A Machine Learning Methodology For Diagnosing Chronic Kidney Disease

Naviya S[1], Sangavi S[2], Anitha M[3]
[1, 2] Dept of Computer Science and Engineering
[3] Assistant Professor, Dept of Computer Science and Engineering
[1, 2, 3] Kingston Engineering College , Vellore-59

Abstract- *Machine learning and Feature extractions are playing a vital role in internet and health domain. Chronic kidney disease (CKD) is a global health problem with high morbidity and mortality rate, and it cover other diseases. Since there are no obvious manifestation during the early stages of CKD, patients often fail to notice the disease. Early observation of CKD enables patients to receive timely treatment to ameliorate the progression of this disease. Machine learning models can effectively aid clinicians achieve this objective due to their fast and accurate recognition performance. In this study, we propose a machine learning methodology for discover CKD.*

The CKD data set was obtained from the University of California Irvine (UCI) machine learning repository, which has a large number of absence of data. KNN imputation was used to fill up in the missing values, which selects several complete samples with the most similar measurements to process the missing data for each incomplete sample. Absence of data are usually seen in actual-life medical situations because patients may miss some measurements for various reasons. After effectively filling out the insufficient data set, six machine learning algorithms (logistic regression, random forest, support vector machine, k-nearest neighbor, and naive Bayes classifier and feed forward neural network) were used to establish models. Among these machine learning models, random forest achieved the best performance with 99.75% diagnosis accuracy. By analyzing the miscalculation generated by the established models, we proposed an combined model that combines logistic regression and random forest by using perceptron, which could attain an average accuracy of 99.83% after ten times of simulation. The goal of this project is to develop an appropriate machine learning tool which can predict kidney disease by using some features. The algorithm that can be used here are KNN, Logistic Regression and Random Forest.

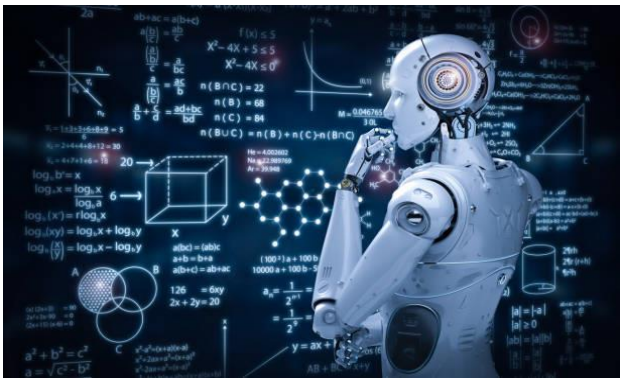
Keywords- Random forest, integrated model, logistic regression, K-Nearest Neighbor, Machine learning.

I. INTRODUCTION

Machine learning and Feature extractions are playing a major in internet and health domain.Chronic kidney disease (CKD) is a global health problem with high morbidity and mortality rate, and it cover other infection. Since there are no obvious manifestation during the early stages of CKD, patients often fail to notice the disease. Early detection of CKD enables patients to receive timely care to ameliorate the progression of this disease.

Machine learning models can effectively aid clinicians achieve this objective dueto their fast and accurate recognition performance. In this study, we propose a machine learning methodology for determining CKD. The CKD data set was acquired from the University of California Irvine (UCI) machine learning repository, which has a large number of absence of data. KNN imputation was used to fill in the absence of data, which selects several complete samples with the most similar measurements to process the absence of data for each incomplete sample.

Absence of data are usually seen in real-life medical situations because patients may miss some measurements for various reasons.



II. LITERATURE SURVEY

1) The Work done by Pankaj Chittora, Sandeep Chaurasiai, Prasun Chakrabarti, Gaurav Kumawati, Tulika Chakrabarti

”Prediction of Chronic Kidney Disease - A Machine Learning Perspective”

Chronic Kidney Disease is one of the most critical illness nowadays and proper diagnosis is required as soon as possible. Machine learning technique has become reliable for medical treatment. With the help of a machine learning classifier algorithms, the doctor can detect the disease on time. For this perspective, Chronic Kidney Disease prediction has been discussed in this article. Chronic Kidney Disease dataset has been taken from the UCI repository. Seven classifier algorithms have been applied in this research such as artificial neural network, C5.0, Chi-square Automatic interaction detector, logistic regression, linear support vector machine with penalty L1 & with penalty L2 and random tree. The important feature selection technique was also applied to the dataset. For each classifier, the results have been computed based on (i) full features, (ii) correlation-based feature selection, (iii) Wrapper method feature selection, (iv) Least absolute shrinkage and selection operator regression, (v) synthetic minority over-sampling technique with least absolute shrinkage and selection operator regression selected features, (vi) synthetic minority over-sampling technique with full features. From the results, it is marked that LSVM with penalty L2 is giving the highest accuracy of 98.86% in synthetic minority over-sampling technique with full features. Along with accuracy, precision, recall, F-measure, area under the curve and GINI coefficient have been computed and compared results of various algorithms have been shown in the graph. Least absolute shrinkage and selection operator regression selected features with synthetic minority over-sampling technique gave the best after synthetic minority over-sampling technique with full features. In the synthetic minority over-sampling technique with least absolute shrinkage and selection operator selected features, again linear support vector machine gave the highest accuracy of 98.46%. Along with machine learning models one deep neural network has been applied on the same dataset and it has been noted that deep neural network achieved the highest accuracy of 99.6%.

[2]The Work done by:G Nandhini ; J Aravinth” Chronic kidney disease prediction using machine learning techniques”

Early diagnosis and characterization are the important components in determining the treatment of chronic kidney disease (CKD). CKD is an ailment which tends to damage the kidney and affect their effective functioning of excreting waste and balancing body fluids. Some of the complications included are hypertension, anemia (low blood count), mineral bone disorder, poor nutritional health, acid base abnormalities, and neurological complications. Early and error-free detection of CKD can be helpful in averting further deterioration of patient's health. These chronic diseases are prognosticated using various types of data mining classification approaches and machine learning (ML) algorithms. This Prediction is performed using Random Forest (RF) Classifier, Logistic Regression (LR) and K-Nearest Neighbor (K-NN) algorithm and Support Vector Machine (SVM). The data used is collected from the UCI Repository with 400 data sets with 25 attributes. This data has been fed into Classification algorithms. The experimental results show that K-NN, LR, SVM hands out an accuracy of 94%, 98% and 93.75% respectively. The RF classifier gives out a maximum accuracy of 100%

[3]The Work done byImesh Udara Ekanayake; Damayanthi Herath” Chronic Kidney Disease Prediction Using Machine Learning Methods ”

Chronic Kidney Disease (CKD) or chronic renal disease has become a major issue with a steady growth rate. A person can only survive without kidneys for an average time of 18 days, which makes a huge demand for a kidney transplant and Dialysis. It is important to have effective methods for early prediction of CKD. Machine learning methods are effective in CKD prediction. This work proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attributes selection. Out of the 11 machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes. The research also considers the practical aspects of data collection and highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction.

III. EXISTING SYSTEM

- In this system different machine learning algorithms are used individually accuracy is used to compare the models
- In this case we can use machine learning algorithms KNN, Random Forest or Logistic regression

3.1 DISADVANTAGE

- The accuracy of detecting the CKD in early stage is less.
- We can’t risk with the human life.So we are to propose another method to solve these problems.
- By individually using the machine learning algorithm provides less accuracy

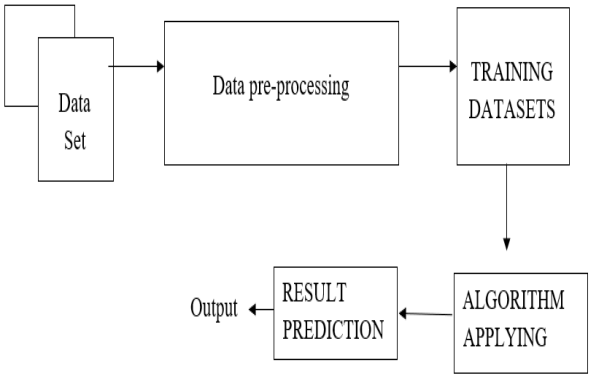
IV. PROPOSED SYSTEM

- In the proposed system k-fold cross validation is used to compare the performance of different Machine learning tools.
- In this case we used test data accuracy to tell which model is good.

4.1 ADVANTAGES

- In the proposed model pre trained VGG16 is used as base model for transfer learning.
- As transfer learning is used so no of training parameters are reduced whichreduces the time complexity and improves the performance.
- Our proposed system will accurately discover the affected area from the original area.
- This system will efficiently mark the damaged area from original image.

V. SYSTEM ARCHITECTURE



System architecture diagram of CKD

A system architecture is the visual representation conceptual model that defines the behaviour, structure and more views of a system. An architecture description is a formal explanation and representation of a architecture, organized in a way that supports reasoning about the behaviors and structure of the system

VI. ALGORITHM

6.1 RANDOM FOREST

Random forest algorithm constructs numerous decision trees to perform as an ensemble of classification and regression process. A number of decision trees are constructed using a random subsets of the training data sets. A large collection of decision trees provide higher accuracy of results. The runtime of the algorithm is comparatively quick and also accommodates lack of data. Random forest randomizes the algorithm and not the training data set. The decision class is the mode of classes generated by decision trees.

6.2 LOGISTIC REGRESSION

- Logistic regression is a supervised learning classification algorithm used to forecast the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there will be only two possible classes.

6.3 K-Nearest Neighbor

In pattern recognition, the K-Nearest Neighbor algorithm (KNN) is a non-parametric procedure used for classification and regression. In both cases, the inputs consists of the K closest training examples in the feature space. K-NN is a type of instance-based learning. In K-NN Classification, the output is a class Membership. Classification is done by a majority voting of neighbours. If K = 1, then the class is said to be single nearest neighbor. In a common weighting scheme, individual neighbour is alloted to a weight of 1/d if d is the distance to the neighbour. The shortest distance between any two neighbours is always a direct line and the distance is known as Euclidean distance [7]. The restriction of the K-NN algorithm is it's sensitive to the local configuration of the data. The process of transforming the input data to a set of features is known as the Feature extraction. In Feature space, extraction is taken place on raw data before applying KNN algorithm.

VII. LIST OF MODULES

7.1 Dataset preparation and preprocessing:

Data is the base for any machine learning project. The second stage of project implementation is complicated and involves data collection, selection, preprocessing, and transformation. Each of these phases can be split into many steps.

7.2 Image Preprocessing:

Image processing is split into analogue image processing and digital image processing.

Digital image processing is the use of a computer algorithm to execute image processing on digital images. As a subfield of digital signal processing, digital image processing has many merits over analogue image processing. It allows a

much wider range of algorithms to be applied to the input data — the objective of digital image processing is to improve the image data (features) by suppressing unwanted distortions and/or enhancement of some important image features so that our AI-Computer Vision models can gain from this improved data to work on.

7.3Data Augmentation:

Amongst the popular deep learning applications, computer vision tasks such as image classification, object detection, and segmentation have been highly victorious. Data augmentation can be effectively used to teach the DL models in such applications. Some of the simple modifications applied to the image are; geometric transformations such as Flipping, Rotation, Translation, Cropping, Scaling, and color space transformations such as color casting, Varying brightness, and noise injection. Figure 1. Shows the real image and the images after applying some of these transformations. The python code used for applying the transformations are shown in theappendix-1.

7.4 Data splitting:-

A dataset used for machine learning should be split into three subsets — training, test, and validation sets.

7.4.1Training set

A data scientist uses a training set to coach a model and define its optimal parameters — parameters it has to learn from the data.

7.4.2Test set:

A test set is essential for an evaluation of the trained model and its potential for generalization. The latter means a model's capacity to identify patterns in new unseen data after having been coached over a training data. It's crucial to use different subsets for teaching and testing to avoid model over fitting, which is the inability for generalization we mentioned above.

7.5 Modeling Evaluation

During this phase, a *data scientist* trains many models to define which one of them provides the most accurate forecast. It's time to train the model with this finite number of images. fast.ai offers many architectures to use which makes it very simple to utilize transfer learning.

VIII. DATASET

The CKD data set utilized in this study was acquired from the UCI machine learning repository. The data set contains 400 samples. In this CKD data set, each sample has 24 predictive variables or features (11 numerical variables and 13 categorical (nominal)variables) and a categorical response variable (class). Each class has two values, namely, CKD (sample with CKD) and not CKD (sample without CKD). In the 400 samples, 250 samples belong to the category of CKD, whereas 150 samples belong to the category of not CKD. It is worth mentioning that there is a huge number of missing values in the data

Details of each variable in the original CKD data set.

Variables	Explain	Class	Scale	Missing Rate
age	Age	Numerical	age in years	2.25%
bp	Blood Pressure	Numerical	in mm/Hg	3%
sg	Specific Gravity	Nominal	(1.005,1.010,1.015,1.020,1.025)	11.75%
al	Albumin	Nominal	(0,1,2,3,4,5)	11.5%
su	Sugar	Nominal	(0,1,2,3,4,5)	12.25%
rbc	Red Blood Cells	Nominal	(normal,abnormal)	38%
pc	Pus Cell	Nominal	(normal,abnormal)	16.25%
pcc	Pus Cell clumps	Nominal	(present,notpresent)	1%
ba	Bacteria	Nominal	(present,notpresent)	1%
bgr	Blood Glucose Random	Numerical	in mgs/dl	11%
bu	Blood Urea	Numerical	in mgs/dl	4.75%
sc	Serum Creatinine	Numerical	in mgs/dl	4.25%
sod	Sodium	Numerical	in mEq/L	21.75%
pot	Potassium	Numerical	in mEq/L	22%
hemo	Hemoglobin	Numerical	in gms	13%
pcv	Packed Cell Volume	Numerical	-	17.75%
wbcc	White Blood Cell Count	Numerical	in cells/cumm	26.5%
rbcc	Red Blood Cell Count	Numerical	in millions/cmm	32.75%
htn	Hypertension	Nominal	(yes,no)	0.5%
dm	Diabetes Mellitus	Nominal	(yes,no)	0.5%
cad	Coronary Artery Disease	Nominal	(yes,no)	0.5%
appet	appet	Nominal	(good,poor)	0.25%
pe	Pedal Edema	Nominal	(yes,no)	0.25%
ane	Anemia	Nominal	(yes,no)	0.25%
class	Class	Nominal	(ckd,notckd)	0%

IX. JOURNAL OF HEALTHCARE ENGINEERING

Journal of Healthcare Engineering 3

Table 1: The stages of development of CKD.

Stage	Description	Glomerular filtration rate (GFR) (mL/min/ 1.73 m^2)	Treatment stage
1	Kidney function is normal	≥90	Observation, blood pressure control
2	Kidney damage is mild	60–89	Observation, blood pressure control and risk factors
3	Kidney damage is moderate	30–59	Observation, blood pressure control and risk factors
4	Kidney damage is severe	15–29	Planning for end-stage renal failure
5	Established kidney failure	≤ 15	Treatment choices

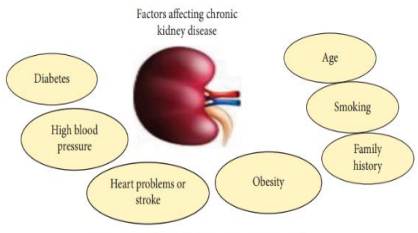
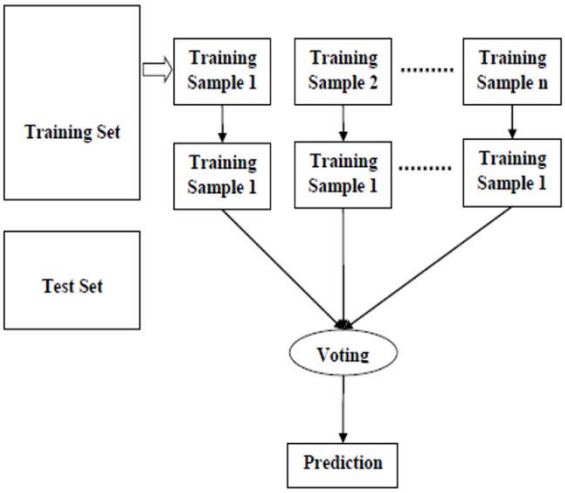


Figure 1: Factors affecting chronic kidney disease.

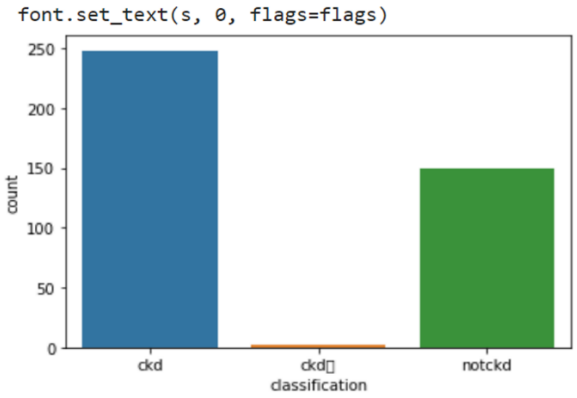
X. ARCHITECTURAL DIAGRAM

An architecture diagram is a **graphical representation of a set of idea**, that are part of an architecture, including their principles, elements and components.

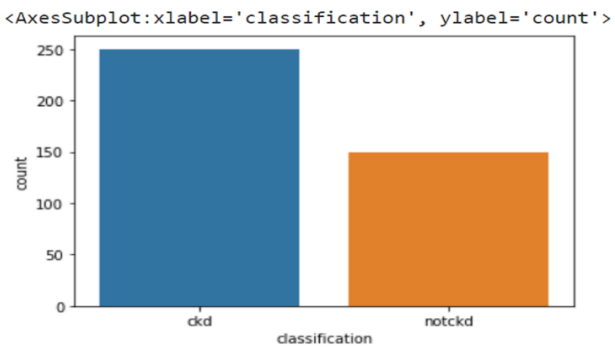


Architectural diagram for CKD

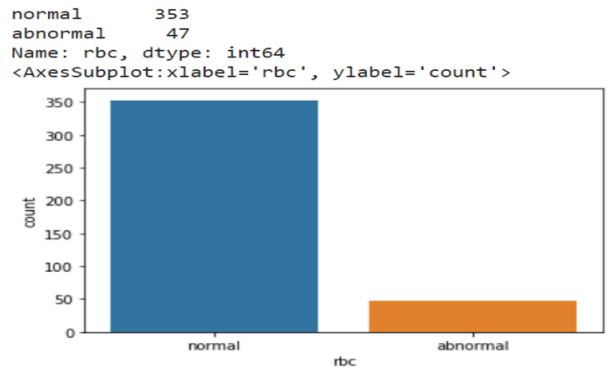
XI. EXPERIMENT ANALYSIS



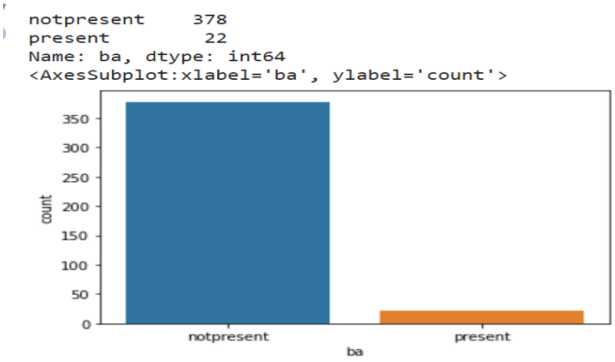
Class label count plot before replacing



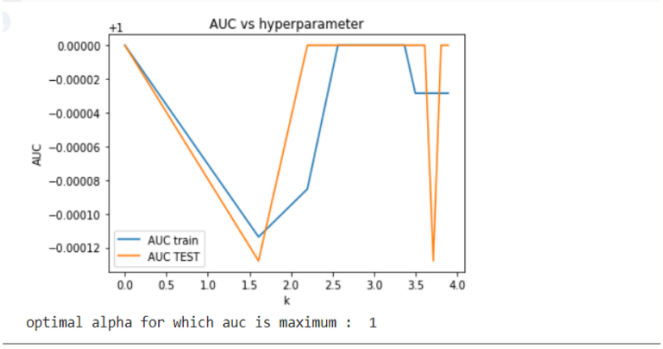
Class label count plot after replacing

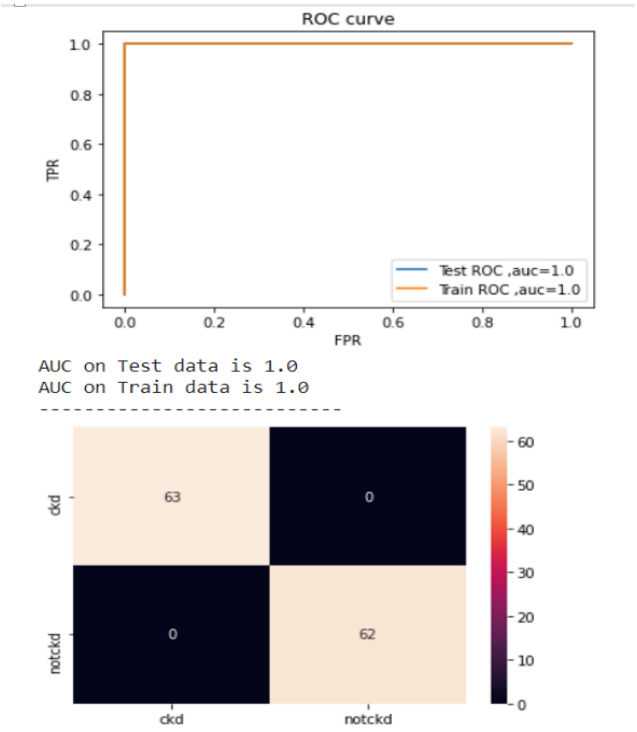


RBC feature count plot



BA feature count plot





XII. CONCLUSION

This review explained ML approaches for the detection of chronic diseases. Moreover, numerous visualization techniques/mappings were summarized to recognize the manifestation of diseases. Although much significant development was observed during the last three to four years, there are still some research gaps which are narrated below

- In most of the researches (as described in the previous sections), PATIENT dataset was used to evaluate the accuracy and performance of the respective DL models/architectures. Although this dataset has a loads of images of several plant species with their diseases, it has a simple/plain background. However, for a practical scenario, the actual environment should be considered.
- Hyperspectral/multispectral imaging is an emerging technology and has been used in numerous areas of research. Therefore, it should be used with the efficient DL architectures to discover the plants' diseases even before their manifestations are clearly apparent.
- A more well organized way of visualizing the spots of disease in plants should be launched as it will save costs by avoiding the needless application
- The seriousness of plant diseases changes with the passage of time, therefore, DL models should be improvedto enable them to discover and classify diseases during their total cycle of occurrence.
- DL model/architecture should be well organized for many illumination conditions, so the datasets should not only indicate the real environment but also contain images taken in distinct field scenarios.
- A comprehensive study is required to acknowledge the factors affecting the detection of plant diseases, like the classes and size of datasets, learning rate, illumination, and the like.

XIII. ACKNOWLEDGEMENT

The authors would like to thank **Ms.M.ANITHA** for her suggestions and excellent guidance throughout the project period.

REFERENCES

[1] Al-Bashish D, M. Braik and S. Bani-Ahmad, 2011. Detection and classification of leaf diseases using K-means-based segmentation and neural networks based classification. Inform. Technol. J., 10: 267-275. DOI:10.3923/itj.2011.267.275, January 2011.

[2] Armand M.Makowski "Feature Extraction of diseased leaf images", Fellow, IEEE Transactions on information theory Vol.59, no.3 March-2013

[3] H.Al-Hiary, S. Bani-Ahmad, M.Reyalat, M.Braik and Z.Al Rahamneh, Fast and Accurate Detection and Classification of Plant Diseases, International Journal of Computer Applications (0975-8887), Volume 17-No.1.March 2011.

[4] DaeGwan Kim, Thomas F. Burks, Jianwei Qin, Duke M.Bulanon, Classification of grapefruit peel diseases using color texture feature analysis, International Journal on Agriculture and Biological Engineering, Vol:2, No:3,September 2009. Open access at http://www.ijabe.org

[5] Jasmeet Kaur, Dr.Raman Chadha, Shvani Thakur, Er.Ramanpreet Kaur. A Review Paper on Plant Disease Detection using Image Processing and Neural Network Approach. International Journal of Engineering Sciences & Research Technology. April 2016. ISSN: 2277-9655.

[6] Diptesh Majumdar, Dipak Kumar Kole, Aruna Chakraborty, Dwijesh Dutta Majumder. REVIEW: DETECTION & DIAGNOSIS OF PLANT LEAF DISEASE USING INTEGRATED IMAGE PROCESSING APPROACH. International Journal of Computer Engineering and Applications. June 2014. Volume VI; Issue- III.

[7] S. S. Sannakki, V. S. Rajpurohit. An Approach for Detection and Classification of Leaf Spot Diseases Affecting Pomegranate Crop. International Journal of Advance Foundation and Research in Computer. January 2015, Volume 2, Special Issue (NCRTIT 2015), ISSN 23484853.

[8] DavoudAshourloo, Hossein Aghighi, Ali Akbar Matkan, Mohammad Reza Mobasheri, and Amir Moeini Rad," An Investigation Into Machine Learning Regression Techniques for the Leaf Rust Disease Detection Using Hyperspectral Measurement" 2016 IEEE.

[9] Mr. MelikeSardogan “Plant Leaf Disease Detection and Classification based on CNN with LVQ Algorithm” 2018 3rd International Conference on Computer Science and Engineering (UBMK) 2018 IEEE.

[10] K. P. Ferentinos, “Deep learning models for plant disease detection and diagnosis”, Computers and Electronics in Agriculture, vol. 145, pp. 311-318, 2018