E-ISSN 2581 – 7957
P-ISSN 2277 – 3916
CVR Journal of Science and Technology, Volume 25, December 2023
DOI: 10.32377/cvrjst2505

Enhancing Plant Leaf Identification: A Comparative Study of Machine Learning Models

D. Bhanu Prakash[1] and G. Santhosh Kumar[2]
[1]Assoc. Professor, CVR College of Engineering/ECE Department, Hyderabad, India
Email: pbhanududi@gmail.com
[2]Sr. Asst. Professor, CVR College of Engineering/ECE Department, Hyderabad, India
Email: santhoshemwave@gmail.com

Abstract: **In this paper, we present a comprehensive analysis of the application of various machine learning models to the task of plant leaf identification. A Weiner filter is applied for noise reduction in the data, and morphological operations are utilized for feature extraction. Subsequently, evaluated the performance of eight different classification models: Quadratic Discriminant Analysis (QDA), Extra Trees Classifier, Random Forest Classifier, Linear Discriminant Analysis, SGD Classifier, Bagging Classifier, Perceptron, and AdaBoost Classifier. These models are assessed in terms of their accuracy, balanced accuracy, ROC AUC, F1 Score, and time taken for predictions. The results reveal that QDA emerges as the top-performing model, achieving remarkable accuracy and balanced accuracy of 93%, along with an F1 Score of 93%. Extra Trees Classifier and Linear Discriminant Analysis also exhibit strong performance with high accuracy and balanced accuracy scores. The SGD Classifier, Bagging Classifier, and Perceptron yield competitive results as well. However, the AdaBoost Classifier falls short in terms of accuracy and F1 Score, indicating challenges in plant leaf identification. The Random Forest Classifier, while achieving an accuracy of 87%, shows slightly lower balanced accuracy and F1 Score.**

Index Terms- **Plant leaf identification, Wiener Filter, Extra Trees classifier, Bagging classifier, Balanced accuracy.**

I. Introduction

In addition to humans, plants are also important for other living creatures. They play a significant role in maintaining the world's climate and biodiversity. They can transform the light energy that comes from the sun into food for humans and other living things [1]. Unlike plants, animals cannot produce their own food. They rely on vegetation to supply them with the energy they need to survive. Plants also provide all the oxygen that organisms need. More coal and gas used by humans are extracted from the plants that lived hundreds of years ago. Unfortunately, people are destroying these natural environments, which will lead to the emergence of different plant types and the yearly death of plants [2].

Various effects of ecological disasters can also be seen in the form of land flooding, desertion, and weather anomalies [3]. These can result in lower survival rates for people and their habitat.

The field of machine learning and computer vision has gained more attention in the recognition of plants. There have been numerous studies on how to classify different kinds of plants. Before the invention of digital cameras and computer systems, people had to thoroughly study the classification of medical plants [4].

In the past, the lack of experience when it comes to the classification of plants has led to fatal errors, which have increased the mortality rate of patients [5]. Using artificial intelligence, machine vision, and digital videos, it has been shown that the classification of plants has improved significantly. This has motivated computer scientists and botanists to improve the systems used for plant classification. Computer vision and image processing are still being studied in various areas since they have numerous practical applications [6].

The plant leaves are regarded as essential features when it comes to performing the classification process. These provide computer models with valuable information about the plant. Furthermore, the textures found on the leaves are also known to play a vital role in determining the classification of the plant.

Due to the increasing importance of plants, various measures have been implemented to safeguard their resources. Understanding the characteristics of plants is very important to protect their populations.

Most of the time, the non-professional scientists are focused on performing plant classification [7]. This process involves identifying the various types of plants. There are around 4 lakhs plant types. These are given with names and recognized by experts.

The field of plant classification is regarded as the most demanding part of the biological and the agriculture industry. It involves identifying new plant species and developing a computerized system for their administration. There are also various requirements that are needed to perform the classification process for the benefit of agriculturists.

Plant recognition is a process that involves identifying all the plants and performing a plunging arrangement based on their similarities. This is beneficial for various applications, such as environmental protection and education.

There are various challenges that occur while performing the plant classification process. To overcome these issues, researchers use a sample image of the plant leaf to help them identify the plant category. This method also eliminates the need for the plant to be identified using the whole body. The goal of this paper is to use leaf recognition to automatically identify plants.

II. Literature Survey

A new technique [8] for classification of plant leaves based on their various invariants. The seven new invariants were developed using an area-oriented approach. The other

Received on 17-10-2023, Revised on 14-11-2023, Accepted on 28-11-2023.

CVR Journal of Science and Technology, Volume 25, December 2023
DOI: 10.32377/cvrjst2505

six were performed using a newer one based on the Geometric distribution of the first two Hu moment invariants.

Authors [9] developed small-scale disease clusters on plant leaves to detect the early signs of plant diseases using ANN. They used a contrast enhancement method before implementing the model. Later, they categorized the collected data into the various features of the model. The researchers then used a wrapper-based approach to select the best features of the model. The ANN was able to classify the collected data into two categories: normal and abnormal. It was developed on low-end smartphones for farmers.

The researchers [10] developed a diagnosis system that allows a computer-aided study of the various leaves of medicinal plants. It was able to classify the performance of different types of plants by distinguishing their texture features. The system was then used to extract the necessary features for five classes of leaves.

In 2008, authors [11] proposed a variety of classification models that were made using LBP operator and filters, including global and local features. The local attributes of the images were then obtained using LBP. They [12] presented a method that allows the identification of plant species using the images of occluded leaves and a dataset of complete leaves. They then used the b-spline curve as a 2D point representation of the data.

Yang [13] presented a novel method of identifying plant leaves by incorporating the shape and texture characteristics. The plant leaf classification framework utilized the MTD technique to study the plant leaves' shape information, and the LBP-HF extraction procedure was used to extract the texture feature. Weighted distance was used to calculate the texture and shape attributes of the images of the plant leaves. The chi-square and L1 distances were then utilized to determine these features.

Authors [14] introduced a new model for plant identification. It was divided into three phases: image gathering, pre-processing, and feature extraction. The latter stage involved removing the irregularities, noise, and irregularities from the collected images. A high-resolution camera was utilized for capturing the images, and various morphological constraints were then extracted from the data.

III. Methodology

A plant's identification can be performed automatically using a computer to learn how to identify the leaves. This method can also be used to determine the species of the plant. The efficiency of different plant identification techniques has been compared with that of molecular biology and cell biology. The flexibility and robustness of photographic sampling leaves can be attributed to the use of digital cameras. The general steps in identifying plant leaves are shown in Figure 1.

A. Dataset

The classification model uses the data collected from the D-leaf dataset. The D-leaf database [15] contains data samples taken from different kinds of tropical plants. It has a collection of over 30 leaf images from each of the 43 species. Figure 3.3 shows the database's sample images.

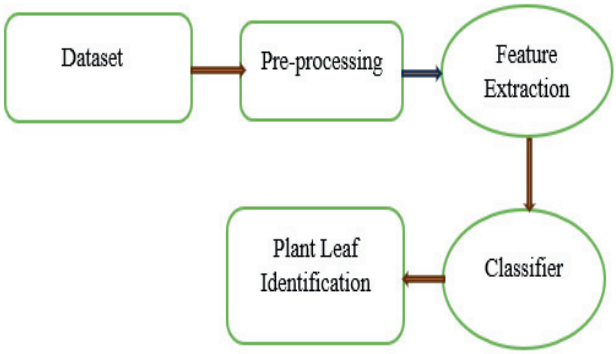

Figure 1. Block diagram of proposed method

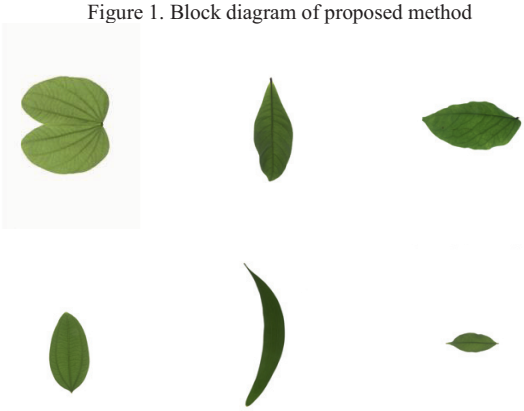

Figure 2. sample images of D-leaf dataset

B. Pre-Processing

The technique known as Wiener filtering [16] considers the image's original characteristics and the noise's statistical model. It can reduce the noise by estimating its original source.

C. Feature extraction

The characteristics of an object or plant that are related to its size, shape, and structure are referred to as morphological features [17]. They are often extracted from images of a plant.

- Counting the number of pixels in the leaf's boundary can determine its total area.
- A leaf's perimeter or outline is the length of its boundary.
- The aspect ratio of a leaf's major and minor axes is shown by comparing their length to the length of its bounding ellipse.
- The circularity of a leaf is computed by comparing its area to that of a circle with a similar perimeter.
- Commonly used to refer to the inverse of circularity, the roundness quotient is calculated as 1/circle.
- The elongation or length of a leaf is determined by comparing its major and minor axis lengths.
- The convex hull area is the area of the smallest polygon that can enclose a leaf.

- The solidity of a leaf is determined by comparing its area to the convex hull's area.
- The Feret's diameter is the maximum distance between a leaf's boundary points.
- The number of linked components in a leaf's image can be identified by the Euler Number, a topological feature. It can be utilized to differentiate complex and simple leaf shapes.
- A leaf skeletonization algorithm can be used to extract the central axis of a leaf. This process is useful in determining the leaf's branching patterns.
- The spread or compactness of a leaf is calculated by comparing its area to its perimeter.
- The leaf's curvature can be measured at various points along its curved surface to get a better understanding of its shape's complexity.
- The symmetry of a leaf can be studied to gain a deeper understanding of its bilateral symmetry, a characteristic present in numerous plant species.
- A leaf's aspect ratio is determined by comparing its width to the height of its smallest boundary.
- The different textures exhibited by a leaf, such as contrast, energy, and entropy, can provide data on its overall structure and appearance.

D. Classifiers

A machine learning classifier is a fundamental component of supervised learning [18]. It helps predict and make decisions based on the input data. These models or algorithms are trained to categorize the data into predefined classes or categories according to the features and patterns they learn.

(i) Extra Tree Classifier: The goal of the Extra Trees Classifier is to provide a more random and robust decision tree structure. It includes features such as random selection and feature splitting. This type of classifier can help improve the robustness and generalization of various data types [19].

(ii) Logistic Regression: This algorithm is a linear classification method that predicts the likelihood of a given outcome based on the logistic function. This is generally applicable to problems where the link between outcomes and features is linear [20].

(iii) Linear Discriminant Analysis: A linear discriminant analysis is a type of classification that seeks to find features that are most likely to separate groups. It is related to PCA and is useful in extracting features and reducing dimensionality [21].

(iv) SGD Classifier: The SGD classifier is a type of linear classification that's trained using the stochastic gradient descent method. It can be used for various applications, such as searching for groups of features [22].

v) Bagging Classifier: A bagging classifier is an ensemble method that combines several base models. It can reduce variance by averaging each model's predictions across different sets of data [23].

(vi) Perceptron: Linear classifier known as the Perceptron is simple and can be used to make decisions on features that are in a linear combination. It is a widely used model for machine learning [24].

(vii) AdaBoost Classifier: It is a framework that combines various weak classifiers to form a powerful one. It aims to improve accuracy by focusing on examples that have been misclassified [25].

(viii) Quadratic Discriminant Analysis (QDA): A QDA is a method for classifying data points that considers the various characteristics of each class's covariance matrix. The algorithm uses Bayes' Theorem to classify the data points. Although a QDA is useful for identifying complex decision boundaries, it is not ideal for handling large sets of features [26].

IV. Results

Table I shows the performance of different machine learning models when it comes to identifying plant leaves. The explanation for each column is provided below.

The column named model highlights the different kinds of machine learning approaches utilized to identify plant leaves.

The percentage of correctly classified plant leaves is referred to as accuracy, and it is a standard metric used in classification.

TABLE I.
Classifier's Performance on D-Datasets

Model	Accuracy (%)	Balanced Accuracy (%)	ROC AUC	F1 Score (%)	Time Taken (sec.)
Quadratic Discriminant Analysis	0.93	0.93	None	0.93	0.04
Extra Trees Classifier	0.88	0.88	None	0.88	0.28
Random Forest Classifier	0.87	0.86	None	0.86	0.70
Linear Discriminant Analysis	0.84	0.84	None	0.83	0.05
SGD Classifier	0.84	0.83	None	0.83	0.15
Bagging Classifier	0.81	0.81	None	0.81	0.33
Perceptron	0.77	0.77	None	0.77	0.06
AdaBoost Classifier	0.11	0.09	None	0.03	0.40
Support Vector Machine	0.82	0.81	None	0.81	0.13
K-Nearest Neighbors	0.83	0.82	None	0.82	0.07
Decision Tree	0.79	0.78	None	0.79	0.05
Naïve bayes	0.75	0.75	None	0.75	0.02

CVR Journal of Science and Technology, Volume 25, December 2023
DOI: 10.32377/cvrjst2505

The Balanced Accuracy metric considers the imbalanced classes in the dataset. It is calculated by averaging the sensitivity and specificity of the classes.

The ROC AUC is a measure of the model's ability to differentiate between classes. It considers the trade-off between the false positive and true positive rates. A value of 1 indicates that the model is perfect at discriminating, while a value of 0.5 indicates random guessing.

The F1 Score, which is a harmonic representation of recall and precision, is useful when trying to balance these two measures in an imbalanced dataset. A higher score indicates that the model has a better chance of achieving a better balance.

The time taken by each model to make predictions is shown in this column. It shows the model's computational efficiency and is useful in large-scale applications.

This classification system has a balanced and high accuracy rate, which shows it can identify plant leaves. Its F1 Score is also impressive at 93%, indicating that it can strike a good balance between recalling and precision. It can do calculations in only 0.04 seconds.

The Extra Trees classifier has an excellent accuracy rate of 88% and a balanced accuracy of 88%. Its F1 Score of 88% indicates that it can strike a balance between recalling and precision, and it can perform predictions in around 0.28 seconds.

Although it has an overall accuracy rate of 87%, the Random Forest classifier has a slightly lower balance of 86. Its F1 score of 86% indicates that it can manage between recalling and precision while still being able to perform accurate predictions.

This linear discriminant analysis can provide an accuracy of 84% and a balanced accuracy of 84%. Its F1 score of 83% indicates that it can maintain a good balance between recalling and precision. It is very fast, taking only 0.05 seconds.

The SGD Classifier is comparable to a linear discriminant analysis in terms of its accuracy and balanced accuracy. It has an F1 score of 83%, and it takes around 0.15 seconds to complete.

The Bagging Classifier can achieve an accuracy of 81% and a balanced accuracy of 81%. Its F1 score of 81% indicates that it can maintain a good balance between recalling and precision. It takes around 0.33 seconds to complete predictions.

The Perceptron can provide an F1 score of 77% and a balanced accuracy of 77%. This implies that it can maintain a balance between recalling and precision while still being efficient. It only took around 0.06 seconds to perform its tasks.

The Adaboost classifier has a balanced accuracy of 9%) and a lower accuracy of 11.1%. But its F1 score of 3% indicates that it has a hard time identifying plant leaves. Its prediction rate of 0.40 seconds is slower than other models.

The Support Vector Machine (SVM) in the provided table exhibits a respectable performance with an accuracy of 82%, a balanced accuracy of 81%, and an F1 score of 81%, making it a reliable choice for classification tasks.

The K-Nearest Neighbors (KNN) model demonstrates competitive results, achieving an accuracy of 83%, a balanced accuracy of 82%, and an F1 score of 82%. Its relatively short training time of 0.07 seconds suggests computational efficiency.

The Decision Tree classifier performs adequately, with an accuracy of 79%, a balanced accuracy of 78%, and an F1 score of 79%. Decision Trees are known for their interpretability, making them valuable in certain applications.

Lastly, the Naïve Bayes classifier shows a moderate performance with an accuracy of 75%, a balanced accuracy of 75%, and an F1 score of 75%, coupled with a very short training time of 0.02 seconds, indicating simplicity and efficiency, though the model's assumptions may affect its suitability for diverse datasets.

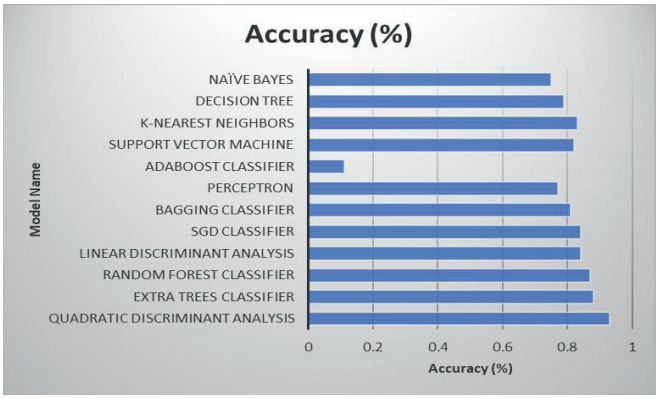

Figure 3. Model vs Accuracy

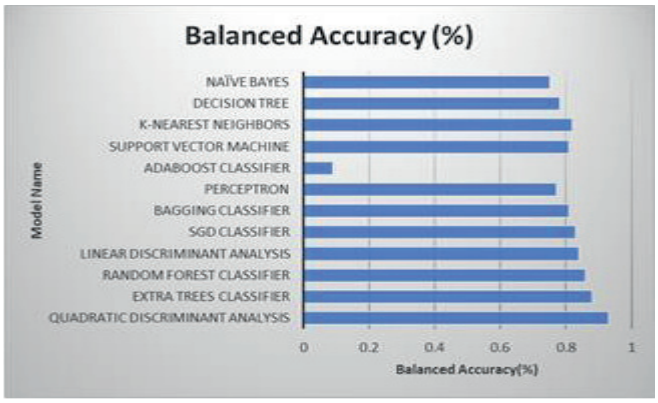

Figure 4. Model vs Balanced Accuracy

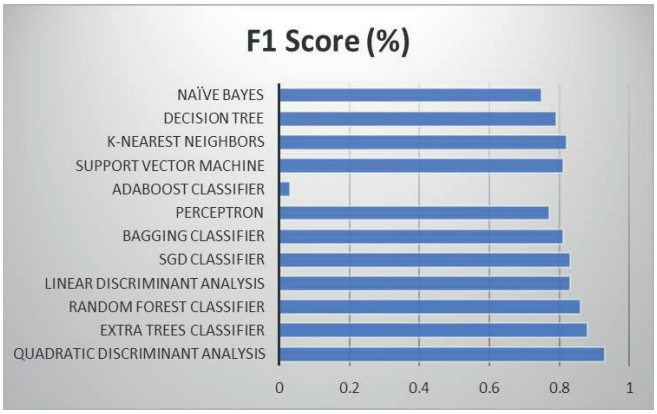

Figure 5. Model vs F1 Score

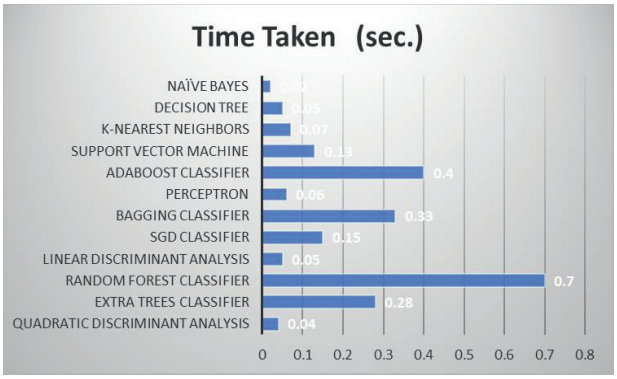

Figure 6. Model vs Time

Figures 3 to 6 indicate the performance measures of various classifiers. The best-performing model when it comes to identifying plant leaves is the Quadratic Discriminator Analysis. It has balanced accuracy and high accuracy. Other models such as the Linear Discriminant and the Extra Trees Classifier provide strong results. But it's crucial to consider the computational efficiency of the model when choosing one for practical applications.

V. Conclusions

This paper provides a comprehensive examination of machine learning models for the challenging task of plant leaf identification. Through a structured approach involving pre-processing, feature extraction, and classification, we shed light on key factors influencing the success of this domain-specific problem. Our study underscores the critical role of the chosen classification model. Notably, Quadratic Discriminant Analysis (QDA) emerged as a standout performer, achieving remarkable accuracy, balanced accuracy, and F1 Score of 93%. Extra Trees Classifier and Linear Discriminant Analysis also demonstrated robust capabilities, making them attractive choices for plant leaf identification tasks.

However, we observed certain challenges associated with specific models. The AdaBoost Classifier struggled to correctly identify plant leaves, resulting in lower accuracy and F1 Score, while the Random Forest Classifier, although delivering an accuracy of 87%, exhibited comparatively lower balanced accuracy and F1 Score.

In summary, this research contributes to the understanding of effective methodologies for plant leaf identification and underscores the adaptability of these techniques to diverse image classification applications. This work will serve as a valuable reference for future endeavors in the realm of computer vision and machine learning.

References

[1] M. Kumar, S. Gupta, X. -Z. Gao and A. Singh, "Plant Species Recognition Using Morphological Features and Adaptive Boosting Methodology," IEEE Access, vol. 7, pp. 163912-163918, 2019.

[2] J. Huixian, "The Analysis of Plants Image Recognition Based on Deep Learning and Artificial Neural Network," IEEE Access, vol. 8, pp. 68828-68841, 2020.

[3] Bhanuprakash Dudi & Dr. V. Rajesh (2023) A computer aided plant leaf classification based on optimal feature selection and enhanced recurrent neural network, Journal of Experimental & Theoretical Artificial Intelligence, 35:7, 1001-1035.

[4] J. Yang et al., "Excitation Wavelength Analysis of Laser-Induced Fluorescence LiDAR for Identifying Plant Species," IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 7, pp. 977-981, July 2016.

[5] L. Li, S. Zhang and B. Wang, "Plant Disease Detection and Classification by Deep Learning—A Review," IEEE Access, vol. 9, pp. 56683-56698, 2021.

[6] B. Liu, C. Tan, S. Li, J. He and H. Wang, "A Data Augmentation Method Based on Generative Adversarial Networks for Grape Leaf Disease Identification," IEEE Access, vol. 8, pp. 102188-102198, 2020.

[7] Y. Wu, X. Feng and G. Chen, "Plant Leaf Diseases Fine-Grained Categorization Using Convolutional Neural Networks," IEEE Access, vol. 10, pp. 41087-41096, 2022.

[8] Mohamed Ben Haj Rhouma, Joviša Žunić, Mohammed Chachan Younis, "Moment invariants for multi-component shapes with applications to leaf classification", Computers and Electronics in Agriculture, vol. 142, pp. 326-337, November 2017.

[9] N. Pham, L. V. Tran and S. V. T. Dao, "Early Disease Classification of Mango Leaves Using Feed-Forward Neural Network and Hybrid Metaheuristic Feature Selection," IEEE Access, vol. 8, pp. 189960-189973, 2020.

[10] Diksha Puri, Abhinav Kumar, Jitendra Virmani & Kriti "Classification of leaves of medicinal plants using laws' texture features," International Journal of Information Technology, 2019.

[11] S. Sladojevic, Marko Arsenovic, Andras Anderla, D. Culibrk, Darko Stefanovic "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," Computational Intelligence and Neuroscience, Vol. 2016, 2016.

[12] A. Chaudhury and J. L. Barron, "Plant Species Identification from Occluded Leaf Images," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 3, pp. 1042-1055, 1 May-June 2020.

[13] Chengzhuan Yang, "Plant leaf recognition by integrating shape and texture features," Pattern Recognition, Vol. 112, April 2021.

[14] Skanda H N, Smitha S Karanth, Suvijith S, Swathi K S, "Plant Identification Methodologies using Machine Learning Algorithms," International Journal of Engineering Research & Technology (IJERT), Vol. 8 Issue 03, March-2019.

[15] Dudi, B., Rajesh, V. Optimized threshold-based convolutional neural network for plant leaf classification: a challenge towards untrained data. J Comb Optim 43, 312–349 (2022).

[16] Dogariu L-M, Benesty J, Paleologu C, Ciochină S. An Insightful Overview of the Wiener Filter for System Identification. Applied Sciences. 2021; 11(17):7774.

[17] Hirata NST, Papakostas GA. On Machine-Learning Morphological Image Operators. Mathematics. 2021; 9(16):1854.

[18] Cherian, I.., Agnihotri, A.., Katkoori, A. K.., & Prasad, V. . (2023). Machine Learning for Early Detection of Alzheimer's Disease from Brain MRI. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 36–43.

[19] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. Mach Learn 63, 3–42 (2006).

[20] Alzen, J.L., Langdon, L.S. & Otero, V.K. A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. IJ STEM Ed 5, 56 (2018).

[21] Adebiyi MO, Arowolo MO, Mshelia MD, Olugbara OO. A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. Applied Sciences. 2022; 12(22):11455.
[22] Kabir, Fasihul, et al. "Bangla text document categorization using stochastic gradient descent (sgd) classifier." 2015 International Conference on Cognitive Computing and Information Processing (CCIP). IEEE, 2015.
[23] Sreng S, Maneerat N, Hamamoto K, Panjaphongse R. Automated Diabetic Retinopathy Screening System Using Hybrid Simulated Annealing and Ensemble Bagging Classifier. Applied Sciences. 2018; 8(7):1198.
[24] Morariu, Daniel, Radu Crețulescu, and Macarie Breazu. "The weka multilayer perceptron classifier." International Journal of Advanced Statistics and IT&C for Economics and Life Sciences 7.1 (2017).
[25] Zhang, Yanqiu, et al. "Research and application of AdaBoost algorithm based on SVM." 2019 IEEE 8th joint international information technology and artificial intelligence conference (ITAIC). IEEE, 2019.
[26] Ghojogh, Benyamin, and Mark Crowley. "Linear and quadratic discriminant analysis: Tutorial." arXiv preprint arXiv:1906.02590 (2019).