demonstrates the power of machine learning in pattern recognition and classification problems, making the Iris dataset an ideal benchmark for testing and comparing different algorithms.

The development of a prediction model based on the Iris dataset involves several crucial steps, including data preprocessing, exploratory data analysis (EDA), model selection, training, evaluation, and optimization. Initially, the dataset is cleaned and visualized to understand the distribution of features and identify any potential outliers. Feature engineering techniques may also be applied to enhance model performance. Following this, different machine learning models are trained on the dataset, each with unique strengths and weaknesses in handling classification tasks. Performance metrics such as accuracy, precision, recall, and F1-score are used to evaluate and compare these models. Additionally, hyperparameter tuning techniques like grid search or random search can be employed to optimize model parameters and improve predictive accuracy. The Iris dataset, despite its small size, serves as an excellent case study for understanding fundamental machine learning concepts and their practical applications. This research not only highlights the effectiveness of various classification models but also provides insights into how machine learning can be applied to real-world classification problems, paving the way for more complex predictive modeling tasks.

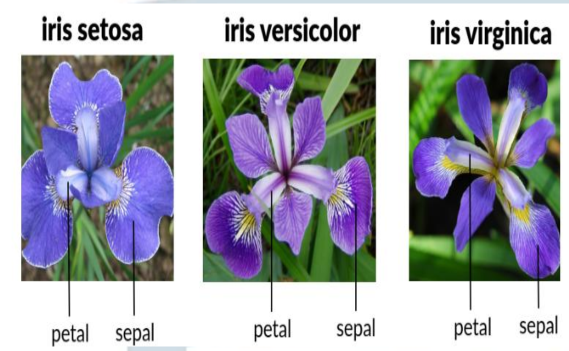

Figure 1: Iris Dataset and Analysis using Machine learning techniques

Iris Dataset

The Iris dataset is a widely used dataset in machine learning and statistics, often employed for classification and clustering tasks. Introduced by British statistician and biologist Ronald Fisher in 1936, it consists of 150 samples of iris flowers from three species: Iris setosa, Iris versicolor, and Iris virginica. Each sample is described by four numerical features—sepal length, sepal width, petal length, and petal width—making it an excellent dataset for exploring supervised learning algorithms. Due to its small size, simplicity, and clear distinctions between classes, the Iris dataset is frequently used for educational purposes, benchmarking machine learning models, and demonstrating data visualization techniques.

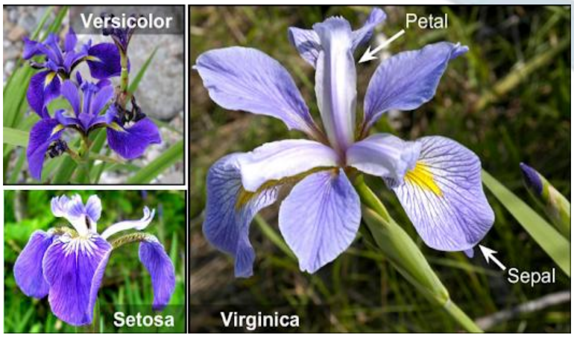

Figure 2: Iris dataset

In fact, three of these iris species look similar, but the difference in measurements can be used to classify them. This data set is a classic example of supervised learning. The input variables are sepal length and width and petal length and width; each row represents an instance or observation. The output variable is Iris-setosa, Iris-versicolor, or Iris-virginica; each column represents a class label.

Machine Learning

Machine learning is a branch of artificial intelligence that enables computers to learn from data and make predictions or decisions without being explicitly programmed. It involves the development of algorithms that identify patterns in data and improve their performance over time. Machine learning is widely used in various applications, such as image recognition, speech processing, fraud detection, and recommendation systems. It can be categorized into supervised learning, where models are trained on labeled data, unsupervised learning, which finds hidden patterns in unlabeled data, and reinforcement learning, where an agent learns by interacting with an environment to maximize rewards. As machine learning continues to evolve, it is driving advancements in automation, healthcare, finance, and many other industries.

The rapid growth of machine learning is fueled by the increasing availability of big data, improvements in computational power, and advancements in algorithms. Deep learning, a subset of machine learning inspired by the human brain's neural networks, has significantly improved the accuracy of tasks such as natural language processing, image recognition, and autonomous driving. However, machine learning also faces challenges, including the need for large datasets, potential biases in algorithms, and ethical concerns related to privacy and decision-making transparency. Researchers are continuously working on making models more efficient, interpretable, and fair. As the field progresses, machine learning is expected to play a crucial role in shaping the future of technology, driving innovation in fields like robotics, personalized medicine, and smart systems.

II. LITERATURE REVIEW

Yahaya Nayaya Tahir et.al (2024) - The Anfis model trained on the Iris Plant dataset exhibits a promising ability to capture complex relationships within the data, as evidenced by its strong correlation values of 0.965 for training and 0.961 for testing. The utilization of Gaussian membership functions and hybrid optimization methods enhances the model's adaptability to continuous data and contributes to efficient training. However, the slightly lower correlation on the testing dataset implies a potential issue of overfitting, where the model may be memorizing the training data rather than generalizing well to unseen instances. To address this concern and enhance the model's robustness, further investigation into regularization techniques or model refinement may be necessary [01].

Jenan Jader Msaad et.al (2024) - Classification is the task of assigning object to one of several predefined categories. Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set to train the model where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model to classify the target,. The classification analysis would generate a model that could be used to find a class (target). A neural network consists of patterns represented in terms of numerical values attached to the nodes of the graph and transformations between patterns achieved via simple message-passing algorithms. Certain of the nodes in the graph are generally distinguished as being input nodes or output nodes, and the graph as a whole can be viewed as a representation of a multivariate function linking inputs to outputs. Numerical values (weights) are attached to the links of the graph, parameter zing the input/output function and allowing it to be adjusted via a learning algorithm [02].

Chya Fatah Aziz et.al. (2023) - Supervised Machine Learning algorithm has an important approach to Classification. We are predicting the deal type of the Iris plant using various algorithms of machine learning. Iris plants are determined by numerous factors such as the size of the length and width of the property. A horticultural skill announces that some of the plants are different in some physical appearances like size, shape, and color. Hence it is difficult to recognize any species. Versicolor, Setosa, and Virginica have three identical subspecies of The Iris flower species. This paper uses machine learning algorithms to recognize all classes of the flower with an accuracy degree of %100 for KNN, %95 for RF, %97 for DT, and %98 for LR. The Iris dataset is frequently available, and it is implemented using Scikit tools. and build the prediction model for Plants. Here, algorithms of machine learning such as Logistic Regression (LR), Decision Tree (DT), K Nearest Neighbor (KNN), and Random Forest (RF) are employed to construct a predictive model [03].

Zahraa Faiz Hussain et.al (2023) - Data mining is known as the process of detection concerning patterns from essential amounts of data. As a process of knowledge discovery. Classification is a data analysis that extracts a model which describes an important data classes. One of the outstanding classifications methods in data mining is support vector machine classification (SVM). It is capable of envisaging results and mostly effective than other classification methods. The SVM is a one technique of machine learning techniques that is well known technique, learning with supervised and have been applied perfectly to a vary problems of: regression, classification, and clustering in diverse domains such as gene expression, web text mining. In this study, A newly mode for classifying iris data set using SVM classifier and genetic algorithm to optimize c and gamma parameters of linear SVM, in addition principle components analysis (PCA) algorithm was use for features reduction [04].

Ferdi Özbilgin et.al (2023) - Coronary Artery Disease (CAD) occurs when the coronary vessels become hardened and narrowed, limiting blood flow to the heart muscles. It is the most common type of heart disease and has the highest mortality rate. Early diagnosis of CAD can prevent the disease from progressing and can make treatment easier. Optimal treatment, in addition to the early detection of CAD, can improve the prognosis for these patients. This study proposes a new method for non-invasive diagnosis of CAD using iris images. In this study, iridology, a method of analyzing the iris to diagnose health conditions, was combined with image processing techniques to detect the disease in a total of 198 volunteers, 94 with CAD and 104 without. The iris was transformed into a rectangular format using the integral differential operator and the rubber sheet methods, and the heart region was cropped according to the iris map. Features were extracted using wavelet transform, first-order statistical analysis, a Gray-Level Co-Occurrence Matrix (GLCM), and a Gray Level Run Length Matrix (GLRLM). The model's performance was evaluated based on accuracy, sensitivity, specificity, precision, score, mean, and Area Under the Curve (AUC) metrics. The proposed model has a 93% accuracy rate for predicting CAD using the Support Vector Machine (SVM) classifier. With the proposed method, coronary artery disease can be preliminarily diagnosed by iris analysis without needing electrocardiography, echocardiography, and effort tests. Additionally, the proposed method can be easily used to support telediagnosis applications for coronary artery disease in integrated telemedicine systems [05].

M. Hanefi Calp et.al (2023) - Plant species were classified on the Iris dataset using Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), and

K-Means algorithms. In this process, models were developed for each method, success rates were obtained, and a model with a minimum error rate was introduced. The dataset of the study was obtained from the Kaggle website. The classification process was applied repeatedly on the iris dataset, and the classification or prediction with the minimum error rate was aimed at the established models. In the study process, first of all, the dataset was obtained, prepared, and visualized. Models were created using the Jupiter Notebook editor via the Anaconda desktop GUI. Then, the models were analyzed and the most successful algorithm was selected. As a result, according to the prediction/classification models, it was seen that the most successful model was obtained with the KNN algorithm, and the most unsuccessful model was obtained with the ANN algorithm [06].

Jie Sun et.al (2022) - The open-set recognition of irises and proposes a deep learning-based open-set iris recognition method. This method introduces the distance factor in the network structure and loss function and achieves open-set iris recognition through two pieces of training. We use the iris datasets CASIA-Iris-Twins and CASIA-Iris-Lamp to construct an open iris dataset and conduct experiments on the existing iris recognition algorithm and open-set image recognition method. The necessity and feasibility of open-set iris recognition are verified. Moreover, experiments show that the proposed method has good open-set iris recognition performance, can effectively distinguish unknown classes of iris samples, and has little effect on the recognition ability of known class samples [07].

Andrey Kuehlkamp et.al (2022) - Iris recognition of living individuals is a mature biometric modality that has been adopted globally from governmental ID programs, border crossing, voter registration and de-duplication, to unlocking mobile phones. On the other hand, the possibility of recognizing deceased subjects with their iris patterns has emerged recently. In this paper, we present an end-to-end deep learning-based method for postmortem iris segmentation and recognition with a special visualization technique intended to support forensic human examiners in their efforts. The proposed postmortem iris segmentation approach outperforms the state of the art and – in addition to iris annulus, as in case of classical iris segmentation methods – detects abnormal regions caused by eye decomposition processes, such as furrows or irregular specular highlights present on the drying and wrinkling cornea. The method was trained and validated with data acquired from 171 cadavers, kept in mortuary conditions, and tested on subject-disjoint data acquired from 259 deceased subjects. To our knowledge, this is the largest corpus of data used in postmortem iris recognition research to date [08].

Swathi Gowroju, et. al (2022) - Authors presented Iris recognition is a secure and best-chosen biometric application in the digital world because of its unique characteristics. Day by day, the digital world plays a significant role in human life for various applications. The applications are vastly spread over secure applications of the nation such as border control applications, criminal investigations, postmortem studies, access the digital equipment, smart homes, smart appliances, smart cars etc. Due to the digitalization of the world, all the research communities, scientists, and industries are focusing on the biometric-based secured iris recognition system. The researcher has implemented various algorithms based on traditional and neural network architectures. In this scenario, this paper gives a brief on different techniques and algorithms used by researchers to predict the age of human people using the iris [09].].

Table 1: Cooperative analysis of previous Method

Author(s) & Year	Title / Study Focus	Methodology / Model Used	Dataset / Platform
Hassan Shabani Mputu et al. (2024)	Tomato quality sorting and grading using hybrid deep learning and machine learning models	Pre-trained CNNs used for feature extraction; traditional ML classifiers (SVM, RF, KNN) for classification. Best hybrid model: InceptionV3 + SVM	Tomato image dataset captured using NVIDIA Jetson TX1; also tested on a public tomato dataset
Olarewaju Mubashiru Lawal et al. (2024)	Lightweight fruit detection for real-time low-power devices	Developed lightweight models YOLO-Lite, YOLO-Liter, YOLO-Litest based on YOLOv5 framework	Tested on low-power computing platforms / computer platform
Umer Amin et al. (2023)	Automated fruit freshness classification using transfer learning	Fine-tuned AlexNet CNN with transfer learning; adjusted learning rate, batch size, and hyperparameters	Three publicly available fruit datasets

Harmandeep Singh Gill et al. (2023)	Fruit recognition and classification using deep learning and feature-based classification	Used CNN, RNN, and LSTM for feature extraction/selection; compared classifiers SVM, ANFIS, FFNN	Fruit image datasets (not explicitly detailed in summary)
Linhui Wang et al. (2023)	Portable intelligent citrus pest detection system using lightweight deep learning	Improved SSD model with optimized feature extraction and miniaturized prediction convolution kernels; proposed MobileNetV3 + RPBM	Portable embedded device for citrus orchard pest detection
Bindu Puthentharayil Vikraman et al. (2022)	Transfer-learning-based automatic date fruit classification system with electromechanical sorting	Transfer learning-based image classification, sensor-triggered image capture, microcontroller, MIT App, and piston-based sorting mechanism	Conveyor-belt based date fruit sorting system with electromechanical hardware

III. METHOD

Introduction to the Iris Dataset

The Iris dataset is composed of 150 samples divided into three distinct classes: Setosa, Versicolor, and Virginica. Each sample has four features (Sepal Length, Sepal Width, Petal Length, and Petal Width) that describe the physical characteristics of the flowers. The dataset is considered ideal for machine learning tasks due to its well-balanced class distribution and the absence of missing data, making it a great candidate for model development. The purpose of the Iris Species Prediction System is to use these features to predict which species the flower belongs to based on user input. This prediction system employs a variety of machine learning models, such as Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), each offering distinct advantages depending on the nature of the data.

End-to-End Workflow

The Iris Species Prediction System follows a structured and systematic approach, divided into various stages. These stages include data pre-processing, model selection, prediction generation, and result presentation.

IV. CONCLUSION

The Iris Species Prediction System achieved exceptional results, with all three models performing at a high level of accuracy. The seamless integration of preprocessing, model training, and prediction display ensures a user-friendly and efficient experience. Among the models, SVM emerged as the most reliable, making it the recommended choice for future applications of this system.

The prediction model based on the Iris dataset using machine learning techniques effectively classifies iris species with high accuracy. By leveraging algorithms such as logistic regression, decision trees, support vector machines, and neural networks, the model demonstrates the power of supervised learning in pattern recognition and classification tasks. Feature selection, data preprocessing, and hyper parameter tuning further enhance the model's performance. The results validate the applicability of machine learning in botanical classification and suggest potential extensions to other datasets and real-world applications, such as automated species identification in ecological research and agriculture

The success of this model highlights the importance of choosing the right algorithm based on dataset characteristics and problem requirements. The Iris dataset, being well-structured and balanced, serves as an excellent benchmark for testing and comparing various machine learning techniques. Future research can explore ensemble learning methods, deep learning approaches, and hybrid models to further improve classification accuracy. Additionally, integrating real-time classification with mobile applications or IoT devices could expand its practical applications in environmental monitoring and biodiversity conservation. Overall, this study reinforces the potential of machine learning in predictive analytics and sets the foundation for more advanced classification models in diverse domains.

The insights gained from this study can be applied to more complex datasets with higher dimensionality and real-world variability. By incorporating feature engineering techniques and advanced optimization strategies, future models can achieve even greater precision and robustness. Additionally, explainability and interpretability of machine learning models remain crucial, especially in scientific research and decision-making processes. Implementing techniques such as SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) can provide deeper insights into model predictions, fostering trust and transparency. Lastly, as machine learning continues to evolve, integrating automation and real-time learning capabilities will further enhance predictive modeling, making it a valuable tool for various fields, including agriculture, medicine, and environmental science.

V. FUTURE SCOPES

The future scope of a prediction model based on the Iris dataset using machine learning is promising, especially in the fields of botanical research, agriculture, and artificial intelligence education. By refining classification algorithms, researchers can develop more accurate and efficient models for plant species identification, which can be extended to a broader range of flora beyond the Iris dataset. Integrating deep learning techniques, such as convolutional neural networks (CNNs), can further enhance pattern recognition and species differentiation.

Additionally, advancements in edge computing and IoT devices could enable real-time plant identification in smart farming applications. The model's adaptability to more complex datasets makes it a valuable tool for machine learning practitioners, serving as a benchmark for testing new classification algorithms. Moreover, its potential integration with automated systems in environmental monitoring and biodiversity conservation could contribute to sustainable development efforts. As machine learning continues to evolve, the Iris dataset prediction model can serve as a foundation for developing more sophisticated classification frameworks in various scientific and industrial domains.

REFERENCES

[1] Yahaya Nayaya Tahir1 ; Umar Aminu1 ; Hamma Safiou “Utilizing Adaptive Neuro-Fuzzy Inference System (Anfis) For Complex Pattern Recognition And Species Prediction For Iris Dataset” Vol. 5 No. 5 September, 2024.

[2] Jenan Jader Msaad, Ruwaidah F. Obaied , Alaa Majeed Shnin “Network Classifier on Iris data Using MATLAB” Vol 8, Issue 3 (2024) P: 54 – 46, ISSN: 2522-3321.

[3] Mashael M. Khayyat, Nuha Zamzami, Li Zhang, Michele Nappi , Muhammad Umer. "Fuzzy-CNN: Improving personal human identification based on IRIS recognition using LBP features." Volume 83, June 2024, 103761.

[4] Rituparna Nath, Arunima Devi “Machine Learning Algorithms Used for Iris Flower Classification” DOI: 10.4018/979-8-3693-2260-4.ch010, 2024.

[5] Chya Fatah Aziz, Banan Jamil Awrahman “Prediction Model based on Iris Dataset Via Some Machine Learning Algorithms” VOL. 10 NO. 2,2023.

[6] Zahraa Faiz Hussain, Hind Raad Ibraheem, Mohammad Alsajri, Ahmed Hussein Ali, Mohd Arfian Ismail , Shahreen Kasim, Tole Sutikno “A new model for iris data set classification based on linear support vector machine parameter's optimization” Vol. 10, No. 1, February 2023, pp. 1079~1084 ISSN: 2088-8708, DOI: 10.11591/ijece.v10i1.pp1079-1084.

[7] Ferdi Özbilgin , Çetin Kurnaz and Ertan Aydın “Prediction of Coronary Artery Disease Using Machine Learning Techniques with Iris Analysis” Volume 13 Issue 6, 13 March 2023.

[8] M. Hanefi Calp, Vijender Kumar Solanki “Classification of Plant Species with Iris Dataset Using ANN, KNN and K-Means Algorithms” Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering pp. 7–10 DOI: 10.15439/2023R63 ISSN 2300-5963 ACSIS, Vol. 38 ,2023.

[9] Jie Sun, Shipeng Zhao, Sheng Miao, Xuan Wang and Yanan Yu “Open-set iris recognition based on deep learning” Accepted: 17 March 2022.

[10] Andrey Kuehlkamp, Aidan Boyd Adam Czajka Kevin Bowyer Patrick Flynn “Interpretable Deep Learning-Based Forensic Iris Segmentation and Recognition” 2022.

[11] Swathi Gowroju, Aarti & Sandeep Kumar "Review on secure traditional and machine learning algorithms for age prediction using IRIS image” Published: 28 June 2022,Volume 81, pages 35503–35531, (2022).

[12] Gunjan Ahuja, Muskan Aggarwal, Jashn

Tyagi, Onkar Mehra "Identification Of Different Species Of Iris Flower Using Machine Learning Algorithms" International Research Journal of Engineering and Technology (IRJET), Volume: 10 Issue: 01 | Jan 2023, ISSN: 2395-0072.

[13] Jie Sun Shipeng Zhao Sheng Miao Xuan Wang Yanan Yu "Open-set iris recognition based on deep learning" 17 March 2022, 2022;16:2361–2372.

[14] Bahzad Taha Chicho, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, Dilovan Assad Zebari "Machine Learning Classifiers Based Classification For IRIS Recognition" VOL. 1 NO. 2 (2021).

[15] Smita Khade; Shilpa Gite; Sudeep D. Thepade; Biswajeet Pradhan; Abdullah Alamri "Detection of Iris Presentation Attacks Using Hybridization of Discrete Cosine Transform and Haar Transform With Machine Learning Classifiers and Ensembles" Volume: 9, ISSN: 2169-3536, 24 December 2021.

[16] Zahra Rahim Samia, Huda Kadhim Tayyeh, Mohammed Salih Mahdic "Survey of Iris Recognition using Deep Learning Techniques" Journal of Al-Qadisiyah for Computer Science and Mathematics Vol. 13(3) 2021 , pp Comp. 47–56 47, 08 /08/2021.