Academic Journal of Nawroz University (AJNU), Vol.11, No.4, 2022
 e-ISSN: 2520-789X
https://doi.org/10.25007/ajnu.v11n4a1636

An Approach based Iris Flower Species Recognition Using Machine Learning Classifiers

Amar Yehya Hussien[1]
[1]Department of Statistics, Van Yuzuncu Yil University

ABSTRACT

For data instances, classification is used to forecast group membership. Techniques for machine learning are being introduced to make the classification problem simpler. The development of a model to categorize Iris blossoms using an Artificial Neural Network (ANN) and Support Vector Machine (SVM) is the main goal of this paper. The Iris flower data set is an example of a multivariate data set. It was first presented by Ronald Fisher, a British statistician and biologist. Multi-parameter analysis of taxonomic is considered a main problem of iris flower recognition. The difficulty is in differentiating between the iris flower species (Setosa, Versicolor, and Verginica) based on the dimensions of the sepal and petal. The Iris data set would be classified by looking for patterns in the sepal and petal sizes of the Iris flower, and then determining how the pattern was predicted to form the class of the Iris flower. Experimental results illustrated that both ANN and SVM can classify iris flowers successfully by obtaining 98.66% and 97.79% of accuracy, respectively.
Keywords: Machine Learning, Deep Learning, AI, Iris Flower.

1. Introduction

The world is surrounded by a variety of flora and animals in the modern world. People not understanding their significance in our daily lives for numerous uses is the problem. There are lots of ayurvedic herbs, healing flowers, and nutritious foods in the area. This initiative benefits many people by taking into account flowers as both decorations and remedies. Even the aroma is used in numerous therapeutic applications in addition to medicine [1]. The first section of this project is called Flower Classification, and it involves categorizing all the flowers according to their sepal, petal, and ovary dimensions. Due to its clarity and simplicity, the Iris dataset is frequently used in machine learning and data science courses. Creating a method in which existing dataset-based data mining technologies are used to display all of the information about a flower's details [2]. Data mining combines a range of trends, clever methods, algorithms, and software to sift user data and extract information from enormous data warehouses. This strategy will help businesses evaluate outcomes, anticipate possible trends, and forecast user behavior. Data mining uses four methods and phases for useful data extraction. A database is a collection of data from several sources, a sizable database that may contain issue definitions. Data discovery is a technique for gleaning important information from enormous amounts of mysterious data [4]. Modeling, which is the third stage, comprises designing and assessing multiple templates. Validated models are implemented in the last stage of data mining approaches. Businesses may employ data mining techniques to transform unusable data into facts. Recognizing every aspect of consumer behavior will help businesses improve their communication strategies and boost sales. Additionally, this information needs to be appropriately categorized in order to maximize its value [5].

The goal of machine learning is to create computer programs that can learn to improve and adapt as they encounter new information. Predictive analytics or statistical learning are other names for the research area that lies at the nexus of statistics, Artificial Intelligence (AI), and computer science. Machine learning falls into two major types. There are two types: supervised learning and unsupervised learning. This particular paper focuses on supervised learning. The objective of supervised learning is to infer a function from labeled training data. There are a number of training examples in the training data. In supervised learning, each example consists of a pair of desired output values and input objects. An inferred function is generated by a supervised learning algorithm from the training data, which may then be used to map fresh samples. Regression and classification issues are within the category of supervised learning challenges. When the output variable is a category, such as "illness" or "no disease," the situation is known as a classification problem [6, 1].

An approach for identifying Iris flower species is described in this research. It operates in two stages: testing and training. A Machine Learning (ML) method is loaded with the dataset of the training stage during training, and labels are given to the data. The predictive model also indicates the species to which the iris flower belongs. Therefore, the predicted Iris species have been identified. This research focuses on machine learning-based IRIS flower classification. The problem statement focuses on identifying IRIS flower species using measurements of their floral attributes. The IRIS data set would be classified by finding patterns in the petal and sepal sizes of the IRIS flower and determining how the class of the IRIS flower was predicted by evaluating the pattern. In this study, we use data to train the machine learning model, and when previously unknown data is found, the algorithm predicts the species based on what it has learned from trained data.

2. Literature Review

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. Additionally, it only employs k-Nearest Neighbors, random forests, and decision trees as classifiers.

The Gaussian Naive Bayes technique is used by Zainab Iqbal [7] 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 [8] 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 [9] 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 [10] 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 [11] 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 [12] 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). 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%.

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. After examining the characteristics of the iris flowers, we classify the iris data set by identifying patterns, and we then project how the patterns will be processed to create the class of iris flowers. 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. Developed Method

A data mining technique called classification divides data instances into a few different classes. The numerous models that built ML identification techniques have all been developed to surpass one another. They all employ statistical techniques, including, but not limited to, DT, SVM, and ANN. These techniques look at the available data in various ways to guess [13]. Figure 1 displays the conceptual framework that this work has built. The classification of iris flower species is done using a

technique that is described. Testing and training are the two stages. Machine learning models are fed data sets during the training stage. Which category the iris flower belongs to is indicated by the predictive model. In this study, we use ML techniques to classify the many kinds of iris flowers. By analyzing the sepal and petal sizes of the flowers, we must find patterns in the iris data set in order to categorize it. The classification of the iris flower is then formed by looking closely at the pattern to make the forecast. The machine learning model is trained by giving it data sets, and if any unexpected data is found or seen, it will forecast the species of flower based on what it has learned from training the data. Our work is to identify iris species based on floral characteristics.

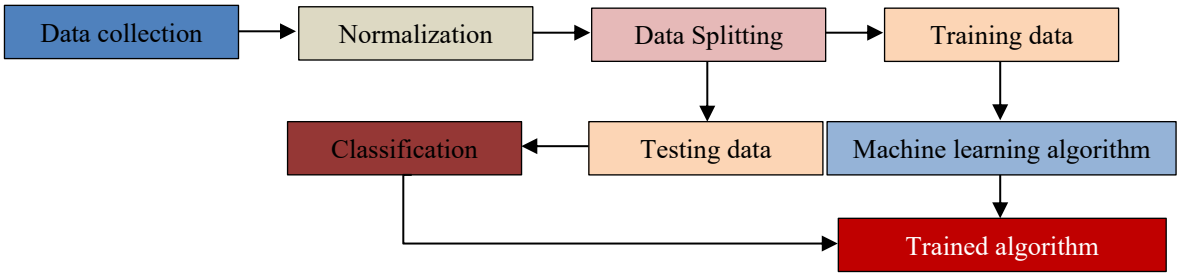

Figure 1: Steps of the General Block Diagram

3.1 Normalization

One of the most popular methods for preparing data is normalization, which enables us to convert the values of the dataset's numerical columns to a common scale. Although it is not required for all available data for machine learning, normalization is used anytime a dataset's properties have a wide range. A machine learning model's performance and dependability are improved as a result. This study gave us the inspiration to utilize the formula below, which allowed us to transform the value into a suitable range.

$$Z_{ab} = \frac{Z_{ab} - c_{min}}{c_{max} - c_{min}} \quad (1)$$

Compute the c_{min} minimum value of the column. Subtract the c_{min} from each of the data, computer the value of (c_{max} - c_{min}), and computer the new value by dividing the two previous results.

Where c_{min} = minimum value of the column, c_{max} =maximum value of the column, and Z_{ab} = Z (data item) present at a^{th} row and b^{th} column.

3.2 Machine Learning Algorithm

ML is more efficient at validating the data and their behavior and exploring information. When data is accessible, divide it into training and test datasets, then train an algorithm to investigate where the data will be in the future. Machine learning employs two different types of methods: descriptive techniques methods (unsupervised methods), which uncover hidden patterns or intrinsic structures in input data, and predictive methods (supervised technique methods), which train a framework on known input and output data in order to predict future outputs. The goal of predictive approaches is to create a model that, in the face of ambiguity, produces predictions based on data. A predictive technique trains a system to produce an accurate classification for the response to new data using a known set of input data and a known response to the data (output). To remain in control over sensitive data and to meet regulatory criteria, data classification is crucial. Understanding the data laws and rules that are used to categorize the data is crucial since machine learning, a subset of artificial intelligence can assist in identifying and categorizing data [14].

A. Artificial Neural Network

A conceptual system or mathematical model based on biological neural networks, or a replication of a biological neural system, is known as an Artificial Neural Network (ANN). It uses a connectionist method of computation to handle information and is made up of a network of artificial neurons. An ANN is often an adaptive system that modifies its structure in response to information coming from the outside or inside the network [15] during the learning phase. Neural networks are non-linear statistical data modeling tools, to put it more simply. They can be applied to identify patterns in data or to model intricate relationships between inputs and outputs. Similar to the extensive network of neurons in the human brain, a neural network is a linked collection of nodes [16].

The simplest ANNs imitate the function of the brain by using a layered structure of interconnected ANs. This type of architecture is also referred to as a Dense or Fully Connected Network since each neuron in an ANN receives the activation of every AN of the preceding layer as input parameters (Figure 2). A Feed-Forward Network would be another term for this architecture, which only connects layers in order. Deep ANNs are ANNs having multiple layers between the input and output layer (also known as hidden layers), and training such networks is referred to as deep learning. Typically, supervised learning is used for training, which involves manually annotating the data before to training and using backpropagation to change the weights for each input [17].

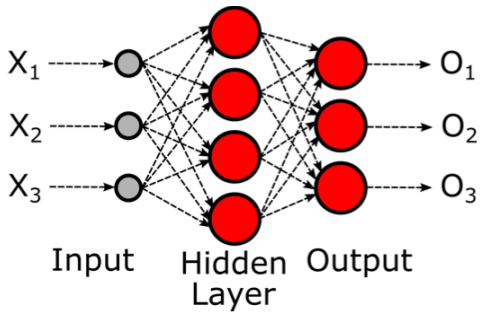

Figure 2: Basic Structure of ANN [16]

B. Support Vector Machine

The main application of SVM in medical data processing is identification. In a multidimensional setting, SVM functions as a separator of distinct data sets. It has the ability to classify data into binary and multiclass categories. Clinical coding, which converts medical data into standardized statistical code, depends on identification. To assess crucial data, identification, for instance, splits the data into diagnostic or process codes. The identification is based on different health assessment-related characteristics [18]. The main objective of this strategy, as shown in Figure 3, is to project nonlinearly identifiable samples to a higher-dimensional space using various kernel functions. Recently, the popularity of SVM has drawn a lot of attention to kernel techniques. The connection between linearity and nonlinearity mainly relies on kernel functions. SVM employs the idea of transforming the input domain into a space with many dimensions to enhance the identification function [19].

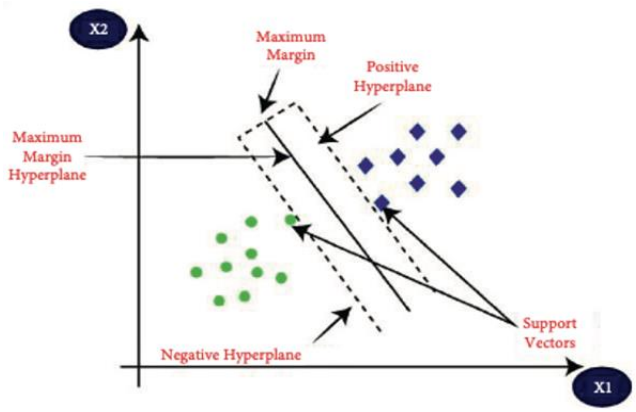

Figure 3: Support Vector Machine [23]

A task that can be divided into two categories is binary categorization. SVM can be effectively employed when the data contains exactly two categories. SVM begins by categorizing data by locating hyperplanes and dividing data points according to a few characteristics. There are numerous hyperplanes; the best has the largest margins and more interdependent data.

3.3 Training Method

Our method is trained using the dataset to make reliable output predictions. By extracting information from this dataset, we concentrate on categorizing the iris flower class. The processed data includes an analysis of each parameter. Data preparation is necessary for the machine learning process in order to transform the data into a format that the machine can interpret. We may say that the algorithm can now quickly analyze the data features. The data is transformed into binary format for our project. The algorithm now performs the necessary calculations. In order for us to understand the output, it is converted from binary to hexadecimal format. Hexadecimal codes are used to identify colors. The aim is to categorize the flowers according to their flower characteristics.

4. Evaluation Results

The iris dataset that we gathered is made up of 150 tuples with four flower parameters, including sepal length, sepal width, petal length, and petal width. Setosa, Versicolor, and Virginia are the 3 distinct flower species that are represented. The dataset has been divided into training and testing, 75% of the dataset is the training dataset, while 25% is the testing dataset.

4.1 Dataset

The UCI Machine Learning Repository served as the source of the dataset for this study. A multivariate data set is the Iris flower data set, often known as Fisher's Iris data set. The data set includes 50 samples from each of the three Iris species (Iris virginica, Iris versicolor, and Iris setosa). Each sample was measured (in centimeters) for the following four characteristics: sepal length, sepal width, sepal length, and sepal length. The Iris species are depicted in Figure 4.

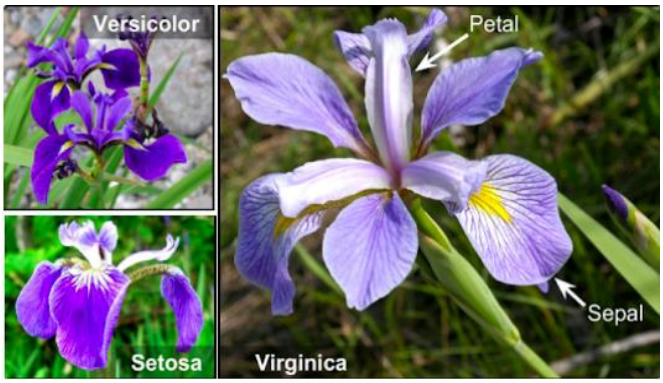

Figure 4: Types of Irises Flowers [11]

The observational data with 150 items are included in the Iris flower data collection. Since the data frame contains 150 items that fall into one of the three target categories and four features (sepal width, sepal length, petal width, and petal length). In this stage, we analyze the dataset mathematically to evaluate the performance of the method. Examples from the IRIS dataset is shown in Table 1 as illustrations.

Table 1: IRIS Dataset-based Sample Instances

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
6.5	3.0	5.2	2.0	virginica
6.2	3.4	5.4	2.3	virginica
5.9	3.0	5.1	1.8	virginica

4.2 Evaluation Results

The input data is the 150x4 matrix from the IRIS dataset, which may be acquired from the UCI repository at www.ics.uci.edu. 75 of these 150 instances were utilized for testing, while 75 were utilized for training. The IRIS dataset's columns and rows each indicate a characteristic or property of the iris flower. Sepal length, sepal width, petal length, and petal breadth are the four characteristics of an iris flower. Each sample in supervised learning has a class label, therefore the Iris dataset contains three classes: Setosa, Versicolor, and Virginia. On the Iris dataset, we can learn the following interpretable rule.

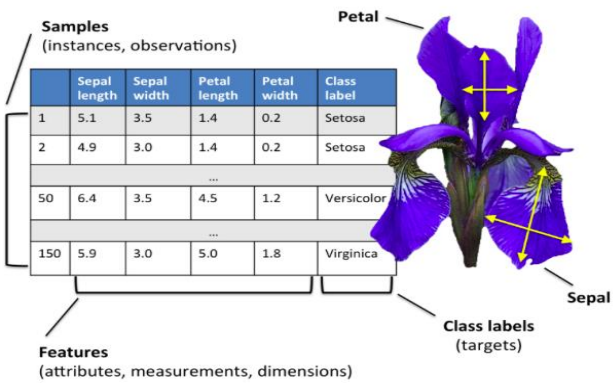

Figure 5: Representing IRIS Dataset

The preceding rule can be parsed in the manner listed below. If an iris flower complies with the criterion, it is assumed to be an Iris Versicolor. The rule is a disjunction of literals in each of the three clauses that make up the conjunction. A literal can make binary decisions if the sepal length is greater than 6.3. According to the aforementioned rule, a clause is satisfied when at least one literal is met, and the rule is met when every clause is met. In this work, we are interested in learning logical formulations that represent categorization rules. Equations (2–4) were used to calculate accuracy, as shown below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Pre = \frac{TP}{(TP+FP)} \times 100\% \quad (4)$$

The terminology used in the above equations that were listed in the confusion matrix is clarified in more detail below: True Positive (TP) is the result of the sample properly predicting the positive class; False Positive (FP) is an outcome in which the sample correctly predicts the positive class, whereas False Negative (FN) is an event in which the sample wrongly predicts the negative class. True Negative (TN) is an outcome in which the sample accurately predicts the negative class.

Table 2: Evaluation Results on IRIS Classification Based data Splitting with 75% Training and 25% Testing

Classifier	Accuracy	Recall	Precision
ANN	98.66	96.32	98
SVM	96.79	93.31	93.88

ANN and SVM, are two classifiers that were assessed using this method using IRIS datasets. The findings show that the classifiers offer various resolutions on various datasets due to functional differences. The evaluation of the accuracy, precision, and recall-based method was described in Table 2. The ANN algorithm outperforms both ANN and SVM in terms of performance. This study's final finding is that ANN outperforms SVM in terms of accuracy.

Table 3: Evaluation Results on IRIS Classification Based data Splitting with 80% Training and 20% Testing

Classifier	Accuracy	Recall	Precision
ANN	96.72	95.09	96.72
SVM	95.03	93.01	92.31

Based on the classifier employed in this study, ANN and SVM classification algorithms were documented in linked works in this article, illuminating the significant tasks that the researchers set with each approach examined. In this section, the findings from this study were contrasted with those from studies based on research from related works. According to related work, a study [20] employed j48 and RF on IRIS datasets to boost their efficiency and achieved 95.83% accuracy for j48 and 95.55% accuracy for RF. The study in [21] presents a method for iris classification using different machine learning techniques including LDA, PCA, RF, and LR. The same dataset has been used in the evaluation; results showed that LDA among them achieved higher accuracy by obtaining 100% whereas other techniques achieved lower accuracy compared to this work by obtaining 86%, 94%, and 96% respectively. Moreover, a study in [22] used only DT for iris classification by obtaining 98%. This method obtained higher accuracy compared to this study when using SVM while when using ANN this work obtained higher accuracy.

5. Conclusion

In this research, we trained our data using a variety of strong algorithms. The best results can only be obtained by data processing, and as we can see from the findings above, those results are quite pleasing. In two of the four models mentioned above, which have accuracy scores of 98.66% and better respectively, it is possible to predict the species of the iris flower. We can draw the conclusion that it will be feasible to identify the species of any flower in the future with the right information on its characteristics.

References

[1] Bhutada, S., K. Tejaswi and S. Vineela. Flower Recognition Using Machine Learning. International Journal Of Researches In Biosciences, Agriculture And Technology, 4(2), 67-73, 2021.

[2] Khalid, L. F., Abdulazeez, A. M., Zeebaree, D. Q., Ahmed, F. Y., & Zebari, D. A. (2021, July). Customer churn prediction in telecommunications industry based on data mining. In 2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA) (pp. 1-6). IEEE.

[3] Haji, S. H., Abdulazeez, A. M., Zeebaree, D. Q., Ahmed, F. Y., & Zebari, D. A. (2021, July). The Impact of Different Data Mining Classification Techniques in Different Datasets. In 2021 IEEE Symposium on Industrial Electronics & Applications (ISIEA) (pp. 1-6). IEEE.

[4] P. Galdi and R. Tagliaferri, "Data mining: accuracy and error measures for classification and prediction," Encycl. Bioinforma. Comput. Biol., pp. 431–6, 2018.

[5] Chicho, B. T., Abdulazeez, A. M., Zeebaree, D. Q., & Zebari, D. A. (2021). Machine learning classifiers-based classification for IRIS recognition. Qubahan Academic Journal, 1(2), 106-118.

[6] Rao, T. S., Hema, M., Priya, K. S., Krishna, K. V., & Ali, M. S. (2021). Iris Flower Classification Using Machine Learning. Network, 9(6).

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

***[8]** 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.*

[9] 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

[10] K R Rathy, Arya Vaishali, "Classification of Dataset using Efficient Neural Fuzzy Approach", vol. 099, August 2019.
[11] D. Decoste, E. Mjolsness. 2001. "State of the art and future prospects by using Machine Learning", vol. 320, 2013.
[12] Y. Lakhdoura and R. Elayachi, "Comparative Analysis of Random Forest and J48 Classifiers for 'IRIS' Variety Prediction," Glob. J. Comput. Sci. Technol., 2020
[13] L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 446–452, 2015.
[14] Hassan, C. A. U., Khan, M. S., & Shah, M. A. (2018, September). Comparison of machine learning algorithms in data classification. In *2018 24th International Conference on Automation and Computing (ICAC)* (pp. 1-6). IEEE.
[15] Taher, K. I., Abdulazeez, A. M., & Zebari, D. A. (2021). Data Mining Classification Algorithms for Analyzing Soil Data. *Asian Journal of Research in Computer Science*, 17-28.
[16] Zafeiris, D.; Rutella, S.; Ball, G.R. An Artificial Neural Network Integrated Pipeline for Biomarker Discovery Using Alzheimer's Disease as a Case Study. *Comput. Struct. Biotechnol. J.* **2018**, *16*, 77–87.
[17] 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.
[18] Abdulqadir, H. R., Abdulazeez, A. M., & Zebari, D. A. (2021). Data mining classification techniques for diabetes prediction. Qubahan Academic Journal, 1(2), 125-133.
[19] 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.
[20] Y. Lakhdoura and R. Elayachi, "Comparative Analysis of Random Forest and J48 Classifiers for 'IRIS' Variety Prediction," *Glob. J. Comput. Sci. Technol.*, 2020.
[21] D. Rana, S. P. Jena, and S. K. Pradhan, "Performance Comparison of PCA and LDA with Linear Regression and Random Forest for IRIS Flower Classification," *PalArchs J. Archaeol. EgyptEgyptology*, vol. 17, no. 9, pp. 2353–2360, 2020.
[22] K. Sarpatwar *et al.*, "Privacy Enhanced Decision Tree Inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 34–35.
[23] Abdulazeez, A. M., Zeebaree, D. Q., Zebari, D. A., & Hameed, T. H. (2021). Leaf Identification Based on Shape, Color, Texture and Vines Using Probabilistic Neural Network. *Computación y Sistemas*, *25*(3), 617-631.