An Ingenious Approach to Investigate the Prediction Rate and Accuracy for Plant Leaf Disease Identification Using Decision Tree Algorithm over K-Nearest Neighbor

L.Lakshmi Narendra M[1] , K. Malathi[2*]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, 602105.

[2*]Project Guide,Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India, 602105.

ABSTRACT

Aim: The study's goal is to identify plant leaf disease to find the best accuracy utilizing machine learning techniques such as K-Nearest Neighbor (KNN) and Innovative Decision Tree (DT). **Methods and Materials:**The data set in this paper utilizes the publicly available Kaggle data set for plant leaf disease detection. The sample size of classification of leaf disease detection with improved accuracy rate was sample 80 (Group 1=40 and Group 2 =40) and calculation is done with G-power 0.8 through alpha and beta qualities are 0.05, 0.2 with a confidence interval at 95%. Accuracy is performed with the dataset from the Kaggle library. The two groups are K-Nearest Neighbor (N=20) and InnovativeDecision Tree algorithms (N=20). **Results:** An Innovative DT is used for detection of Plant Leaf disease. Accuracy is analyzed based on disease images of 92.37% where the KNN has the accuracy of 75.63%. The two algorithms DT and KNN are statistically satisfied with the independent sample T-Test (α=.001) value ($p<0.05$) . **Conclusion:** Identification of plant leaf disease significantly seems to be better in Innovative DT than KNN.

Keywords:K-Nearest Neighbor , Innovative Decision Tree, Machine Learning, Image processing, Plant Leaf Disease, Detection.

INTRODUCTION

Identification of plant leaves disease is one of the serious issues that each framework faces.Now a days,there is no application based framework yet to recognize plant leaf infection (Ahmed et al. 2019). It is vital to acknowledge plant leaf diseases that are classified as varied diseases . If diseases are , then it's potential to acknowledge the diseases by an individual (Jaisakthi et al. 2019).The plants would have completely different diseases and visual manifestations on the plant leaves. it's impossible to observe the sickness with the human eye and therefore the disease can cause the plants to become health problem that ends up in decrease the yield of the crop once the crop production is attenuated the economy of the country decrease each country will rely upon the agriculture to feed their individuals (Prem et al. 2018). Identification of plant plant disease would have the information labeling,model training and model abstract thought. It has wide applications in numerous fields like scientific research (Liu and Wang 2021; Prem et al. 2018).

Identification of Plant Leaf Disease is conducted by researchers in order to increase business. In all, 20 related articles have been published in IEEE, and 6 have been published in Google Scholar like ResearchGate, Sciencedirect. (Dhaware and Wanjale 2017) identification of plant leaf disease with the implementation of decision trees tells regarding the finding of assorted unwellness that depends on three conditions host plants liable to disease and therefore the ralated atmosphere and viable infectious agent has got the accuracy of 95.87%. (Jaisakthi et al. 2019) the plant leaf disease identification with several images processed by the image process supported the disease detection of assorted plants. There are numerous image process options to discover color based mostly like RGB, native options on pictures like scale-invariant feature transformation (SIFT) area unit a number of the classification models of the plant leaf disease with the accuracy of 86.59% . (Ennouni, Sabri, and Aarab 2021) recognizable proof of plant leaf illness characterized as foundation image utilizing snatch cut division strategy .From the division of leaf a part of infected district with fluctuated division can get 83.54% of decision tree algorithm results. (Azim et al. 2021) implementation of the plant leaf disease with determined ways for locating the illness and also the house of infected.The plant can cause the deficiency of heath and also the plant to show dead set be less helpful as 82.25% . (Dhaware and Wanjale 2017) has implemented the best accuracy of plant leaf disease classification that has provided results by victimization dataset with sample image and that are massive and unbiased. The trained dataset had an accuracy of 85.69%.(Bhavikatti et al. 2021; Karobari et al. 2021; Shanmugam et al. 2021; Sawant et al. 2021; Muthukrishnan 2021; Preethi et al. 2021; Karthigadevi et al. 2021; Bhanu Teja et al. 2021; Veerasimman et al. 2021; Baskar et al. 2021)

Based on the literature survey, the KNN has very little accuracy, the correctness of the leaf is shown in a very low percentage while analyzing the plant disease and therefore the manual input isn't attainable to feature to the dataset. The study aims to enhance the accuracy of the morbid leaf , raise the correctness proportion of the identification of plant plant disease, and scale back knowledge while training and testing the dataset.

METHODS AND MATERIALS

The suggested work is being studied in the OOAD laboratory, Computer Science and Engineering Department, Saveetha School of Engineering, SIMATS, Chennai. Sample size is calculated by using clincalc.com by keeping G power (Kane, Phar, and BCPS) the calculation is performed utilizing G-power and The minimum power of the analysis is set at 0.8 with alpha, and the maximum allowed error is set at 0.5 with beta quality with threshold value as 0.05% and Confidence Interval is 95%. Mean and standard deviation has been calculated based on the previous literature for size calculation. The two groups are used, namely K-Nearest Neighbor (N=10) as an existing model as a group. 1 and Decision Tree (N=10) as a Proposed model as a group. 2.

DATA PREPARATION

The Decision Tree is to identify the plant leaf diseases that are stored in the dataset, to train and test through the Kaggle dataset. The dataset includes 1000 data in the form of images which are taken as a sample from the University of America students with their plant leaf disease. There are 750 training images and 250 test images (Caldeira, Santiago, and Teruel 2021). The sample images of leaves present in the dataset have been shown in Fig. 1.

K-Nearest Neighbor(KNN)

KNN classifier characterizes the illnesses such as alternaria alternata, anthracnose, bacterial scourge, leaf spot, and plant infection.Plant plant disease detection and classification using laptop vision and machine learning strategies. The raw image of a leaf is pre-processed, segmental and options like form, color, texture, vein etc. are extracted (Jasim and AL-Tuwaijari 2020). KNN is considered as a non-parametric method and for distribution of the data, it does not make any underlying assumptions (Singh and Minj 2019). Compute the distance between the inquiry example and every one the training examples (Sharma 2021) . kind the distance and regulate neighboring neighbors centered on the K-th minimum distance ,Fig. 2 represents the working flow of the K-Nearest Neighbor.

KNN algorithm steps as follows ,

1. The image will pre processing
2. For Feature Extraction, use the GLCM algorithm
3. Training the dataset
4. Detection of disease

Decision Tree

The Innovative Decision Tree is a tree type structure that partitions the entire dataset into essentially unrelated areas and every area incorporates a category mark that depicts all of the data focuses connected with the dataset. The Decision Tree depends on the possibility of non-parametric administered learning strategy (Jaisakthi et al. 2019). The arrangement tree portrays the formation of a " twofold choice tree". Fig. 3 flow has been mentioned within the below steps .

DT algorithm steps as follows ,

1. Collect the input image from the user
2. Pre process the image with RGB to gray conversion
3. Feature extraction
4. Detection of disease

For comparing both the models, the dataset has been trained with five different sample sizes. the accuracy values are recorded.

The system configuration is used for the algorithm to run in a 64 - bit Operating System, 4GB RAM PC, and using Windows 10, Google Colab, and Microsoft Office for software specification.

The training model's performance is estimated using the data that has been split for training and testing to validate the dataset. Then load and reshape the data arrays to categorize the numbers. Normalize the pixel values of grayscale images All the layers will be functioned through the ReLU activation function to the categorical cross_entrophy to find the loss function. The model will be evaluated with the fit() function which has the metrics function to validate the accuracy of the data.

Statistical Analysis

SPSS is used for statistical analysis of K-Nearest Neighbor (KNN) algorithm and Innovative Decision Tree algorithm. The independent sample t test was performed to find the standard deviation, and standard error mean statistical significance between the groups, and then comparison of the two groups with the SPSS software.

RESULTS

The DT algorithm forms the layers with all the images of each number, whenever it runs at different times due to the initialization of sample size (N=20). The layers formed due to the iterations, the accuracy value changes with the duration of running time and produces the accuracy concerning the period which is shown in Table 1. DT has better accuracy than the KNN due to the activation functions and metrics, the KNN algorithm has not compatible with the advanced activation functions which are only restricted to the adam, adaleta, and adagrad which takes more time and the functions are not taking the whole data to analyze the diseased leaf in the dataset whereas the DT takes the data and forms layers with each leaf is individually and finally gives the result. Concerning the activation functions, the Accuracy has changed and has proven that DT is better than the KNN.

Table 1 represents the data collection from the N=20 samples of the dataset for KNN with the size of 28*28 pixels to gain accuracy (%) and DT to gain accuracy (%) is calculated based on eqn. 1.

$$Accuracy = (TP + TN)/\,(TP + FP + TN + FN) \quad (1)$$

Where,TP = True Positive
TN - True Negative
FP - False Positive
FN - False Negative

Loss: A scalar worth that endeavor to limit during our preparation of the model. The lower the misfortune, the nearer our expectations are to the genuine names.

The IBM SPSS version 21 statistical software is used for our study. Shape and size are independent variables, and size is the dependent variable accuracy (%). For our study Identification of Plant Leaf Disease

The datasets are created in SPSS with N=20 being the sample size for K Nearest Neighbor and Decision Tree. The grouping variable is GroupID, and the testing variable is accuracy. GroupID is given as 1 for KNN and group 2 for DT. Group Statistics is applied for the Statistical Package for the Social Sciences (SPSS) dataset and shown in Table 2. By performing the statistical analysis group statistics represents the comparison of the accuracy of Identification of plant leaf disease of KNN and DT. The DT algorithm had the highest accuracy (92.67)). CTC had the lowest accuracy (89.07) in Table 2.

Table 3 represents the Independent Sample T-Test, which is used for sample collections by fixing the level of significance as 0.005 with a confidence interval of 95 %. After applying the SPSS calculation, DT has accepted a statistically significant value($P<0.05$). From Fig. 4 it was represented by a simple bar Mean of Accuracy KNN error range (0.82 - 0.91) error range (2-4) and DT error range (0.91 - 0.92) .

DISCUSSION

Our overall results indicate that there are some variances in the accuracy values due to the advancements of the activation functions which proved that the Decision Tree with an accuracy of 92.37% is better than the Identification of plant leaf disease with an accuracy of 75.63% in recognizing the Leaf disease. In this case, there is a statistically significant difference. innovative Identification of plant leaf disease accuracy of two algorithms having the significant accuracy value of 0.001(p<0.005 Independent Sample t-Test).

The Innovative decision tree (DT) The clustering technique is used to segment input photos. The decision tree classification is applied which will classify the input image into two classes. This improves disease detection accuracy while also categorizing the data into different classifications. Furthermore, based on disease detection the system sprays fertilizers and pesticides, which reduces human work innovative optimized performance to 85.67% (Chopda et al. 2018)).The There was a pre-processing step, noise management, picture improvement, and transformation. Then, characteristics based on form, texture, and color were extracted. After normalizing the data, five machine learning algorithms were applied to the dataset for categorization. Finally, the algorithms' classification accuracies were determined using the Multilayer Perceptron algorithm's accuracy curves 75.5% of accuracy (Aurangzeb et al. 2020). A model was built in to classify the disease based only on the extracted percentage of the RGB value of the diseased region of rice leaf using image processing. The RGB percentages were input into a Naive Bayes classifier, which classified the illnesses into three groups: Bacterial leaf blight, Rice blast, and Brown spot. The model's accuracy in classifying disorders is more than 89 percent (Sharma 2021). To distinguish between healthy and unhealthy leaves, Random Forest, an ensemble learning approach, is utilized. The authors employed the Histogram of Oriented Gradient to extract picture characteristics (HOG). Their work was stated to be 92.33 percent accurate (Ahmed et al. 2019).

In the future, improve this classification for further development in Decision Tree Architecture and the applications of some big complex noisy data. Improve this system to recognize various diseases . Systems are to be developed to analyze plant leaf disease .

Future work will include in picture preparation and spreading use of the model via preparing the for plant sickness acknowledgment on more extensive land regions (Sladojevic et al. 2016). Furthermore, the work might be on the shadowing methods which produce best outcomes. Broadening this exploration, there is a desire to accomplish a significant effect on economical turn of events, influencing crop quality for people in the future.

CONCLUSION

In this research, the innovative identification of plant leaf disease performed using the Kaggle dataset seems to be better accuracy (92.37%) using the innovative Decision Tree than the K-Nearest Neighbor (75.63%). The clarity of plant leaf disease found with good accuracy is achieved.

DECLARATIONS

Conflict of interests

There are no conflicts of interest in this paper.

Authors Contributions:

Author LLN was involved in conceptualization, data collection, data analysis, manuscript writing. Author KM was involved in conceptualization, guidance, and critical review of the manuscript.

Acknowledgments:

The authors would like to thank Saveetha School of Engineering and Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the infrastructure required to complete this study effectively.

Funding:

Thanks to the following organizations for their financial assistance, which allowed us to complete the study.

1.Chipontime Technologies Pvt.Ltd, Bangalore

2. Saveetha University

3. Saveetha Institute of Medical and Technical Sciences.

4. Saveetha School of Engineering.

REFERENCES

1. Ahmed, Kawcher, Tasmia Rahman Shahidi, Syed Md Irfanul Alam, and Sifat Momen. 2019. "Rice Leaf Disease Detection Using Machine Learning Techniques." *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*. https://doi.org/10.1109/sti47673.2019.9068096.
2. Aurangzeb, Khursheed, Farah Akmal, Muhammad Attique Khan, Muhammad Sharif, and Muhammad Younus Javed. 2020. "Advanced Machine Learning Algorithm Based System for Crops Leaf Diseases Recognition." *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*. https://doi.org/10.1109/cdma47397.2020.00031.
3. Azim, Muhammad Anwarul, Mohammad Khairul Islam, Md Marufur Rahman, and Farah Jahan. 2021. "An Effective Feature Extraction Method for Rice Leaf Disease Classification." *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. https://doi.org/10.12928/telkomnika.v19i2.16488.
4. Baskar, M., R. Renuka Devi, J. Ramkumar, P. Kalyanasundaram, M. Suchithra, and B. Amutha. 2021. "Region Centric Minutiae Propagation Measure Orient Forgery Detection with Finger Print Analysis in Health Care Systems." *Neural Processing Letters*, January. https://doi.org/10.1007/s11063-020-10407-4.
5. Bhanu Teja, N., Yuvarajan Devarajan, Ruby Mishra, S. Sivasaravanan, and D. Thanikaivel Murugan. 2021. "Detailed Analysis on Sterculia Foetida Kernel Oil as Renewable Fuel in Compression Ignition Engine." *Biomass Conversion and Biorefinery*, February. https://doi.org/10.1007/s13399-021-01328-w.
6. Bhavikatti, Shaeesta Khaleelahmed, Mohmed Isaqali Karobari, Siti Lailatul Akmar Zainuddin, Anand Marya, Sameer J. Nadaf, Vijay J. Sawant, Sandeep B. Patil, Adith Venugopal, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "Investigating the Antioxidant and Cytocompatibility of Mimusops Elengi Linn Extract over Human Gingival Fibroblast Cells." *International Journal of Environmental Research and Public Health* 18 (13). https://doi.org/10.3390/ijerph18137162.
7. Caldeira, Rafael Faria, Wesley Esdras Santiago, and Barbara Teruel. 2021. "Identification of Cotton Leaf Lesions Using Deep Learning Techniques."

Sensors 21 (9). https://doi.org/10.3390/s21093169.

8. Chopda, Jayraj, Hiral Raveshiya, Sagar Nakum, and Vivek Nakrani. 2018. "Cotton Crop Disease Detection Using Decision Tree Classifier." *2018 International Conference on Smart City and Emerging Technology (ICSCET)*. https://doi.org/10.1109/icscet.2018.8537336.
9. Dhaware, Chaitali G., and K. H. Wanjale. 2017. "A Modern Approach for Plant Leaf Disease Classification Which Depends on Leaf Image Processing." *2017 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/iccci.2017.8117733.
10. Ennouni, Assia, My Abdelouahed Sabri, and Abdellah Aarab. 2021. "Plant Diseases Detection and Classification Based on Image Processing and Machine Learning." *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-3-030-72588-4_20.
11. Jaisakthi, S. M., P. Mirunalini, D. Thenmozhi, and Vatsala. 2019. "Grape Leaf Disease Identification Using Machine Learning Techniques." *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*. https://doi.org/10.1109/iccids.2019.8862084.
12. Jasim, Marwan Adnan, and Jamal Mustafa AL-Tuwaijari. 2020. "Plant Leaf Diseases Detection and Classification Using Image Processing and Deep Learning Techniques." *2020 International Conference on Computer Science and Software Engineering (CSASE)*. https://doi.org/10.1109/csase48920.2020.9142097.
13. Karobari, Mohmed Isaqali, Syed Nahid Basheer, Fazlur Rahman Sayed, Sufiyan Shaikh, Muhammad Atif Saleem Agwan, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. "An In Vitro Stereomicroscopic Evaluation of Bioactivity between Neo MTA Plus, Pro Root MTA, BIODENTINE & Glass Ionomer Cement Using Dye Penetration Method." *Materials* 14 (12). https://doi.org/10.3390/ma14123159.
14. Karthigadevi, Guruviah, Sivasubramanian Manikandan, Natchimuthu Karmegam, Ramasamy Subbaiya, Sivasankaran Chozhavendhan, Balasubramani Ravindran, Soon Woong Chang, and Mukesh Kumar Awasthi. 2021. "Chemico-Nanotreatment Methods for the Removal of Persistent Organic Pollutants and Xenobiotics in Water - A Review." *Bioresource Technology* 324 (March): 124678.
15. Liu, Jun, and Xuewei Wang. 2021. "Plant Diseases and Pests Detection Based on Deep Learning: A Review." *Plant Methods* 17 (1): 1–18.
16. Muthukrishnan, Lakshmipathy. 2021. "Nanotechnology for Cleaner Leather Production: A Review." *Environmental Chemistry Letters* 19 (3): 2527–49.
17. Preethi, K. Auxzilia, K. Auxzilia Preethi, Ganesh Lakshmanan, and Durairaj Sekar. 2021. "Antagomir Technology in the Treatment of Different Types of Cancer."

Epigenomics. https://doi.org/10.2217/epi-2020-0439.

18. Prem, G., M. Hema, Laharika Basava, and Anjali Mathur. 2018. “Plant Disease Prediction Using Machine Learning Algorithms.” *International Journal of Computer Applications*. https://doi.org/10.5120/ijca2018918049.
19. Sawant, Kashmira, Ajinkya M. Pawar, Kulvinder Singh Banga, Ricardo Machado, Mohmed Isaqali Karobari, Anand Marya, Pietro Messina, and Giuseppe Alessandro Scardina. 2021. “Dentinal Microcracks after Root Canal Instrumentation Using Instruments Manufactured with Different NiTi Alloys and the SAF System: A Systematic Review.” *NATO Advanced Science Institutes Series E: Applied Sciences* 11 (11): 4984.
20. Shanmugam, Vigneshwaran, Rhoda Afriyie Mensah, Michael Försth, Gabriel Sas, Ágoston Restás, Cyrus Addy, Qiang Xu, et al. 2021. “Circular Economy in Biocomposite Development: State-of-the-Art, Challenges and Emerging Trends.” *Composites Part C: Open Access* 5 (July): 100138.
21. Sharma, Shivam. 2021. “KNN - The Distance Based Machine Learning Algorithm.” May 15, 2021. https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/.
22. Singh, Ramesh Kumar, and Jasmine Minj. 2019. “Detection of Bacterial and Fungal Leaf Diseases Using Image Processing and Machine Learning Techniques.” *International Journal of Computer Sciences and Engineering*. https://doi.org/10.26438/ijcse/v7i3.11261129.
23. Sladojevic, Srdjan, Marko Arsenovic, Andras Anderla, Dubravko Culibrk, and Darko Stefanovic. 2016. “Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification.” *Computational Intelligence and Neuroscience* 2016 (June): 3289801.
24. Veerasimman, Arumugaprabu, Vigneshwaran Shanmugam, Sundarakannan Rajendran, Deepak Joel Johnson, Ajith Subbiah, John Koilpichai, and Uthayakumar Marimuthu. 2021. “Thermal Properties of Natural Fiber Sisal Based Hybrid Composites – A Brief Review.” *Journal of Natural Fibers*, January, 1–11.

TABLES AND FIGURES

Table 1. Data collection from the N=10 samples of the dataset for KNN with the size of 28*28 pixels to gain accuracy (%) and DT to gain accuracy(%)

Samples (N)	**K-Nearest Neighbor**	**Decision Tree**
	Accuracy(%)	**Accuracy(%)**
1	51.29	92.70
2	63.54	92.14

3	69.25	92.17
4	73.59	92.57
5	77.12	92.65
6	78.30	92.45
7	80.24	92.25
8	88.80	92.57
9	85.87	92.78
10	89.70	92.47

Table 2. Comparison of the accuracy of Identification of plant leaf disease of KNN and DT.DT algorithm had the highest accuracy (92.67%). KNN had the lowest accuracy (75.63%)

Groups		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	KNN	10	75.6320	11.82487	3.73935
	DT	10	92.3750	.20887	.06605

Table 3. The Independent Sample T-Test is used for the sample collections by fixing the level of
significance as 0.05 with confidence interval as 95 %. After applying the SPSS calculation, DT has accepted a statistically significant value(P<0.05).

		Levene's test for equality of variances		**T-test for equality means**						
		f	sig	t	df	Sig. (2-tailed)	Mean difference	Std.Error difference	95% confidence interval	
									Lower	Upper
Accuracy	Equal variances assumed	16.098	.001	-4.477	18	.000	-16.743	3.73993	-24.600	-8.8856
	Equal variance not assumed			-4.477	9.006	.002	-16.743	3.7399	-25.202	-8.2834

Fig. 1. Identification of plant leaf disease

FLOWCHART

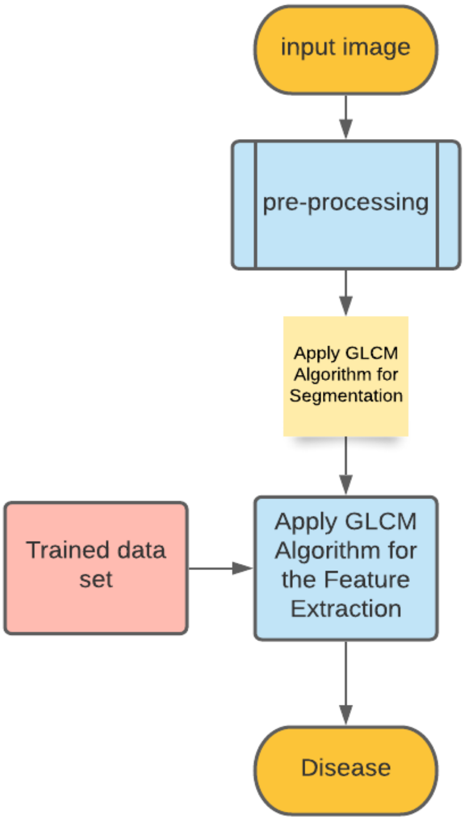
input image
pre-processing
Apply GLCM Algorithm for Segmentation
Trained data set
Apply GLCM Algorithm for the Feature Extraction
Disease

Fig. 2. Flow chart of K-Nearest Neighbor

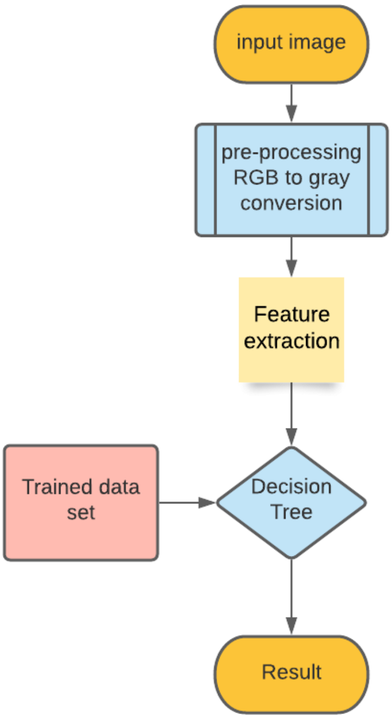

Fig. 3. Flow chart of Decision Tree

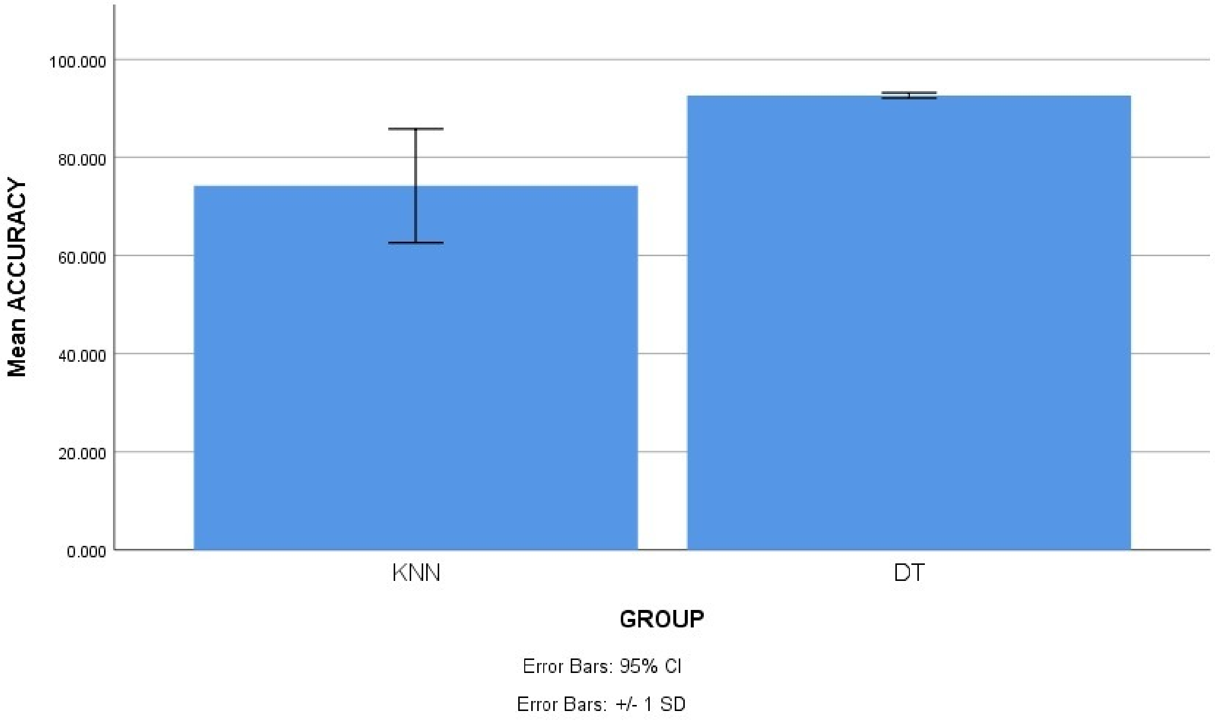

Fig. 4. Simple Bar Mean of Accuracy KNN error range (0.82 - 0.91)and Decision Tree error range (0.91 - 0.92) with Mean accuracy of detection ± 1 SD.X Axis as KNN and Y-Axis as DT Mean accuracy of detection ± 1 SD.