Computer Science, Engineering and Technology

Vol: 1(2), June 2023

REST Publisher; ISSN: 2583-9179 (Online)

Website: https://restpublisher.com/journals/cset/

DOI: https://doi.org/10.46632/cset/1/2/2

A review on Cotton Leaf Disease Detection with Balanced and Imbalanced Data

[*] Prashant Udawant, Aditya R Joshi, Shell Mahajan, Atharv A Swarge
SVKM's NMIMS MPSTME, Shirpur, Maharashtra, India.
*Corresponding Author Email: Prashant.udawant@nmims.edu.in

Abstract: *Cotton is a major crop in India, and its production is essential to the country's economy. However, cotton plants are susceptible to a number of diseases, which can cause significant yield losses. Early detection of these diseases is crucial, but manual identification can be challenging. This paper reviews the use of machine learning for cotton leaf disease detection. A number of classification methods and algorithms are discussed, including convolutional neural networks (CNNs), random forests, and SMOTE. The results show that machine learning can be used to achieve high accuracy in detecting cotton leaf diseases, even in the case of imbalanced datasets. The paper also discusses the importance of balancing class distributions in order to improve model performance. A number of data-level and classifier-level techniques are evaluated, and the results show that the Weighted Random Forest algorithm is effective in handling imbalanced datasets. The findings of this paper suggest that machine learning is a promising tool for the early detection of cotton leaf diseases. This could help to reduce yield losses and improve the profitability of cotton farming in India.*

Key words: *Cotton Leaf Disease, Machine Learning, CNN, Random Forest, SMOTE, Imbalanced Datasets.*

1. INTRODUCTION

In a country like India, where agriculture serves as one of the most significant economic drivers, cotton cultivation plays a vital role. The nation's economy greatly benefits from the production of cotton, which is a crucial crop. Cotton farming is an essential part of the Indian economy because it provides 2.5% (2022) of the country's GDP and employs about 14% of its population. Environmental factors and weather patterns have a big impact on the productivity and quality of cotton. Plant diseases, on the other hand, are one of the most frequently disregarded elements that might affect the quality of cotton products. Numerous diseases can damage cotton plants, causing aberrant development, modifications to the colour and shape of the leaves, as well as changes to the overall health of the plant. In the realm of agriculture, early detection of these diseases is crucial since they can cause significant yield losses and have a negative influence on the economy. Disease detection on leaves in cotton cultivation is a difficult undertaking that necessitates the application of scientific methods and in-depth observation. The symptoms of many cotton leaf diseases can frequently resemble one another, making manual identification challenging. In order to reliably identify cotton leaf diseases and recommend efficient treatments, an automated method is required. To deal with this problem, machine learning methods like Convolutional Neural Networks (CNN), Random Forest, SMOTE, etc. are frequently used. Large databases of cotton plant photos may be analyzed by these potent algorithms, which can then be trained to distinguish between healthy plants and those with various leaf diseases. Agricultural experts can speed up diagnostics, conduct prompt remedies, and safeguard their cotton crops from potential losses by utilizing machine learning.

2. LITERATURE SURVEY

This paper's main goal is to focus on identifying plant leaf diseases using the texture of leaves. P.Revathi et al. employed homogenize techniques such as the sobel and canny filter in order to find the edges [4].By using Homogeneous Pixel Counting method for Cotton Diseases Detection (HPCCDD) algorithm, the diseases have been categorized very efficiently thus achieving an accuracy of 98.1% [3]. Dheeb Al Bashish et al. developed a neural network classifier based on statistical classification that successfully detected and classified diseases with a 93% precision [5]. Another ANN, the back propagation neural network (BPNN), was attempted by Menukaewjinda et al. [6] for effective grape leaf colour extraction in the presence of a convoluted background. Additionally, they examined the genetic algorithm (GA) and modified self organizing feature maps (MSOFM), and found that both methods enable automatic parameter change for grape leaf disease colour extraction. In order to effectively classify leaf diseases, support vector machines (SVM) have also been discovered to be quite promising [1]. The four main methods for identifying plant diseases are SGDM, K-means clustering, BPNN, and SVM. Zhang Jian used the support vector machine technique to diagnose cucumber diseases. They utilized various kernel functions. The outcomes demonstrated that the SVM method based on the RBF kernel function performed best for classifying cucumber disease when each spot was used as a sample [7].

Jassid Attack: Jassid attacks on cotton leaves are caused by jassids or leafhoppers, attracted to cotton plants due to favourable environmental conditions. Symptoms of jassid infestation include yellowing of leaves, leaf stippling, leaf curling, and reduced plant growth. These attacks have negative effects on cotton plants, such as reduced photosynthesis, yield loss, and increased vulnerability to other pests and diseases.

Aphid attack: Aphid attacks on cotton leaves are primarily caused by the infestation of aphid insects, which are small, soft-bodied pests that feed on plant sap. These attacks usually occur when environmental conditions are favourable, such as during warm and dry periods. Symptoms of aphid infestation on cotton leaves include the presence of sticky honeydew secretions, distorted and curled leaves, yellowing or wilting of foliage, and stunted growth. Aphids can also transmit viral diseases. The overall effect of aphid attacks on cotton leaves is reduced plant vigour and decreased cotton yield.

Cercospora: A common cotton leaf disease caused by the fungal pathogen, Cercospora Gossypina, is known as Cercospora leaf spot [20]. It leads to leaf spot symptoms characterized by small, dark brown to black lesions with yellow halos. The disease can spread rapidly under warm and humid conditions, causing defoliation, reduced photosynthesis, and economic losses for cotton growers.

Alternaria: Alternaria leaf spot is a common disease caused by the fungal pathogen Alternaria Saprophytic [20]. It affects cotton plants and is characterized by dark brown to black lesions on the leaves, which can lead to defoliation, reduced photosynthesis, and increased vulnerability to other pathogens.

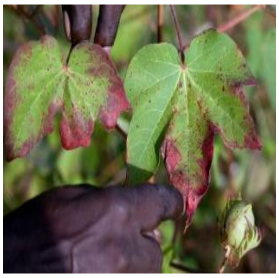
Fig.1: Jassid Attack

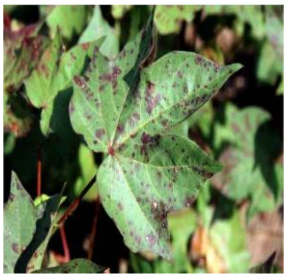
Fig.2: Aphid Attack

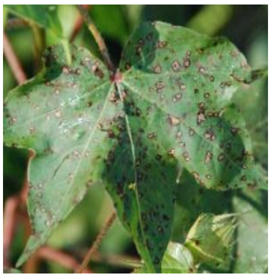
Fig.3: Cercospora

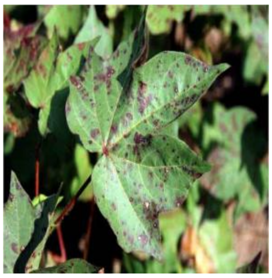
Fig.4: Alternaria

3. MODELS ON BALANCED DATA

When there are nearly equal numbers of examples or samples available for various classes or categories in a dataset, or if not, then the dataset is said to have balanced data in machine learning. In other words, each class is represented similarly, avoiding any notable imbalance that can potentially skew the learning process. It is crucial to have balanced data so that the machine learning algorithm may learn from a wide and representative group of instances from all classes in order to produce accurate and fair models.

KNN : The paper discusses how the KNN algorithm is used to classify unhealthy leaves into different diseases. The KNN algorithm is a machine learning algorithm that works by finding the most similar data points to a new data point and then classifying the new data point based on the categories of the similar data points. In this case, the unhealthy leaves are classified into different diseases based on their similarity to other unhealthy leaves that have already been classified.

The paper shows how the KNN algorithm works in identification of cotton leaf disease in three steps:

- The leaves are first classified either healthy or unhealthy. In the first step, the leaves are classified into two categories: healthy and unhealthy. This step involves a preliminary assessment of the leaves to determine whether they show signs of disease or are considered healthy. Various methods can be used for this initial classification, such as visual inspection by experts, using image processing techniques, or utilizing other sensor-based technologies.

- The unhealthy leaves are then further classified into different diseases using the KNN algorithm. After identifying the unhealthy leaves in the previous step, the next step is to classify them into specific disease categories. This is where the KNN algorithm comes into play. KNN is a supervised machine learning algorithm used for classification tasks. It works based on the principle of finding the k-nearest neighbours to a new data point (in this case, the unhealthy leaf) among the labeled data points (previously identified and categorized unhealthy leaves).

The KNN algorithm saves all available data and categorizes new data points based on their similarity to existing data. It is a widely used supervised type of machine learning algorithm. This means that the KNN algorithm learns from a set of labeled data points (in this case, unhealthy leaves that have already been classified into different diseases). The KNN algorithm then uses this knowledge to classify new data points (in this case, new unhealthy leaves). In this study, various machine learning techniques are used to classify cotton illness. For cotton leaf disease classification, the images are split from the complicated background, and removing the background is tough. This is the modified factorization-based active contour method. We used a modified factorization-based active contour method to remove the background. This method helps in recognizing the required leaf image from the image) was employed in this study to segment the pictures from the background using the cotton leaf image database. First, segmented photos are used to extract the colour and texture properties. It must then be input into machine learning algorithms like K-nearest neighbor, Random Forest, AdaBoost, Naive Bayes, and Multilayer Perceptron [2]. Colour characteristics are sufficient to distinguish between photos of healthy and diseased cotton leaves. Performance metrics like precision, recall, F-measure, and Matthews correlation coefficient were used to assess the classifiers' effectiveness.There were accuracy scores of 93.38%, 90.91%, 95.86%, 92.56%, and 94.21% for Support vector machines, Naive Bayes, Random Forest, AdaBoost, and K-nearest neighbor classifiers, respectively [8]. According to Hossain et al. [10], the Arkansas plant disease database and the Reddit leaf disease database help the KNN classifier perform better classification. After transforming the input RGB image to a l*a*b model, the authors performed colour segmentation. Colour and texture features were retrieved and fed to the KNN classifier, which had a 96.76% accuracy rate. It is among the simplest methods for supervised classification. A new data pointis classified by the KNN algorithm based on similarity to all previously stored data. Because it doesn't automatically learn from the training set but instead saves the dataset and runs an operation on it when classifying, this method is frequently referred to as a lazy learner.

SVM: The classification of the cotton leaf disease spot uses the Multiclass SVM. The colour of the cotton leaf is divided into segments according to the number of weight vectors. To categorize cotton leaf disease, data gathered from the pixels of both healthy and infected leaves is used for training SVM Classifier. Cotton leaf with Alternaria Leaf Spot Fungal Disease (ALSFD), Grey Mildew Cotton Disease (GMCD), and Rust Foliar Fungal Disease (RFFD) was used as the input leaf in this research [21].

TABLE 1. Confusion Matrix for Cotton Leaf Samples [21]

SVM Classifier	**Alternaria Leaf Spot Fungal Disease (ALSFD)**	**Grey-Mildew cotton Disease (GMCD)**	**Rust Foliar Fungal Disease (RFFD)**	**Accuracy**
ALSFD	1123	74	135	84.3%
GMCD	4	1012	9	98.7%
RFFD	32	45	1056	93.2%

The number of samples examined by the SVM classifier is 3490. ALSFD is correctly diagnosed in 1123 samples, while GMCD and RFFD are misclassified in 74 and 135 instances, respectively. ALSFD has an accuracy of 84.3%. For 1012 samples, GMCD is accurately diagnosed; 4 samples are misclassified as ALSFD, and 9 samples are misclassified as RFFD. For GMCD, the accuracy is 98.7%. RFFD is accurately identified in 1056 samples but incorrectly classified in 32 samples as ALSFD and 45 samples as GMCD. For RFFD, the accuracy is 93.2%. The overall accuracy rate is 92.06% [21].

Ensembling of CNN with Random Forest: In this case, the suggested strategy is an ensemble model developed by combining two approaches, namely Convolutional Neural Network (CNN) and Random Forest for feature extraction and classification, respectively [9].

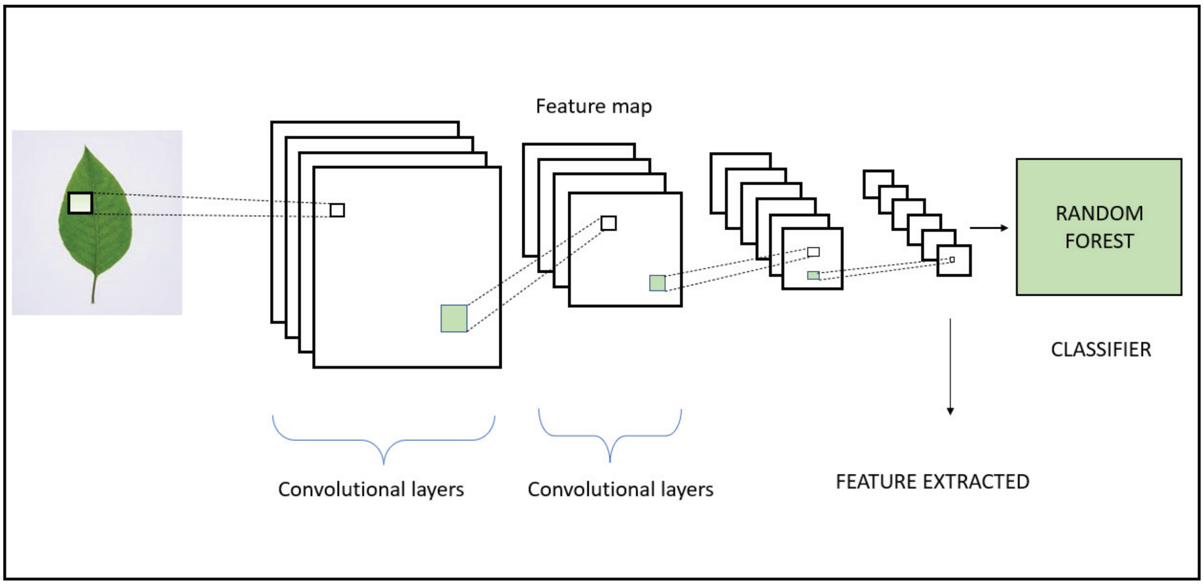

FIGURE 1. Proposed Framework – CNN + RF

Only convolutional layers are utilized in CNN. With a better and deeper understanding, convolutional layers can extract practically all information in terms of size, colour, and shape. The feature mapping process makes use of convolutional layers. These layers are employed for feature extraction. These features will now be flattened, and typically, they must be transmitted to completely linked layers. However, another method used is already an ensemble base method i.e., Random Forest [14]. This will classify for the classification of diseased plants and types of diseases [9]. Convolutional layers assist in removing the diseased area from a plant leaf picture. The max-pooling layers assist in feature mapping and dimensionality reduction, both of which speed up computation. [16] The output of each learner is gathered as a vote in Random Forest, where a number of decision trees are employed as base learners. The input instance is finally identified as belonging to the class with a higher frequency.

4. MODELS ON IMBALANCED DATA

When the classes or categories in the training data are not distributed equally, the dataset is said to be imbalanced in machine learning. There are a lot more examples of one class than the others. As a result, biased models that attempt to anticipate the minority class may end up failing. In recent years, a number of advanced learning algorithms have been proposed to handle the classification issue in the imbalance dataset [11]. Data level methods and algorithm level methods are the two major ways covered for learning imbalanced data. [12]. The main goal of data-level approaches is to balance the class distribution in the training data before supplying it to the learning algorithm. To produce a more balanced dataset, these methods directly alter the data points. Both random undersampling and random oversampling are used in data-driven approaches. Undersampling methods come in many forms, including Random Undersampling (RUS), Tomek Links, etc. Oversampling, the second widely used sampling strategy, involves randomly duplicating items from minority classes until they have a uniform representation. Focused oversampling, synthetic sampling, random oversampling, and certain cutting-edge heuristic techniques like Synthetic Minority Oversampling Technique (SMOTE) are only a few of the oversampling techniques available [15]. SMOTE is a method that creates artificial samples to oversample the minority class's object [15]. Various Data level methods implemented by Kubat&Matwin on Oil data set which has 50 features and 937 samples, out of which 41 are slick oil samples (minority class) and 896 are non-slick oil samples (majority class) [19].

TABLE 2. Performance comparison on oil spill dataset [19]

Methods	Recall	Acc-	Precision	F-measure	G-mean	Wt. Accuracy
One-sided Sampling	76.0	86.6	20.53	32.33	81.13	81.3
SHRINK	82.5	60.9	8.85	15.99	70.88	71.7
SMOTE 500% + Down 50%	89.5	78.9	16.37	27.68	84.03	84.2
SMOTE 500% + Down 100%	78.3	68.7	10.26	18.14	73.34	73.5
BRF cutoff = 0.5	73.2	91.6	28.57	**41.10**	81.19	82.4
BRF cutoff = 0.6	85.4	84	19.66	31.96	84.7	84.7
BRF cutoff = 0.7	92.7	72.2	13.24	23.17	81.18	82.5
WRF weight = 41:896	92.7	82.4	19.39	32.07	**87.4**	**87.6**

The classifier level method, also known as the algorithm driven approach, keeps the training dataset constant while modifying the inference algorithm to make it easier to learn especially about the minority class. This method combines conventional approaches such as one-class classification, thresholding, and cost-sensitive learning with hybrid techniques such as ensemble learning [17]. Thresholding method contains multiple methods which are used in adjusting the decision boundary as the threshold moves. Certain algorithms such as Naive Bayes can produce probability estimates and translate them into forecasts. Cost-sensitive involves changes to the learning rate such that more examples with greater costs contribute overall to update the weighting.

TABLE 3. Learning strategies and limitations of various classifier algorithms for imbalanced data [11][18]

Classifier learning Algorithm	**Learning Strategy**	**Limitations**
Random Forest	Combine the output generated by multiple classifier	Non-linear correlation between independent and dependent variables causes bias
K-Nearest Neighbour	It determines class label based on upper rank class among k nearest neighbours	Carries the higher possibilities of instances from frequent class
Decision Trees	If a problem arises, subdivide training data coercively and shorten sub trees	Many division required To find imbalance
Neural Networks	Weight is adjusted repeatedly to minimize error	Minimize errors only for frequent classes.
Naive Bayes	Consider flexibility among features	Wrong classification of instances of small class
Support Vector Machine	Find the best hyperplane dissociation with the maximal margin.	Sensitive to imbalanced datasets and biased decision boundarytowards minority class

5. RESULT ANALYSIS

This review article primarily examines the many approaches and algorithms used to classify different datasets using both balanced and unbalanced data. We examined a number of algorithms for the classification of balanced data, of which Random Forest and CNN provided 92% and 93% accuracy respectively. However, by combining the CNN and Random Forest algorithms, we were able to achieve the highest accuracy of 99.89%, followed by the KNN algorithm with an accuracy of 96.76% and the Support Vector Machine (SVM) algorithm with an accuracy of 92.06%.

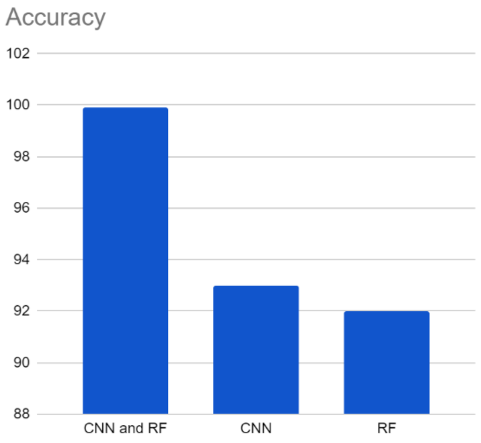

FIGURE 2. Methods vs Accuracy on Balanced Data

Further, we explored methods to handle the issue of imbalanced data for the classification purpose. In the oil dataset, the highest accuracy of 87.6% is achieved by implementing the Weighted Random Forest algorithm with the weight of 41:896.

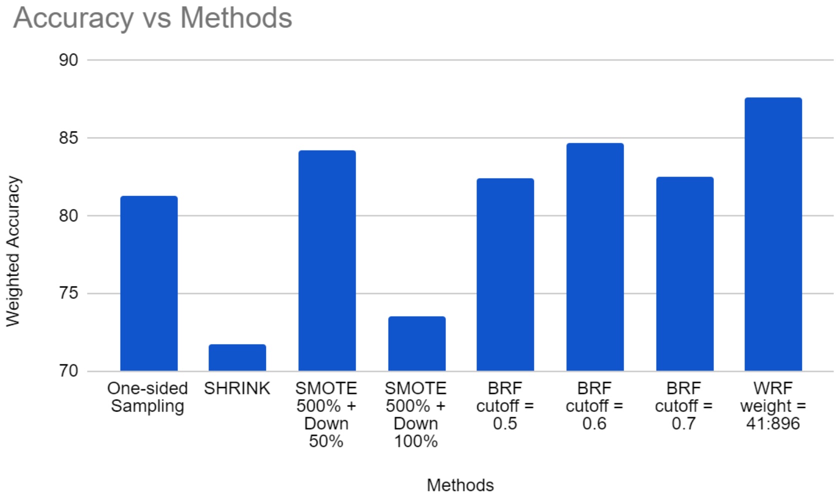

FIGURE 3. Methods vs Accuracy on Imbalanced Data

6. CONCLUSION

This review paper mainly focuses on various classification methods and algorithms for plant leaf disease detection, focusing on both balanced and imbalanced datasets. In order to identify disease using leaf textures, a number of algorithms and methods were investigated, with considerable accuracy gains. The ensemble of Convolutional Neural networks (CNN) and Random Forest exhibited outstanding performance, with a remarkable accuracy of 99.89% for balanced data. Machine learning algorithms can be severely limited by imbalanced training data, resulting in biased models that struggle to accurately predict the minority class. In the study, data-level strategies including random under-sampling, over-sampling with methods like SMOTE, and hybrid methods combining thresholding and cost-sensitive learning were all evaluated. Additionally, classifier-level techniques like ensemble learning were taken into account. In handling the imbalanced oil dataset, the results showed that the Weighted Random Forest algorithm, with a weight of 41:896, had the maximum accuracy of 87.6%. These findings highlight the significance of using the right strategies to properly balance class distributions and lessen the effect of class imbalance on model performance.

7. FUTURE SCOPE

As the agricultural industry continues to adopt machine learning for plant disease detection, addressing imbalanced data remains an important challenge to be solved. Future research in this area should focus on refining existing algorithms, exploring novel approaches, and integrating domain-specific knowledge to further enhance the accuracy and reliability of disease classification models. Researchers and data scientists should focus on discovering creative techniques to deal with imbalanced data and avoiding biased models. To maximize the learning process on imbalanced datasets, advanced ensemble techniques, hybrid algorithms, and innovative data-level methodologies should be developed. We can construct strong and accurate plant leaf disease classifiers by constantly refining and enhancing algorithms to deal with imbalanced data, providing farmers with reliable tools for early disease detection and efficient crop disease management. Finally, these advancements will have a considerable impact on enhancing agricultural productivity and global food security in the years to come.

REFERENCES

[1]. Ms. Rupali S.Zambre et al Int. Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 5(Version 1), May 2014, pp. 92–97

[2]. Kotian, S., Ettam, P., Kharche, S., Saravanan, K., &Ashokkumar, K. (2022). Cotton Leaf Disease Detection Using Machine Learning. In SSRN Electronic Journal. Elsevier BV. https://doi.org/10.2139/ssrn.4159108

[3]. Revathi, P., &Hemalatha, M. (2012). Classification of cotton leaf spot diseases using image processing edge detection techniques. In 2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET). 2012 International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET). IEEE. https://doi.org/10.1109/incoset.2012.6513900

[4]. Mr. Viraj A. Gulhane ,Dr. Ajay A. Gurjar " Detection of Diseases on Cotton Leaves and Its Possible Diagnosis ", International Journal of Image Processing (IJIP), Volume (5) : Issue (5) : 2011.

[5].Al Bashish, D., Braik, M., & Bani-Ahmad, S. (2010). A framework for detection and classification of plant leaf and stem diseases. In 2010 International Conference on Signal and Image Processing. 2010 International Conference on Signal and Image Processing (ICSIP). IEEE. https://doi.org/10.1109/icsip.2010.5697452

[6].Hannuna, S., N, A., Subramanian, S., Prashant, S., Jhunjhunwala, A., & C, N. C. (2011). AGRICULTURE DISEASE MITIGATION SYSTEM. In ICTACT Journal on Communication Technology (Vol. 02, Issue 02, pp. 363–369). ICT Academy. https://doi.org/10.21917/ijct.2011.0050

[7].Zhang Jian, & Zhang Wei. (2010). Support vector machine for recognition of cucumber leaf diseases. In 2010 2nd International Conference on Advanced Computer Control. 2010 2nd International Conference on Advanced Computer Control. IEEE. https://doi.org/10.1109/icacc.2010.5487242

[8].Patil, B. M., &Burkpalli, V. (2021). A Perspective View of Cotton Leaf Image Classification Using Machine Learning Algorithms Using WEKA. In F. Bellotti (Ed.), Advances in Human-Computer Interaction (Vol. 2021, pp. 1–15). Hindawi Limited. https://doi.org/10.1155/2021/9367778

[9].Saxena, S., & Rathor, S. (2023). An Ensemble-Based Model of Detecting Plant Disease using CNN and Random Forest. In 2023 6th International Conference on Information Systems and Computer Networks (ISCON). 2023 6th International Conference on Information Systems and Computer Networks (ISCON). IEEE. https://doi.org/10.1109/iscon57294.2023.10112023

[10].Hossain, E., Hossain, Md. F., &Rahaman, M. A. (2019). A Colour and Texture Based Approach for the Detection and Classification of Plant Leaf Disease Using KNN Classifier. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE. https://doi.org/10.1109/ecace.2019.8679247

[11].Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. In IOP Conference Series: Materials Science and Engineering (Vol. 1099, Issue 1, p. 012077). IOP Publishing. https://doi.org/10.1088/1757-899x/1099/1/012077

[12].Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. In Information Sciences (Vol. 513, pp. 429–441). Elsevier BV. https://doi.org/10.1016/j.ins.2019.11.004

[13].Kotsiantis S, Kanellopoulos D, Pintelas P et al. 2006 GESTS International Transactions on Computer Science and Engineering 30 25–36

[14].Ramesh, S., Hebbar, R., M., N., R., P., N., P. B., N., S., & P.V., V. (2018). Plant Disease Detection Using Machine Learning. In 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C). 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control (ICDI3C). IEEE. https://doi.org/10.1109/icdi3c.2018.00017

[15].Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. In Journal of Artificial Intelligence Research (Vol. 16, pp. 321–357). AI Access Foundation. https://doi.org/10.1613/jair.953

[16].Yadav, D. P., Sharma, A., Singh, M., & Goyal, A. (2019). Feature Extraction Based Machine Learning for Human Burn Diagnosis From Burn Images. In IEEE Journal of Translational Engineering in Health and Medicine (Vol. 7, pp. 1–7). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jtehm.2019.2923628

[17].Buda, M., Maki, A., &Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. In Neural Networks (Vol. 106, pp. 249–259). Elsevier BV. https://doi.org/10.1016/j.neunet.2018.07.011

[18].Duncan, D., Nouaili, N., Garner, R., Salehi, S., & Rocca, M. L. (2022). Key Radiological Features of COVID-19 Chest CT Scans with a Focus on Special Subgroups: A Literature Review. In Current Medical Imaging Reviews (Vol. 19, Issue 5, pp. 442–455). Bentham Science Publishers Ltd. https://doi.org/10.2174/1573405618666220620125332

[19].Chen, C., Liaw, A. & Breiman, L. (2004) Using random forest to learn imbalanced data, Dept. Statistics, Univ. California, Berkeley, CA, Tech. Rep.666.

[20].Bruguière, A., Le Ray, A., Bréard, D., Blon, N., Bataillé, N., Guillemette, T., Simoneau, P., &Richomme, P. (2016). Identifying Natural Products (NPs) as potential UPR inhibitors for crop protection. In Planta Medica (Vol. 81, Issue S 01, pp. S1–S381). Georg Thieme Verlag KG. https://doi.org/10.1055/s-0036-1596783

[21].Kumari, Ch. U., Vignesh, N. A., Panigrahy, A. K., Ramya, L., & Padma, T. (2019). Fungal Disease in Cotton Leaf Detection and Classification using Neural Networks and Support Vector Machine. In International Journal of Innovative Technology and Exploring Engineering (Vol. 8, Issue 10, pp. 3664–3668). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP. https://doi.org/10.35940/ijitee.j9648.0881019.