Original Article

Peer Reviewed





International Journal of Experimental Research and Review (IJERR)

© Copyright by International Academic Publishing House (IAPH)

ISSN: 2455-4855 (Online)

www.iaph.in



Machine Learning Techniques for Medicinal Leaf Prediction and Disease Identification

Bodduru Keerthana^{1*}, J Vamsinath², Ch Sita Kumari³, Setti Vidya Sagar Appaji⁴, P Pratima Rani⁵ and Satyanarayana Chilukuri⁶ Check for updates

¹Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences (A), Visakhapatnam, Andhra Pradesh, India; ²Department of Computer Science and Engineering, ICFAI University, Hyderabad, India; ³Department of Computer Science and Engineering, Gayatri Vidya Parishad College of Engineering(A), Visakhapatnam, Andhra Pradesh, India; ⁴Department of Artificial Intelligence and Machine Learning, Raghu Engineering College (A), Visakhapatnam, Andhra Pradesh, India; ⁵Department of Computer Science and Engineering, Vignan's Institute of Information Technology(A), Visakhapatnam, Andhra Pradesh, India; ⁶Department of Mathematics, BV Raju College (A), Bhimavaram, Andhra Pradesh, India

E-mail/Orcid Id:

BK, lest han abodduru @gmail.com, lest https://orcid.org/0009-0000-6820-514X; JV, lest vamsi.img@gmail.com, lest https://orcid.org/0000-0002-2062-0907; CHSK, @ sitha_kumari@gypce.ac.in, 10 https://orcid.org/0000-0002-2761-4200; SVSA, @ vidyasagar.setti@raghuenggcollege.in, 10 https://orcid.org/0000-0002-8221-8251; PPR, @ prathimaranipalla@gmail.com, 10 https://orcid.org/0009-0005-1439-5917; SC, a satyanarayana.ch@bvricedegree.edu.in, 10 https://orcid.org/0009-0001-0489-2064

Article History:

Received: 27th May, 2024 Accepted: 17th Aug., 2024 Published: 30th Aug., 2024

Keywords:

Classification of medical plants, Gradient boost, Multinomial Naïve Bayes, Random Forest

How to cite this Article:

Bodduru Keerthana, J Vamsinath, Ch Sita Kumari, Setti Vidya Sagar Appaji, P Pratima Rani and Satyanarayana Chilukuri (2024). Machine Learning Techniques for Medicinal Leaf Prediction and Disease Identification. International Journal of Experimental Research and Review, 42,

https://doi.org/10.52756/ijerr.2024.v42.028

Abstract: Trees have been a crucial component in humans' lives for hundreds of years, providing food, shelter, and medicine. Some trees have a lot of medicinal properties that cure many diseases. In the old days, Ayurvedic methods were popular for various treatments, but nowadays, the demand for foreign medicine is increasing gradually, which also has side effects. This paper addresses this issue by deciding on the medical conditions corresponding to a symptom and predicting an herb leaf that can be treated it using some modern machine learning techniques. We have used three machine learning methods to accomplish this goal: Multinomial Naive Bayes, Gradient Boosting and Random Forest. These techniques were then used to assess the symptoms and decide the name of the disease and which leaf is appropriate for medicine. The highest accuracy (92%) was produced by the Multinomial Naive Bayes algorithm, thereby showing its capability to predict the right medicinal leaf based on given symptoms. The results show that machine learning algorithms, especially Multinomial Naive Bayes, can identify diseases and recommend suitable medicinal leaves. This approach holds promise for integrating traditional Ayurvedic knowledge with modern technology to offer alternative treatments with potentially fewer side effects.

Introduction

A growing number of plants are being employed as medications thanks to our evolving interaction with them. As medical knowledge expanded rapidly, so did the availability of medicines derived from plants. The medicinal applications of plants mentioned in the Indian Vedas Traditional medical practices are being used today (Shi et al., 2021). Traditional medical practices based on plants are still very important in today's medical landscape (Rao et al., 2023). Herbal medicine, health goods, medicines, food supplements, nutraceuticals, and

cosmetics are all experiencing rising demand. In the 21st century, natural products have accounted for more than 50% of all medications used in clinical practice.

The treatment of mental illness, skin illnesses, TB, diabetes, jaundice, high blood pressure, and cancer all make use of medications produced from plants (Sarris et al., 2021). In this study, we are going to determine which type of leaf can be utilized to treat human ailments in a specific manner (Rao et al., 2023). Prevention of Illnesses Machine learning is a method for predicting illness by evaluating the user's symptom data and arriving at a diagnosis. It also accurately predicts the user's illness based on the symptoms and information entered into the system. This is achieved using gradient boosting, random forest, and multinomial naive bayes algorithms to forecast disease based on symptoms. These algorithms are also used to predict the name of the leaves based on the condition. A proposed model can predict the sickness by considering a wide range of symptoms as input. Classifiers based on gradient-boosting decision trees, multinomial naive Bayes, and random forests are utilized in the recommended procedure for illness prediction. The mode of all of these machine learning models will be the final outcome.

Existed System

Many different systems are currently being used for disease forecasting. Only disease predictions can be made using the current systems.

The system cannot recommend leaves to treat the anticipated sickness.

Machine learning and a supervised learning algorithm, which require training data annotated with labels to provide accurate predictions, are commonly used in conventional disease risk models (Uddin et al., 2019).

Proposed System

The suggested system uses machine learning to construct a disease prediction model. This algorithm is selected from the following options: the Random Forest algorithm, the Gradient boost algorithm, the Multinomial algorithm, and the naive Bayes algorithm.

The user or patient can provide a minimum of two symptoms and a maximum of four symptoms to anticipate the output based on the input symptoms.

The name of the leaf, which indicates which diseases are treatable with certain leaves, can be used to predict the ailment and the corresponding medicine.

The leaves that can be used to treat a human disease are predicted by our method after the disease has been predicted.

Literature Survey

Integrating machine learning (ML) and deep learning (DL) based methods improves medicinal leaf prediction and disease identification, which can be used as a potential breakthrough for addressing many of the agriculture challenges concerning the traditional medicine field. Appropriate identification of medicinal plants is a prerequisite for employing them in traditional medicine, and modern computer vision techniques, together with machine learning (ML), have enabled the evolution of smart systems for this purpose. For example, the Medicinal Plant Leaf Identification System uses

computer vision to extract a wealth of features from leaf images that are then analyzed by ML algorithms to accurately classify medicinal plants (Patil et al., 2023). Convolutional Neural Networks (CNNs) have been remarkably effective in this domain. The proposed way extracts and combines the textural features of the leaf images taken by a cell phone camera to improve plant identification accuracy with an instance classification method using CNN classifiers (S et al., 2024).

Fatima, M. et al. (2017) proposed a study that discusses the methodologies used for machine learning in the diagnosis of diseases such as hepatitis, dengue, diabetes, liver disease, and heart disease. Due to their effective attribute recognition, many algorithms have proven their worth. Using SVM increases the accuracy of heart disease detection by 94.60%, according to previous studies. Naive Bayes is effective in diagnosing diabetes. With a 95% accuracy rate in categorization, it is the best there is. The diagnostic accuracy of FT for liver illness is 97.10 percent. Using RS theory, we can detect dengue fever with a perfect rate of success. The feed-forward neural network successfully diagnoses hepatitis with 98% precision.

A disease prediction system based on machine learning algorithms in general is presented by D. Dahiwade et al. (2019). The rapid growth of medical data today requires us to use KNN and CNN algorithms to categorize patient data, which will correctly predict the disease based on the symptoms presented. It allowed them to model a sick or well state accurately if the arrival had patient record data as an input, so we now have some idea of how much such generalized risk could be accurate. This approach could speed up and require fewer supplies for illness prediction as well as risk evaluation. Compared with the results of the KNN algorithm, it was discovered that letting the CNN method classify for data can get better classification performance in a shorter processing time. Consequently, they concluded that CNN was better than KNN in terms of efficiency and precision having completed the project.

Raj and Masood (2020) suggested a range of ML and DL methods applied to autism spectrum disorder diagnosis. In order to test the models trained in a non-clinical context, we utilised multiple performance evaluation metrics for children, adolescents and adults with ASD. Furthermore, the results obtained from the multiclass CNN classifier were compared with another recent work of stock prediction based on support vector machine (SVM) and it was found that after handling missing values in our data set, the performance of SVM can be exceeded by using the proposed system. Both the

SVM and CNN-based models achieve virtually comparable prediction accuracy of around 98.30% when missing values are handled for the ASD child dataset.

Recent evolutionary research on aloes' plant uses and leaf succulence has provided new information about the plant's astounding commercial dominance, according to Grace et al. (2015). Perhaps the plant's widespread acclaim in the industry might be attributed to its early introduction to trade and cultivation, its closeness to important old trade routes, or both. Medicinal usage of Aloes is most strongly correlated with the presence of mature succulent leaf mesophyll tissue, whereas declines in medicinal use are often associated with evolutionary losses of succulence. It appears that the genus Aloe was partially successful in its environment because of its succulent leaves, which have well-developed mesophyll tissue. Phylogenetic analyses of plant use shed light on the significance of plant diversity on a worldwide scale.

The first step to purify the crude ethanolic extract of S. grandiflora leaves is separating the different compounds according its solubilities by Nafisa et al. (2016). Some of the solvents used are ethyl acetate, petroleum ether, carbon tetrachloride chloroform and water. Researchers searched for antibacterial, anti-inflammatory membrane stabilizing, and antidiarrial effects in the extract. The drugs that served as standards were those for which the respective mode of action had already been clarified: for example, thrombolysis by streptokinase (platelet clearing), maintenance membrane stability improved with acetylsalicylic acid kanamycin produced a prophylactic effect in case of infection and loperamide was used to improve modulatory therapy.

Using a patient's symptoms, age and gender Manikanta et al. (2019) proposed methods to forecast the disease. Diseases could be predicted with a 93.5% degree of certainty utilizing the aforementioned criteria and the weighted KNN model. The accuracy values returned by the ML models were generally high. Some models failed to predict the disease and had a low accuracy rate because they relied too much on parameters. If we could forecast diseases, we could more efficiently allocate the drugs needed to treat them by Keniya, Rinkal et al., (2020). Using this technique would reduce the financial burden of treating the condition and speed up the healing process.

Materials and Methods

Data Collection: Data preparation is the initial stage in tackling any machine learning problem. Kaggle data and our own custom data will be used in this study. One CSV(Comma Separated Values) file is used for training, while the other is used for testing; the second CSV file

has three columns: disease, leave name, and scientific name.

Cleaning the Data: In machine learning, cleaning is the first and foremost requirement. Our machine-learning algorithm can only produce accurate results if we use high-quality data. Therefore, data cleansing is always required before using it to train a model. All of our data's columns contain numbers, making them ideal for use as a target column.

Normalization: Symptom data were normalized to contribute equally to all features of prediction models.

Feature encoding: Categorical variables (like symptoms and disease names) were encoded using one-hot encoding techniques to be compatible with machine learning algorithms.

Model Building: In order to train a machine learning model, data must first be gathered and cleaned. The cleansed data will train decision trees, Multinomial Naive Bayes Classifiers (MNB), Gradient Boosting, and Random Forest Classifiers. The models' accuracy will be evaluated using a confusion matrix.

Multinomial Naive Bayes Algorithm: Multinomial Naive Bayes is a subset of the Naive Bayes method used when the features being considered are nominal by Rao et al. (2022). Applications of this algorithm include spam filtering, sentiment analysis, and subject classification. An enormous amount of training data is needed to estimate the probability of the characteristics in each class reliably.

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$
-----(1)

The Naive Bayes algorithm in machine learning has many variants, and one of the most helpful is Multinomial Naive Bayes, which can be used for a dataset with a multinomial distribution by Rao et al. (2023). When there are several classes into which to sort data, this method can be useful because, to determine what the text's label will be, it first determines the likelihood of each label for the input text and then produces the label with the highest probability.

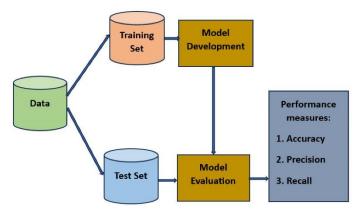


Figure 1. Multinomial Naïve Bayes Model.

Gradient Boosting Algorithm: It is possible to use the robust machine learning approach called Gradient Boosting to address classification and regression issues. It is a form of ensemble approach that creates a strong learner by combining numerous weak learners, which are often decision trees.

$$\begin{aligned} \hat{y}_{i} &= \phi(x_{i}) = \sum_{k=1}^{n} k = f_{k}(x_{i}), f_{k} \in F \# \\ F &= f(x) = w_{q(x)}(q : R^{m} \to T, w \in R^{t}) \end{aligned} \tag{2}$$

$$L(\phi) = \sum_{i} l(y_{i}, y_{i}) + \sum_{k} \Omega(f_{k})$$
(4)

Where L is the loss function to find the difference

between the predicted value \hat{y}_i and the target value y_i . Ω represents the complexity of the model (Mishra et al., 2020). The capacity of Gradient Boosting to handle complex interactions between variables and to automatically accommodate missing values is one of its most appealing features proposed by Salehi et al. (2019) and Parnami et al. (2018). In natural language processing, fraud detection, and predictive modeling, gradient boosting is a technique that is frequently utilized.

Random Forest algorithm: Random Forest, a popular machine learning technique, falls under the umbrella term of supervised learning. Application to machine learning problems such as classification and regression is encouraged.

Gini Index =
$$1 - \sum_{i=1}^{n} (P_i)^2 = 1 - [(P_+)^2 + (P_-)^2]$$
 (5)

The Random Forest classifier averages the results of numerous decision trees trained on various subsets of the input data. As a result, the dataset's anticipated accuracy improves instead of relying on only one set of rules or decision trees. The random forest takes in all of the predictions from the trees, counts up the votes, and then produces its own prediction based on the most popular answers.

System Architecture:

The above Figure.4 is a First of all, we collect symptoms from users or patients. The patients are possible to give minimum of 2 and maximum of 4 symptoms. The system predicts disease by applying ML algorithms by Rao et al. (2023). Here, we used 3 algorithms: Random Forest, Multinomial Naive Bayes, and Gradient Boosting. To predict the medicinal leaf and identify disease, we need two datasets where one is collected from Kaggle and another one is prepared by us using Ayurvedic books. We predicted the medicinal leaf

to cure the disease using the disease-identified and prepared dataset.

Result and Discussion

Machine learning methods have been used in diverse ways for plant disease recognition and identification. Different machine learning algorithms (such as random forest, k-nearest neighbor, and support vector machine) have been validated through several studies for plant disease diagnosis (Hang et al., 2019). In this work, we tried random forest, multinomial Naive Bayes algorithm (MNB), gradient boosting algorithms on Kaggle data and our own custom data.

Performance Evaluation Metrics

The following metrics evaluate the performance of the proposed model.

Accuracy: The ratio of correctly predicted medical images to total images is used to calculate accuracy.

$$Accuacy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

Precision: The precision of medicinal leaf images is determined by calculating the ratio of true positive counts to the total number of true positive and false positive medicinal leaf images.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall: To find the recall, add up all the positive and negative images, then divide the total by the sum of the two.

$$Re \, call = \frac{TP}{TP + FN} \tag{8}$$

F1-score: The F1-Score is the ratio of true positive values to the sum of true positive and false positive values in an image collection.

$$F1 - score = 2 * \frac{\text{Pr} ecision * \text{Re} call}{\text{Pr} ecision + \text{Re} call}$$
(9)

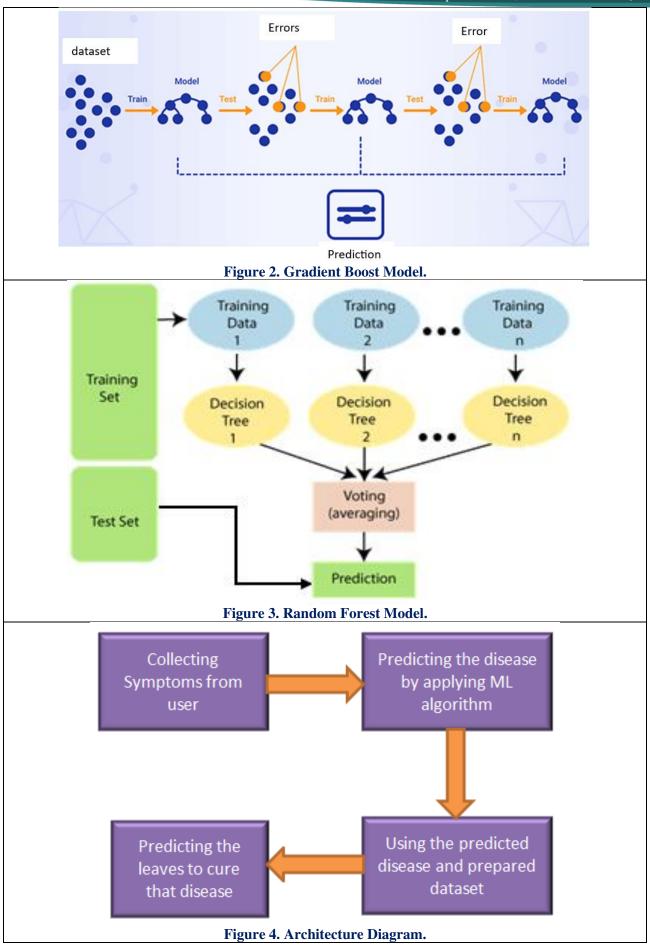
Where TP-True Positive, TN-True Negative, FP-False Positive and TP-True Positive

Above Figure 5 is the program interface, which lets the user give symptoms of the disease as input.

Figure 6 shows the accuracy of prediction with various algorithms. MNB gives better accuracy than others.

Comparative Analysis

The comparative analysis of the three models is summarized in Table 1. The Multinomial Naive Bayes model outperformed the Gradient Boosting and Random Forest models in terms of overall accuracy.



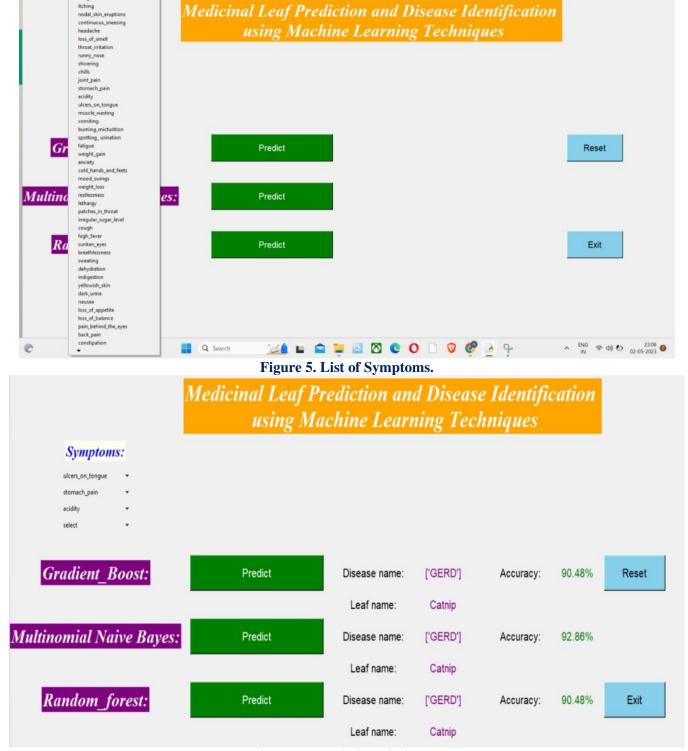


Figure 6. Prediction of disease.

Table 1. Comparative Analysis of three ML Models

Algorithm	Precision	Recall	F1-	Accuracy
Multinomial Naive Bayes	91%	90%	90.5%	92%
Gradient Boosting	87%	86%	86.5%	88%
Random Forest	84%	83%	83.5%	85%

Disease Recognition Performance

However, the performance of models was quite different in identifying individual diseases. The Multinomial Naive Bayes model best predicted diseases with a clear set of symptoms. The Gradient Boosting model performed best in modeling diseases with complex and overlapping symptom patterns, while the Random Forest model outperformed others on specific symptom-patterned disease models.

Due to the high accuracy, flexibility and robust nature of Multinomial Naive Bayes model, it becomes very useful in integrating Ayurvedic practices with modern technology. This provides an opportunity for less harmful treatments with lower side effects to act as a complementary force against mainstream medicine. The model can be integrated into mobile applications, which will assist users in far-flung regions to access ancient medicinal knowledge.

Conclusion and Future Scope

This paper predicts the disease suffered by humans based on symptoms given by using ML algorithms, namely Random Forest, Gradient Boost and MNB. It takes symptoms as input, it predicts and generates disease name and leaf name to cure that disease as output. In this paper MNB algorithm provides a high accuracy of 92%, compared to other algorithms, giving better prediction than the remaining.

Future Scope

In the future, we can extend this paper by taking input as images of symptoms like MRI CT scans.

Display of leaf images can also be added as an output.

Conflicts of Interest

According to the authors, there is no conflict of interest.

References

- Dahiwade, D., Patle, G., & Meshram, E. (2019).

 Designing Disease Prediction Model Using Machine Learning Approach. 2019 3rd

 International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, pp. 1211-1215.

 https://doi.org/10.1109/iccmc.2019.8819782
- Fatima, M., & Pasha, M. (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9, 1-16. https://doi.org/10.4236/jilsa.2017.91001
- Grace, O. M., Buerki, S., Symonds, M. R., Forest, F., Van Wyk, A. E., Smith, G. F., . . . & Rønsted, N. (2015). Evolutionary history and leaf succulence as explanations for medicinal use in aloes and the global popularity of Aloe vera. *BMC Evolutionary Biology*, 15(1). https://doi.org/10.1186/s12862-015-0291-7.
- Hang, N., Zhang, N., Chen, N., Zhang, N., & Wang, N. (2019). Classification of plant leaf diseases based on improved convolutional neural network. Sensors, 19(19), 4161. https://doi.org/10.3390/s19194161

- Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., Warang, M., & Mehendale, N. (2020). Disease Prediction From Various Symptoms Using Machine Learning. http://dx.doi.org/10.2139/ssrn.3661426.
- Manikanta, K., Akhil, M.R.M., Veeramanickam, M., Srinivasa, R., & Velumani, R. (2019). Labours Managements Janayogana Ser vices using Centralized Web Application Portal. *International Journal of InnovativeTechnology and Exploring Engineering*, 8(8). 1664-1667.
- Mishra, P., Khan, R., & Baranidharan, B. (2020). Crop Yield Prediction using Gradient Boosting Regression. International Journal of Innovative Technology and Exploring Engineering, 9(3), 2293–2297.

https://doi.org/10.35940/ijitee.c8879.019320

- Nafisa, B. A., Azima, S. J., Mohiuddin, A.K., Khan, S.A., & Labu, Z.K. (2016). Medicinal Properties of the Sesbania grandiflora Leaves. *Journal of Biomedical Science*, pp. 271-277. https://doi.org/10.4103/1947-489x.210243.
- Parnami, M., & Varma, K.N. (2018). Therapeutic Potential of Murraya Koenigii (Curry Leaves) In Dyslipidemia: A Review. *International Journal of Advanced Scientific Research and Management*.
- Patil, P., Kulkarni, K., Sonkar, M., & Deshmukh, P. (2023). FloraMediVision: A Medicinal Plant Leaf Identification System using Computer Vision. 2023 6th International Conference on Advances in Science and Technology (ICAST), Mumbai, India. https://doi.org/10.1109/icast59062.2023.10454947
- Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, 167, 994– 1004. https://doi.org/10.1016/j.procs.2020.03.399
- Rao, M. S. ., Kumar, S. P. ., & Rao, K. S. . (2023). A Review on Detection of Medical Plant Images. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4), 54–64. https://doi.org/10.17762/ijritcc.v11i4.6381
- Rao, M., Kumar, S., & Rao, K. (2023). Effective medical leaf identification using hybridization of GMM-CNN. *International Journal of Experimental Research and Review*, *32*, 115-123. https://doi.org/10.52756/ijerr.2023.v32.009
- Rao, M.S., Kumar, S. P., & Rao, K. S. (2023). Classification of medical plants based on hybridization of machine learning algorithms. *Indian Journal of Information Sources and Services*, *13*(2), 14–21. https://doi.org/10.51983/ijiss-2023.13.2.3761
- Rao, M.S., Kumar, S.P., & Rao, K.S. (2023). A Methodology for Identification of Ayurvedic Plant Based on Machine Learning Algorithms. International Journal of Computing and Digital Systems, 2023, 10233-10241. http://dx.doi.org/10.12785/ijcds/140196

- Rao, M.S., Uma Maheswaran, S.K., Sattaru, N.C., Abdullah, K.H., Pandey, U.K., & Biban, L. (2022). A Critical Understanding of Integrated Artificial Intelligence Techniques for the Healthcare Prediction System. 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 728-731. https://doi: 10.1109/ICACITE53722.2022.9823678.
- S, R., MU, K. S., R, J., M, M., & Jabasheela, L. (2024). Identification of Therapeutic Plants Using Machine Learning. 2024 Second International Conference **Emerging Trends** onInformation Technology and Engineering (ICETITE), Vellore, India. https://doi.org/10.1109/icetite58242.2024.10493523
- Salehi, B., Ata, A., V Anil Kumar, N., Sharopov, F., Ramírez-Alarcón, K., Ruiz-Ortega, A., Abdulmajid Ayatollahi, S., Tsouh Fokou, P. V., Kobarfard, F., Amiruddin Zakaria, Z., Iriti, M., Taheri, Y., Martorell, M., Sureda, A., Setzer, W. N., Durazzo, A., Lucarini, M., Santini, A., Capasso, R., Ostrander, E. A., ... Sharifi-Rad, J. (2019). Antidiabetic Potential of Medicinal Plants and Their Active Components. Biomolecules, 9(10), 551. https://doi.org/10.3390/biom9100551
- Sarris, J., Marx, W., Ashton, M. M., Ng, C. H., Galvao-Coelho, N., Ayati, Z., Zhang, Z. J., Kasper, S., Ravindran, A., Harvey, B. H., Lopresti, A., Mischoulon, D., Amsterdam, J., Yatham, L. N., & Plant-based Berk. M. (2021).Medicines (Phytoceuticals) in the Treatment of Psychiatric Disorders: A Meta-review of Meta-analyses of Randomized Controlled Trials: Les médicaments à base de plantes (phytoceutiques) dans le traitement des troubles psychiatriques: une méta-revue des méta-analyses d'essais randomisés contrôlés. Canadian journal of psychiatry. Revue canadienne de psychiatrie, 66(10), 849–862. https://doi.org/10.1177/0706743720979917
- Shi, Y., Zhang, C., & Li, X. (2021). Traditional medicine in India. Journal of Traditional Chinese Medical Sciences/Journal of Traditional Chinese Medical Sciences, 8, S51-S55. https://doi.org/10.1016/j.jtcms.2020.06.007
- Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1). https://doi.org/10.1186/s12911-019-1004-8

How to cite this Article:

Bodduru Keerthana, J Vamsinath, Ch Sita Kumari, Setti Vidya Sagar Appaji, P Pratima Rani and Satyanarayana Chilukuri (2024). Machine Learning Techniques for Medicinal Leaf Prediction and Disease Identification. International Journal of Experimental Research and Review, 42, 320-327.

DOI: https://doi.org/10.52756/ijerr.2024.v42.028



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 7 tion-NonCommercial-NoDerivatives 4.0 International License.