Proceedings of 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS 2024)
IEEE Xplore Part Number: CFP24UD1-ART; ISBN: 979-8-3503-7274-8

2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS) | 979-8-3503-7274-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC-ROBINS60238.2024.10534032

Enhancing Accuracy: Iris Flower Classification with Ensemble Models

Nishanthini Mani
Department of Artificial Intelligence and Data Science
KPR Institute of Engineering and Technology
Coimbatore, TamilNadu, India
nishanthinimani525@gmail.com

Premkumar Murugiah
Department of Artificial Intelligence and Data Science
KPR Institute of Engineering and Technology
Coimbatore, TamilNadu, India
premkumarmurugiah1395@gmail.com

Vethavarna Veeraraghavan
Department of Artificial Intelligence and Data Science
KPR Institute of Engineering and Technology
Coimbatore, TamilNadu, India
varna.vgv@gmail.com

Nuzha Razia
Department of Artificial Intelligence and Data Science
KPR Institute of Engineering and Technology
Coimbatore, TamilNadu, India
nuzha.razia@gmail.com

Shashank Ramesh
Software Engineer 2
NextLabs Pte. Ltd.
Singapore
shashankramesh2@gmail.com

ABSTRACT:
This research study investigates the classification of Iris flowers based on their morphological structures, addressing the challenges posed by variations in attributes like size, shape, and color. This study explores various classification techniques and their practical implementations by conducting a comparative analysis using the IRIS dataset. With 21 attributes and three species (Setosa, Versicolor, and Virginica), each comprising 400 samples, this research aims to leverage machine learning algorithms to achieve precise classification. Key steps include data preprocessing, model selection, and hyperparameter tuning, with evaluation metrics to enhance the model performance. Furthermore, this study also explores ensemble-based learning to improve the prediction accuracy. Comparative analysis reveals the accuracy of each model, demonstrating effective predictions with accuracy rates of 95.5% and 97%. In summary, this research study offers a comprehensive overview of Iris flower species classification and prediction using machine learning techniques, contributing to advancements in this domain.

Keywords – IRIS dataset, KNN, Decision Tree, Random Forest, Logistic Regression, Naive Bayes, Ensemble Learning.

I. INTRODUCTION

This research study focuses on the classification of Iris flowers using various machine learning techniques, aiming to accurately determine their species based on identifiable features.

The dataset used, Iris_extended, provides a detailed insight into the botanical attributes of Iris flowers, with samples from three species: Setosa, Versicolor, and Virginica, each comprising 400 samples.

In the realm of machine learning, precise classification of botanical species is crucial. This study employs machine learning models such as K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB) to classify Iris flowers.

Additionally, ensemble techniques are utilized to enhance the accuracy rate of the classification results. The implementation of these machine learning models is done by using prominent tools and libraries such as scikit-learn, NumPy, and pandas.

II. LITERATURE REVIEW

The Iris dataset, well known for its relevance in machine learning domain was introduced by the renowned statistician and biologist Sir Ronald A Fisher in 1936. Fisher used it to demonstrate discriminant analysis, which later paved way for practical applications of statistical methods for classification in the field of machine learning and statistics. Various prominent studies and research projects have been done in this domain and the work have drawn inspiration from a range of studies in this field.

Rana et al. (2020) conducted a comparative study on IRIS flower classification by comparing the performance of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) with Logistic Regression and Random Forest algorithms. The work has concluded that LDA has outperformed PCA, with Logistic Regression and Random Forest yielding similar results when using LDA. However, the study focused solely on IRIS flower classification to limit the generalizability of the findings to other datasets or domains. Additionally, the impact of hyperparameter tuning on algorithm performance was not explored as it was potentially overlooking opportunities for optimization [1].

Mithy, S. A., et al. (2022) explored the classification of the Iris flower dataset using various machine learning algorithms such as SVM, KNN, Logistic Regression, and Random Forest. Their study demonstrated high accuracy rates with SVM. However, the study lacked exploration of more advanced algorithms or ensemble methods, which could further improve the classification accuracy [2].

Bahzad Taha Chicho et al. (2021) demonstrated the effectiveness of various machine learning classifiers for IRIS recognition, emphasizing accurate classification and the significance of choosing the right algorithm for performing specific tasks [3].

Yihan Zhou (2022) compared machine learning models on the Iris dataset, with MLP demonstrating the highest accuracy. While MLP excelled in handling nonlinear problems, challenges such as overfitting and computational intensity were noted for other algorithms like KNN[4].

Hemlata Ohal et al. (2023) focused on using machine learning algorithms to classify IRIS flowers based on their characteristics. The work employed techniques such as SVM, Logistic Regression, and Decision Tree, achieving high accuracy in categorizing flowers. However, selecting the most suitable algorithm for flower classification remains challenging, requiring expertise and experimentation[5].

Shukla et al. (2020) conducted a study on Flower Classification using Supervised Learning, focusing on segmenting flowers from backgrounds and extracting features like color and texture for analysis. Various machine learning algorithms, including Neural Network, Logistic Regression, Support Vector Machine, and k-Nearest Neighbors are used in the study. While these algorithms demonstrated effectiveness in capturing complex patterns and providing interpretable results, challenges such as overfitting, computational intensity, and sensitivity to parameter choices were noted[6].

Bhutada et al. (2021) proposed a novel approach to classifying and recognizing flowers based on text processing. Their system allowed users to input flower details and receive information on medicinal and therapeutic uses. While two algorithms, KNN and Random Forest, were utilized for classification, the limited number of flower species in the dataset potentially restricted the system's scope and accuracy. Additionally, reliance solely on text processing for feature extraction may overlook visual characteristics crucial for enhancing flower recognition accuracy[7].

Chao Chen et al. (2020) focused on dynamic forecasting of flowering periods using multivariable LSTM and ensemble learning classification tasks. Leveraging ensemble learning strategies, the work achieved accurate local flowering predictions by combining time series and classification forecasting[8].

Silky Sachar et al. (2022) delved into deep ensemble learning for automatic identification of medicinal plant leaves, emphasizing the healing properties of plants. Their study surpassed established models in accuracy by leveraging convolutional neural networks and transfer learning[9].

Hasib Uddin et al. (2023) delved into the classification of Bangladeshi medicinal plants using deep learning models, achieving high accuracy rates. The work demonstrated the effectiveness of ensemble learning techniques in accurately classifying medicinal plant species[10].

Grewal et al. (2022) presented a comparative analysis of machine learning models, emphasizing the importance of processing raw data to reveal valuable insights. The study discussed the advantages of using Logistic Regression and Decision Tree Classifier algorithms. However, limitations such as Logistic Regression's restriction to discrete variable prediction and Decision Tree Classifier's susceptibility to overfitting were noted. The research underscored the significance of selecting the most optimal algorithm based on dataset characteristics to enhance decision-making processes in various applications[11].

ByungJoo Kim (2019) explored ensemble methods in classification problems, aiming to enhance model stability and accuracy. Their study showcased the effectiveness of ensemble models in outperforming traditional classifiers, despite potential complexity in implementation and resource requirements[12].

Zhibin Wang et al. (2022) introduced a dynamic ensemble selection technique tailored for CNNs in flower classification, optimizing classifier selection and integration. Their method enhances classification accuracy by sorting classifiers based on recognition credibility, thereby avoiding exhaustive classifier searching. However, the implementation of dynamic ensemble selection may introduce complexity due to sorting and selection algorithms[13].

Md Fozle Rabbi et al. (2023) presented an ensemble-based deep learning model, FlowerConvNet, for multi-class flower recognition. Their approach outperformed traditional CNN models in classifying diverse flower species, addressing the challenge of accurate identification using image processing techniques[14].

Computational complexity in implementing ensemble models[12], overfitting due to complex models or limited dataset[3][4][6][11] and suitable algorithm selection[5][7]8][9][10] are the common research issues identified from the literature study.

III. METHODOLOGY

Here, the ensemble learning method is used to enhance the accuracy of classifying Iris flower based on the species. Ensemble learning is a machine learning technique that involves the integration of multiple individual models to create a stronger and more accurate predictive model. The objective is to refine the classification process and achieve better accuracy in identifying different types of Iris flower. This is done by exploring various machine learning algorithms such as SVM, Logistic Regression, Decision tree, Naive Bayes, Random Forest and K neighbor classifier. The overall process flow of the work is represented in Fig. 1.

The novelty of the proposed work lies in the implementation of a two-level stacking ensemble model that effectively combines the predictive strengths of diverse base models to enhance generalization and predictive accuracy. In both the proposed models, diverse base models such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Random Forest (RF), and Naive Bayes (NB) are utilized to capture different aspects of the input data.

At the first level, each base model is individually trained on the input data to generate predictions, leveraging their unique modeling capabilities. These predictions serve as input features for a meta-model, which is introduced at the second level of the stacking ensemble. The meta-model, typically Logistic Regression (LR) or Decision Tree (DT), learns to combine the diverse predictions from the base models to make the final decision.

By incorporating predictions from multiple base models and leveraging the complementary strengths of each, the stacking ensemble aims to improve predictive accuracy and generalization compared to individual models. This approach allows the ensemble to effectively capture complex relationships within the data and make more robust predictions. Overall, the proposed work represents a novel approach to ensemble learning, offering a promising avenue for enhancing classification performance in various domains.

The software packages considered in the work are Scikit-learn which is a popular machine learning library in Python that provides tools for data pre-processing, model building, and model evaluation. This is used for implementing machine learning algorithms, ensemble techniques, and evaluation metrics. Also, Kaggle platform is utilized for practical availability and facilitating comprehensive testing of the implemented machine learning algorithms and ensemble techniques.

A) Dataset

Iris_extended dataset sourced on Kaggle *(https://www.kaggle.com/code/samybaladram/iris-dataset-extended-quick-data-exploration/input).* The dataset comprises of 1200 samples and 21 attributes with 400 samples allotted to each of three species Setosa, Versicolor and Virginica.

Each sample is associated with various attributes, including:

1. Species: The class label indicating the species of the Iris flower.
2. Soil Type: The type of soil in which the Iris flower is grown.
3. Sepal Length: The length of the sepals of the flower.
4. Sepal Width: The width of the sepals of the flower.
5. Petal Length: The length of the petals of the flower.
6. Petal Width: The width of the petals of the flower.
7. Sepal Area: The area of the sepals, calculated as length multiplied by width.
8. Petal Area: The area of the petals, calculated as length multiplied by width.
9. Sepal Aspect Ratio: The ratio of sepal length to sepal width.

10. Petal Aspect Ratio: The ratio of petal length to petal width.
11. Sepal to Petal Length Ratio: The ratio of sepal length to petal length.
12. Sepal to Petal Width Ratio: The ratio of sepal width to petal width.
13. Sepal Petal Length Difference: The difference between sepal length and petal length.
14. Sepal Petal Width Difference: The difference between sepal width and petal width.
15. Sepal Area Square Root: The square root of the sepal area.
16. Petal Area Square Root: The square root of the petal area.
17. Area Ratios: Various ratios calculated based on the areas of sepals and petals.

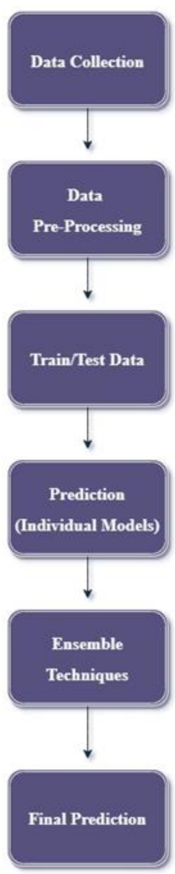

Fig. 1. Process Flow

These attributes referred as "features" are instrumental in facilitating accurate classification of Iris flowers based on the species.

B) Data Preprocessing

This step involves preprocessing the Iris_Extended dataset, that is, filling the null values and label encoding so that it doesn't pose a hindrance to the machine learning process. The null values are filled in with '-2' by employing the fillna() method in pandas and label encoding is a method by which categorical columns are converted into numerical ones so that the values can be fitted by machine learning models. Here the values 0, 1 and 2 have been assigned to Setosa, Versicolor and Virginia respectively. Elevation, petal_curvature_mm, petal_texture_trichomes_per_mm2 and leaf_area_cm2 features has been dropped as it poses very little relevance to the classification process.

From the preprocessed dataset, the attributes(features) are employed to build a model to facilitate the process of classification. In order to test whether the model is working or not, the data is split into two subsets namely training set and test set. Training dataset is the bigger subset of the dataset, which is used to train or fit the machine learning model and test data is the subset used to evaluate the model. The dataset has been split in the 70:30 ratio such that 70% of the data has been assigned to train subset and 30% to test subset. The feature Species has been assigned to Y and rest of the attributes has been assigned to X.

C) Comparative Analysis of ML Models

There are 6 machine learning algorithms used to build the proposed models. The accuracy rate of each algorithm are checked using actual and predicted values. The machine learning algorithms used are Support Vector Machine, Logistic Regression, Decision Tree, Naive Bayes, Random Forest and K-Nearest Neighbors.

D) Ensemble Learning

Ensemble learning is a machine learning technique that integrates multiple machine learning models into a single model with an aim to increase the model's performance.

There are mainly 3 kinds of ensemble learning methods. They are bagging, boosting and stacking. Stacking is employed for heterogeneous base learner type to improve the accuracy rate whereas bagging is used to reduce variance and overfitting for homogeneous base learner types and boosting is used to reduce bias for homogenous base learner types.

In this work, stacking method is implemented to improve the performance of the model as shown in Fig. 2. In stacking, multiple models are trained individually and their predicted value is given as input to a meta model. This meta model is our curated model and it gives the final output (i.e.) classification of Iris flower.

E) Proposed Models

For the Iris flower classification task, ensemble learning methods were employed to enhance the accuracy of the classification process. Specifically, the following ensemble model was considered:

Stacking is a technique where multiple models are trained individually, and their predicted values are used as input to a meta-model. This meta-model combines the predictions of the base models to generate the final output.

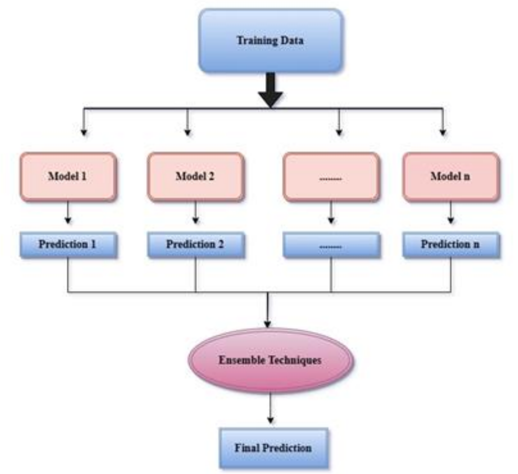

Fig. 2. Stacking Model

In this particular study, the stacking method was implemented to enhance the performance of the model, showcasing the power of ensemble techniques in surpassing the accuracy of individual machine learning models. Techniques such as bagging, boosting, and stacking were considered, with stacking being the chosen method for this specific task due to its ability to improve accuracy rates.

In constructing the proposed stacking ensemble model 1 shown in Fig. 3, Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Random Forest (RF) are incorporated as diverse base models. Each of these models was individually trained on the input data to generate predictions. Following this, a meta-model utilizing Logistic Regression (LR) was introduced in the second tier, incorporating the predictions derived from SVM, KNN, and RF as input features.

In the stacking ensemble model 2 shown in Fig. 4, Naive Bayes (NB), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) are employed as base models. Each base model was trained on the input data to make individual predictions. Subsequently, a Decision Tree (DT) meta-model was introduced at the second level, taking the predictions from NB, SVM, and KNN as input features. In both the proposed models, the meta-model was trained on the target variable, incorporating the diverse predictions from the base models. This two-level stacking approach aimed to harness the unique strengths of each base model and leverage the combined predictive power in the final decision made by the ensemble. The ensemble's performance was evaluated on a validation set, demonstrating the potential for improved generalization and predictive accuracy compared to individual models.

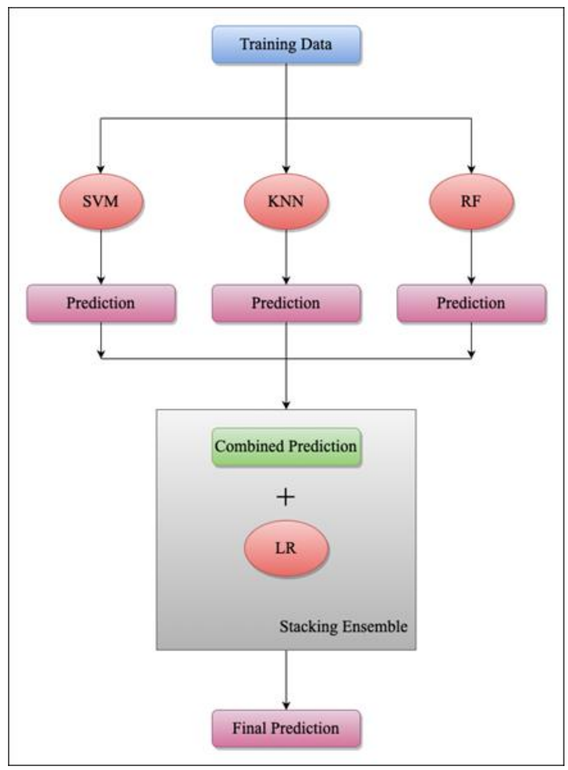

Fig. 3. Proposed Model 1

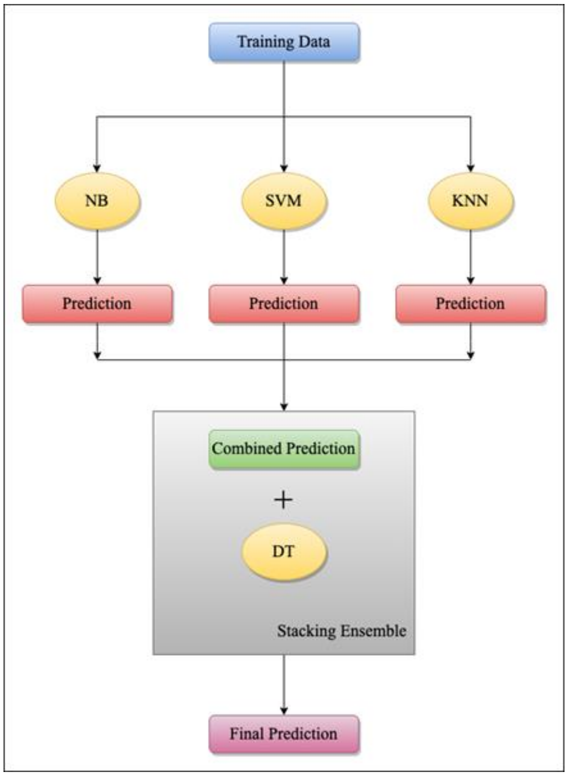

Fig. 4. Proposed Model 2

IV. RESULTS

Improved accuracy was noted with the incorporation of ensemble techniques, outperforming individual machine learning models. The experiments systematically assessed various models, such as Support Vector Machine, Naive Bayes, Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors. Particularly noteworthy were Logistic Regression and Decision Tree, emerging as the leading performers with accuracy rates of 96.47% and 96.39%, respectively, as depicted in Fig. 5.

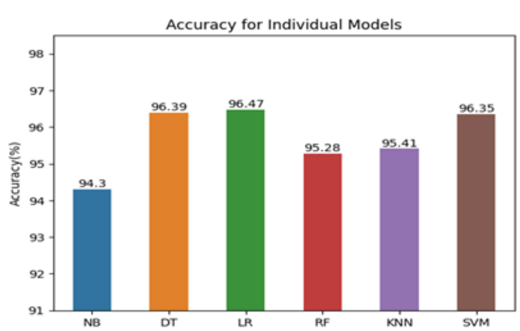

Fig. 5. Accuracy of Individual Models

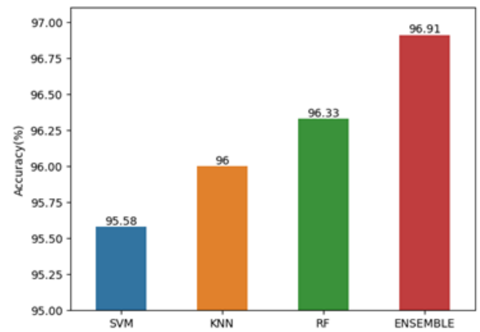

Fig. 6. Accuracy of Proposed Model 1

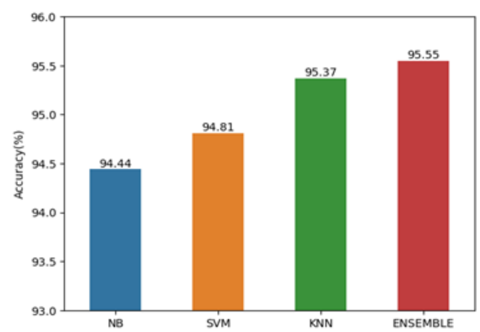

Fig. 7. Accuracy of Proposed Model 2

By employing ensemble techniques, the proposed model 1 resulted in a notable accuracy of 97% approx., as illustrated in Fig. 6, while the proposed model 2 achieved an accuracy of 95.55%, as depicted in Fig. 7. The confusion matrix of the both the proposed models are presented below in Fig. 8 and Fig. 9.

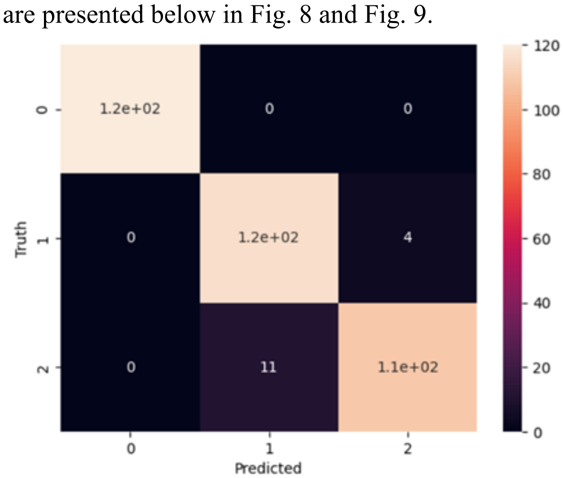

Fig. 8. Confusion Matrix of Proposed Model 1

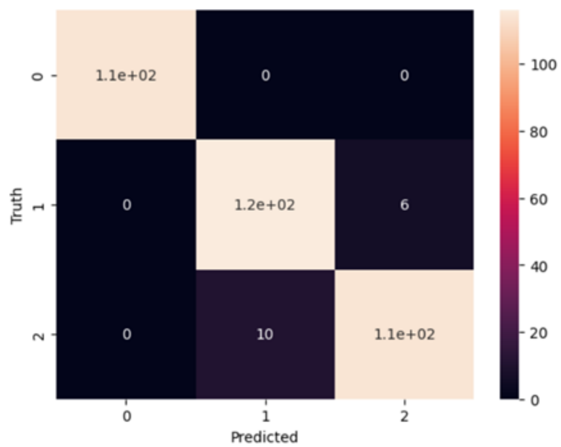

Fig. 9. Confusion Matrix of Proposed Model 2

The accuracy, precision, recall and F1-score of the proposed models are listed in the table TABLE I. Also, the accuracy of the proposed model is compared with earlier works. Though the individual models demonstrated satisfactory performance, the utilization of ensemble techniques proved effective in enhancing overall accuracy, as evidenced by the results. The observed results highlight the effectiveness of ensemble techniques in enhancing the accuracy of classification models compared to individual machine learning algorithms.

TABLE I. Evaluation Metrics

Measures	[12]	[13]	[14]	Model 1	Model 2
Accuracy	95.5	95.5	95	96.91	95.55
Precision	-	-	-	0.966	0.954
Recall	-	-	-	0.964	0.961
F1-score	-	-	-	0.969	0.955

These results represent the importance of ensemble techniques in combining the strengths of multiple models to enhance overall performance. While individual models demonstrated satisfactory performance, the utilization of ensemble techniques proved to be effective in further boosting accuracy. Ensemble methods leverage the diversity among individual models to mitigate errors and improve generalization, resulting in more robust and accurate predictions. Overall, the observed results validate the efficacy of ensemble techniques in enhancing the accuracy of Iris flower classification, providing valuable insights for future research and applications in machine learning.

In future, the study will focus on exploring novel ensemble techniques beyond traditional bagging, boosting and stacking methods. Also, research focus on selecting appropriate individual models based on the chosen domain. Generalization of developed ensemble model will also be experimented in near future.

V. CONCLUSION

This work successfully utilized ensemble learning methods by incorporating Machine Learning (ML) algorithms like Support Vector Machine, Logistic Regression, Decision Tree, Naive Bayes, Random Forest and K-Nearest Neighbors. The proposed method of employing ensemble learning significantly improved the precision of classifying Iris flowers based on its species. The effective implementation of these algorithms using scikit-learn and practical testing on the kaggle platform emphasized the reliability and efficiency of the proposed models. Overall, this work demonstrated the ability of innovative approaches in advancing machine learning capabilities with its broader applications and future exploration.

REFERENCES

[1] Rana, Debaraj, Swarna Prabha Jena, and Subrat Kumar Pradhan. "Performance comparison of PCA and LDA with linear regression and random forest for Iris flower classification." PalArch's Journal of Archaeology of Egypt/Egyptology 17, no. 9 (2020): 2353-2360

[2] Mithy, S. A., S. Hossain, S. Akter, U. Honey, and S. B. Sogir. "Classification of the Iris Flower Dataset using Different Algorithms." Int. J. Sci. Res. in (2022).

[3] Chicho, Bahzad Taha, Adnan Mohsin Abdulazeez, Diyar Qader Zeebaree, and Dilovan Assad Zebari. "Machine learning classifiers based classification for IRIS recognition." Qubahan Academic Journal 1, no. 2 (2021): 106-118.

[4] Zhou, Yihan. "Study for Iris Classification Based on Multiple Machine Learning Models." Highlights in Science, Engineering and Technology 23 (2022): 342-349.

[5] Ohal, Hemlata, Payal Dhindale, Eshwari Bhandkar, Tanusri Bokka, and Shravani Shinde. "A Comparative Analysis of Machine Learning Based Algorithms for Iris Flower Classification Application." BioGecko 12, no. 3 (2023): 639-646.

[6] Shukla, Asmita, Ankita Agarwal, Hemlata Pant, and Priyanka Mishra. "Flower classification using supervised learning." Int. J. Eng. Res 9, no. 05 (2020): 757-762.

[7] Bhutada, Sunil, K. Tejaswi, and S. Vineela. "Flower Recognition Using Machine Learning." International Journal of Researches In Biosciences, Agriculture And Technology 4, no. 2 (2021): 67-73.

[8] Chen, Chao, Xingwei Zhang, and Shan Tian. "Research on dynamic forecast of flowering period based on multivariable lstm and ensemble learning classification task." (2020).

[9] Sachar, Silky, and Anuj Kumar. "Deep ensemble learning for automatic medicinal leaf identification." International Journal of Information Technology 14, no. 6 (2022): 3089-3097.

[10] Uddin, A. Hasib, Yen-Lin Chen, Bijly Borkatullah, Mst Sathi Khatun, Jannatul Ferdous, Prince Mahmud, Jing Yang, Chin Soon Ku, and Lip Yee Por. "Deep-learning-based

classification of Bangladeshi medicinal plants using neural ensemble models." Mathematics 11, no. 16 (2023): 3504.

[11] Grewal, Prateek, Prateek Sharma, Anu Rathee, and Shikha Gupta. "Comparative Analysis of Machine Learning Models." EPRA International Journal of Research and Development (IJRD) 7, no. 6 (2022): 62-75.

[12] Kim, ByungJoo. "Ensemble methods applied to classification problem." International Journal of Internet, Broadcasting and Communication 11, no. 1 (2019): 47-53.

[13] Wang, Zhibin, Kaiyi Wang, Xiaofeng Wang, Shouhui Pan, and Xiaojun Qiao. "Dynamic ensemble selection of convolutional neural networks and its application in flower classification." International Journal of Agricultural and Biological Engineering15, no. 1 (2022): 216-223.

[14] Rabbi, Md Fozle, Sraboni Das Dola, Noorjahan Binte Kader Taha, Tamim Al Mahmud, Muhammad Nazrul Islam, Md Abdur Razzaque, and Faiz Al Faisal. "An Ensemble-based Deep Learning Model for Multi-class Flower Recognition." In 2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM), pp. 1-6. IEEE, 2023.