Machine Learning Approach for Indian Crop Yield Prediction

Dr.S.Erana Veerappa Dinesh[1], M. Anusuya[2], M. Vanneeswari[3]
[1]*Asst.Professor, P.S.R.R College of Engineering, Sivakasi*
[2,3]*UG Student, P.S.R.R College of Engineering*

Abstract- In today's world, technology plays a critical part in overcoming challenges and achieving better and maximal results in a variety of fields. The agricultural sector in India has a significant impact on the economy. Agriculture employs half of the country's population. The agriculture business is heavily influenced by its surroundings' natural circumstances, and as a result, it faces a number of difficulties in terms of actual farming techniques. Agriculture techniques in the country are generally ad hoc, and technological advancement is modest. In this industry, effective technology can be employed to boost yield while minimizing the challenges. Farmers typically sow crops based on their market value and potential yield.

***Key Words*: Crop Recommendation, CNN, Data Analysis, Decision tree, Logistic regression, Machine Learning, Random Forest.**

I.INTRODUCTION

Farming is not considered a business in India, yet it has a significant impact on the social lives of those who are involved with it. The many seasons and farming activities are commemorated with a variety of festivals and social gatherings. As a result, agriculture affects a vast portion of the population, either directly or indirectly. In India, however, the status of farmers is not good. Despite employing half of the population, agriculture accounts for barely 20% of India's GDP. Hence, it is in dire need of improvement to make a good and profitable yield and also, a practical need without harming nature. That's where technology comes in and can have major effects on the agricultural sector. Our project aims to tackle the difficulties faced by the farmers and aims to provide a correct crop for the farmers to grow and avoid undesirable results by providing effective solutions using machine learning techniques.

II.LITERATURE SURVEY

Before beginning our endeavor, we consulted the following research publications. When referring to We discovered distinct discoveries in each of these studies, which are mentioned below. Each of the following entries describes the paper's title, the algorithms utilized, and a broad conclusion derived from the study.In [1] Using a machine learning algorithm, predict crop yield. The goal of this research is to use the Random Forest method to predict crop yield based on existing data. The models were built using real data from Tamil Nadu, and the models were tested using samples. The Random Forest Algorithm can be used to accurately predict crop yields.

In [2] Random forests for global and regional crop yield prediction. Because of its high accuracy and precision, ease of use, and adaptability, our generated outputs suggest that RF is an excellent and flexible machine-learning method for crop production projections at regional and global scales. of use, and utility in data analysis. Random Forest is the most efficient strategy and it outperforms multiple linear regression (MLR).In [3]. Crop production forecast using an ensemble machine learning model. The suggested ensemble models AdaNaive and AdaSVM are utilized to project crop production over time in this paper. Implementation done using AdaSVM and AdaNaive. The AdaBoost method improves the efficiency of the SVM and Naive Bayes algorithms.In [4]. Crop yield forecasts based on climate parameters using machine learning. The paper provided in International Conference on Computer Communication and Informatics (ICCCI).Crop Advisor, a user-friendly online portal for estimating the impact of climatic conditions on crop yields, has been developed as part of the current study.

The C4.5 method is used to generate the most influential climatic parameter on agricultural yields of selected crops in Madhya Pradesh's selected districts. The paper is implemented using Decision Tree.In[5]. Prediction On Crop Cultivation. Soil analysis and interpretation of soil test findings are currently done on paper. This has contributed to inaccurate interpretation of soil test results, which has resulted in poor crop recommendations, soil amendments, and fertilizer recommendations to

farmers, resulting in poor crop yields, micronutrient deficiencies in soil, and excessive or insufficient fertilizer application. Formulae to Match Crops with Soil, Fertilizer Recommendation.In [6]. Analysis of Crop Yield Prediction by making Use Data Mining Methods. In this paper the main aim is to create a user friendly interface for farmers, which gives the analysis of rice production based on the available data. Various data mining approaches were utilized to estimate crop yield in order to maximize crop productivity. For example, the K-Means algorithm can be used to predict the pollution factor in the atmosphere.

In [7]. Machine Learning Techniques in the Production of Agricultural Crops. From GPS-based color, images are provided as an intensified indistinct cluster analysis for classifying plants, soil, and residue regions of interest. The document provides a number of parameters that can be used to improve crop output and boost the yield ratio during cultivation.In [8] We give a comprehensive assessment of studies on the use of machine learning in agricultural production systems in this publication. Machine learning (ML) has emerged together with big data technologies, techniques, methods, and high-performance computing to generate new opportunities to unravel, quantify, and analyze data-intensive processes in agricultural operational sectors. By using Support Vector Machines (SVP) the Paper is Implemented. In [9]. Precision Agriculture on an Aerial Platform: A Study to Determine Yield for Crop Insurance Precision agriculture (PA) is the application of geospatial methodologies and remote sensors to identify variances in the field and deal with them utilizing various ways.

The causes of variability of crop growth in an agricultural field might be due to crop stress, irrigation practices, the incidence of pests and disease, etc. Ensemble Learning is used to implement the paper (EL). In [10]. Random Forests for Crop Yield Predictions at the Global and Regional Level Because of its great accuracy, the obtained outputs suggest that RF is an effective and unique machine-learning method for agricultural production projections at regional and global scales. Ensemble Learning is used to implement the paper (EL).In[11]. Data Mining Techniques for Crop Suitability and Fertilizer Recommendation. This paper Random Forest, K- means clustering algorithm is used. The Accuracy of random forest is found to be higher than the ID3(Iterative Dichotomiser 3) algorithm for crop prediction and K- means for fertilizer recommendation.

In[12]. Machine Learning-based Random Forest Algorithm for Soil Fertility Prediction and Grading. This paper is Random Forest, Gaussian Naïve Bayes, Support VectorMachine, and Linear Regression. In the case of Crop Prediction, Random Forest proves to be a better classifier as compared to Gaussian Naïve Bayes and Support Vector Machine while linear regression works efficiently for grading soil.In[13]. Soil Classification using Machine Learning Methods and Crop Suggestion Based on soil series. This paper uses Weighted K-NN, SVM, and Bagged Tree. SVM has given the highest accuracy in soil classification as compared to K-NN and Bagged tree algorithms.In[14]. Machine Learning Algorithms for Intelligent Crop Recommendation Decision Tree, Random Forest, K- NN is used. Accuracy rates were Decision Tree (90.20), K-NN (89.78), Random Forest (90.43)

III.METHODOLOGY

To accomplish the desired goals, the following procedures were taken during the project implementation.

Cleaning and Preprocessing of Data

One of the first tasks is to double-check that the dataset we're working with is correct. There should be no missing values in the dataset, and if there are, they should be filled in with the right values. It's also a good idea to look at the data to see if the features have a normal distribution. Outliers should be eliminated.The skew value of the features should be verified, and if skewness exists, the features should be normalized with transformations. We utilized a dataset with skewness in some of the features. We applied quantile transformation on the data to standardize it.It is also known as scrubbing. Filling in missing numbers, smoothing or deleting noisy data and outliers, and resolving inconsistencies are all part of this task.

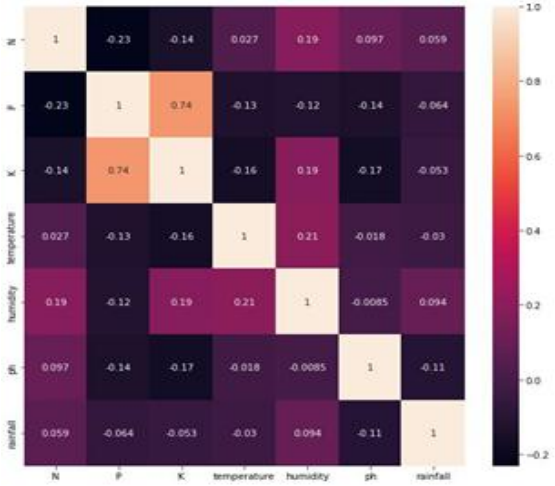
Data Analysis and Visualization

Following the data cleaning and preparation, we do data analysis and visualizations on our dataset. We strive to dig deeper into our data to see if there are any trends or patterns in the data. In order to fully comprehend our dataset, we generated numerous visualizations. We used bar charts, scatter plots,

box plots, and other visualization tools to look at the data and see if there are any trends or patterns that we can utilize to help us implement our project.

Feature Selection

Only those features that are required to decide the sort of crop to produce must be chosen. We've constructed a correlation matrix to show the linear link between each attribute and the others. If features are highly correlated, they should be dropped; however, the features in the matrix below are not highly connected with one another, therefore it makes sense not to eliminate any of them, and we will use all of them to predict the sort of crop to produce.



Model Building

Building the machine learning model is the next phase. While building the machinelearning model, first we need to split our dataset into 2 parts i.e.: training data and test data. We have split the data in the ratio of 70-30. Taking the training data, we apply our machine learning algorithms to the features of the dataset. We have used 4 machine learning algorithms on our training dataset and the algorithms that give us the highest accuracy will be selected on the test dataset.

IV.DATASET

This topic's dataset was obtained via Kaggle. It's a crop suggestion dataset that tells us about different sorts of crops and the characteristics that determine which one is best for growing. we access the dataset through this site https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

Features of the Dataset

N: ratio of Nitrogen content in the soil
P: the ratio of Phosphorous content in the soil
K: the ratio of Potassium content in soil
Temperature: temperature in degree Celsius
Humidity: relative humidity in %
pH: pH value of the soil
Rainfall: rainfall in mm

Domain	Area Code	Area	Element Code	Element	Item Code	Item	Year Code	Year	Unit	Value
Crops	2	Afghanistan	5419	Yield	56	Maize	1961	1961	hg/ha	14000
Crops	2	Afghanistan	5419	Yield	56	Maize	1962	1962	hg/ha	14000
Crops	2	Afghanistan	5419	Yield	56	Maize	1963	1963	hg/ha	14260
Crops	2	Afghanistan	5419	Yield	56	Maize	1964	1964	hg/ha	14257
Crops	2	Afghanistan	5419	Yield	56	Maize	1965	1965	hg/ha	14400

Fig:- Dataset collection

V.MACHINE LEARNING ALGORITHM USED

Random Forest

Random Forest is a supervised ensemble machine learning technique that can be used to classify and predict data. It contains a number of decision trees, and the result is calculated by taking the average among them. Random forest is effective in lowering the effect of over fitting and so producing a more accurate output, as decision trees are prone to over fitting.

1. Choose "k" features at random from a total of "m" features.
2. Calculate the node "d" using the optimal split point among the "k" characteristics.
3. Using the best split, split the node into daughter nodes.
4. Repeat steps 1–3 until the "I" number of nodes is attained repeating steps 1 through 4 "n" times to create "n" trees.

Decision Tree

The Decision Tree method is one of the most widely used machine learning algorithms. It is mostly used to solve classification problems, but it may also be used to solve regression problems. Its operation is based on a basic technique in which a yes/no question is posed and the tree is divided into smaller nodes based on the answer. The nodes can be separated using either Gini impurity (which calculates the measure of impurity) or information gain (calculates the change in the entropy). Over

fitting is a risk with Decision Trees, which can result in reduced accuracy. A random forest algorithm can be used to solve this problem.

1. Put the dataset's best attribute at the top of the tree.
2. Split the training set into subsets. Subsets should be created so that each subset has data with the same attribute value.
3. Repeat steps 1 and steps 2on each subset until you find leaf nodes in all the branches of the tree.

Logistic Regression

It's a tool for resolving classification issues. It employs a sigmoid function to determine the likelihood of observation, after which the observation is assigned to the appropriate class. If an observation's probability is 0 or 1, a threshold value is chosen, and classes with probabilities over the threshold are given the value 1, while classes with probabilities below the threshold are given the value 0.

1. Training data as input
2. In order for me to be able to k
3. For each occurrence of training data di.
4. Set the regression's goal value to ziyj-p(1/djp(1/dj).(1-p(1/dj)) .
5. Set p(1/dj) as the weight of instance dj. (1-p(1/dj)).
6. Apply af(j) to the data using the class value (zj) and the weights (wj).
7. If p(1/dj)>0.5, assign (class label: 1), else (class label: 2).

Pseudocode for logistic regression

LR cell images and related HR cell images Xi are used as input.

Model parameter =W1,W2,W3,B1,B2,B3 (output)

1. θ are initialized by selecting a number at random from the Gaussian Distribution
2. Step 2: n is the number of training photos for i=0 to n/.

Pseudocode for CNN

EFFICIENCY

Step 1:The user logs in to the system in the first step.

Step 2: If the user's login is successful, the user's location is tracked.

Step 3: The system now offers the following two options to the user:

- Prediction Module: The user can select to learn about a specific crop's prediction or to learn about a list of crops and their accompanying productions.
- Fertilizer Module: This module allows the user to determine when it is appropriate to apply fertilizer. Rainfall forecasts for the next 15 days are used to achieve this. If it is likely to rain, the system returns yes; otherwise, it returns no.

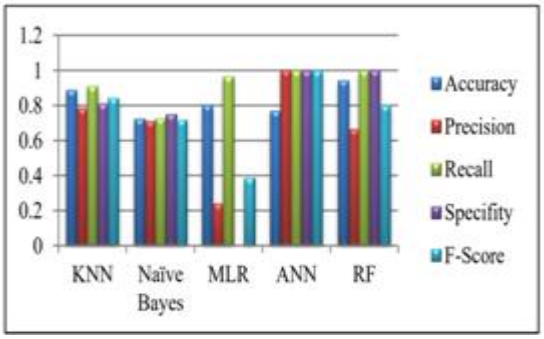


Fig:- Efficiency graph for crop yield prediction

VI.FLOW CHART

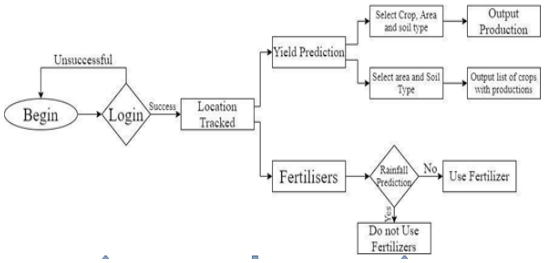


Fig:- Flow chart of crop yield prediction system

VII.IMPLEMENTATION

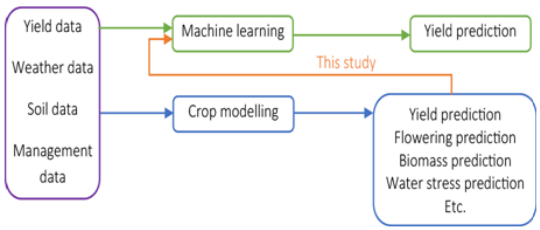


Fig:- Diagram of crop yield prediction

VIII.RESULT

We used our machine learning algorithms to the dataset's features after conducting data cleaning and visualization. Decision Tree, Random Forest, Logistic Regression, and CNN are the four algorithms we used. The N(Nitrogen), P (Phosphorous), and K (Potassium) values of the soil, as well as temperature, humidity, rainfall, and pH value, were chosen as features for the dataset.

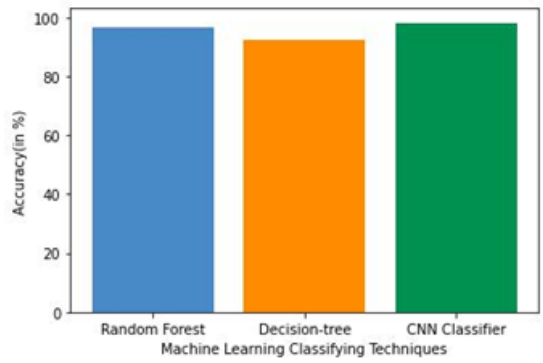


Fig:- Graphical Comparison of accuracies of all the machine learning models

Model	Accuracy Score
Random Forest	96.67%
Decision Tree	92.43%
CNN	98.18%

Table:- Accuracy values of all the machine learning model

IX.CONCLUSION

Our farmers are underperforming in their farming approaches because they do not use technology in their farming practices. As a result, we created this project to urge farmers to embrace modern technologies rather than traditional techniques. Growing a crop necessitates a great deal of knowledge and insight into a variety of factors, such as the contents of the soil, the temperature of the location, the pH of the soil, and so on. As a result, it makes sense for the farmer to adopt newer technologies to make his life easier. We anticipated a 98 percent accuracy in predicting the proper crop to grow using machine learning algorithms, and if the farmer embraces this technology, it will not only make his life easier, but it will also assist him in making environmentally friendly decisions.

X.FUTURE SCOPE

We have currently confined our scope to a crop recommendation system, but this can be expanded in the future to include other areas such as fertilizer recommendations, where a person can learn about the best fertilizer for his crop. Another expanded application of this may be the classification of plant diseases using CNN and then providing treatment for that condition. Agriculture is a relatively untouched field, thus the potential for such a project is enormous. While we have created a functional website for our approach, it might be converted into an app in the future to make it more accessible to farmers. The two People will be able to grasp the website and app if they are created in regional languages. People will be more comfortable using the website and app if they are designed in their native language.

REFERENCE

[1] P. Priya, U. Muthaiah M. Balamurugan. Predicting yield of the crop using machine learning algorithm. International Journal of Engineering Science Research Technology.

[2] J. Jeong, J. Resop, N. Mueller and team. Random forests for global and regional crop yield prediction. PLoS ONE Journal

[3] Narayanan Balkrishnan and Dr. Govindarajan Muthukumarasamy. Crop production Ensemble Machine Learning model for prediction. International Journal of Computer Science and Software Engineering (IJCSSE).

[4] S. Veenadhari, Dr. Bharat Misra, Dr. CD Singh. Machine learning approach for forecasting crop yield based on climatic parameters. International Conference on Computer Communication and Informatics (ICCCI).

[5] Shweta K Shahane, Prajakta V Tawale(2016). “Prediction On Crop Cultivation”. IInternational Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 5, Issue 10.

[6] D Ramesh, B Vishnu Vardhan. Analysis Of Crop Yield Prediction Using Data Mining Techniques. IJRET: International Journal of Research in Engineering and Technology.

[7] Subhadra Mishra, Debahuti Mishra, Gour Hari Santra(2016) “Applications of Machine Learning Techniques in Agricultural Crop Production”. Indian Journal of Science and Technology, Vol 9(38), DOI:10. 17485/ijst/2016/v9i38/95032.

[8] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson ID, Dionysis Bochtis. Machine Learning in Agriculture. Lincoln Institute for Agri-food Technology (LIAT), University of Lincoln, Brayford Way, Brayford Pool, Lincoln LN6 7TS, UK, spearson@lincoln.ac.uk.

[9] Baisali Ghosh. A Study to Determine Yield for Crop Insurance using Precision Agriculture on an Aerial Platform. Symbiosis Institute of Geoinformatics Symbiosis International University 5th & 6th Floor, Atur Centre,

Gokhale Cross Road, Model Colony, Pune – 411016.

[10] Jig Han Jeong, Jonathan P. Resop, Nathaniel D. Mueller, David H. Fleisher, Kyungdahm Yun, Ethan E. Butler, Soo-Hyung Kim. Random Forests for Global and Regional Crop Yield Predictions. Institute on the Environment, University of Minnesota, St. Paul, MN 55108, United States of America.

[11] Ecochem Online. (2009). Soil Health and Crop yields. Last modified January 28th 2009. Retrieved on March 4th 2009 from http://ecochem. com/healthy_soil. html Food and Agricultural Organization. (2006). The state of Agricultural Commodity Markets. 37-39.

[12] Aditya Shastry, H. A Sanjay And E. Bhanushree,(2017)"Prediction of crop yield using Regression Technique", International Journal of computing r12 (2):96- 102, ISSN:1816-914] E. 14]E. Manjula, S.

[13] Djodiltachoumy, (2017)"A Model for Prediction of Crop Yield", International Journal of Computational Intelligence and Informatics, Vol. 6: No. 4.

[14] Mrs. K. R. Sri Preethaa, S. Nishanthini, D. SanthiyaK. Vani Shree,(2016)"CropYield Prediction", International Journal On Engineering Technology and Sciences – IJETS™ISSN(P): 2349-3968, ISSN (O):2349-3976 Volume III, Issue III.

[15] Jharna Majumdar, Sneha Naraseeyappa and Shilpa Ankalaki,(2017) "Analysis of agriculture data using data mining techniques: application of big data" Majumdar et al. J Big Data DOI 10. 1186/s40537-017- 0077-4

[16] D. Ramesh and B. Vardhan,(2015) "Analysis of crop yield prediction using data mining techniques", International Journal of Research in Engineering and Technology, vol. 4, no. 1, pp. 47-473.

[17] Yethiraj N G,(2012)"Applying data mining techniques in the field of Agriculture and allied sciences", Vol 01, Issue 02.

[18] Zelu Zia (2009). "An Expert System Based on Spatial Data Mining used Decision Tree for Agriculture Land Grading". Second International Conference on Intelligent Computation Technology and Automation.

[19] Archana Chougule, Vijay Kumar Jha and Debajyoti Mukhopadhyay(2019) "Crop Suitability and Fertilizers Recommendation using Data Mining Techniques" Springer Nature Singapore Pte Ltd.

[20] Keerthan Kumar T G, Shubha C, Sushma S A(2019)" Random Forest Algorithm for Soil Fertility Prediction and Grading using Machine Learning "International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-1.

[21] Sk Al Zaminur Rahman, S.M. Mohidul Islam, Kaushik Chandra Mitra" Soil Classification using Machine Learning Methods and Crop Suggestion Based on Soil

[22] Zeel Doshi, Subhash Nadkarni, Rashi Agrawal, Prof. Neepa Shah (2018)"AgroConsultant: Intelligent Crop Recommendation System Using Machine LearningAlgorithms " Fourth International Conference onComputing Communication Control and Automation (ICCUBEA).

[23] B. Balázs, E. Kelemen,et.al., (2021)''Integrated policy analysis to identify transformation paths to more sustainable legume-based food and feed value-chains in Europe,'' Agroecol. Su; stain. Food Syst., vol. 45, no. 6, doi10.1080/21683565.2021. 18 84 165.

[24] S. S. Sana, ''Price competition between green and non green products under corporate social responsible firm,(2020)'' J. Retailing Consum. Services, vol. 55, Art. no. 102118, doi: 10.1016/j.jretconser.2020.102118

[25] D. Bertoni, D. Cavicchioli, F. Donzelli, G. Ferrazzi, D. Frisio, R. Pretolani, E. Ricci, and V. Ventura,(2018) ''Recent contributions of agricultural economics research in the field of sustainable development,'' Agriculture, vol. 8, no. 12, p. 200, doi: 10.3390/agriculture8120200.

[26] A. Anik, S. Rahman, and J. Sarker, (2017)''Agricultural productivity growth and the role of capital in south Asia (1980–2013),'' Sustainability, vol. 9, no. 3, p. 470, doi: 10.3390/su9030470.

[27] N. T. Liliane and M. S. Charles,(2020) ''Factors affecting yield of crops,'' in Agronomy—Climate Change and Food Security. Rijeka, Croatia: IntechOpen, doi: 10.5772/intechopen.90672.

[28] Elavarasan, D. R. Vincent, V. Sharma, A. Y. Zomaya, and K. Srinivasan,(2018) ''Forecasting yield by integrating agrarian factors and machine learning models: A

survey,'' Comput. Electron. Agricult., vol. 155, pp.257–282, doi: 10.1016/ j.compag.2018.10.024.

[29] Kannan, M. Prabhakaran S and P. Ramachandran (2011)."Rainfall forecasting using data mining technique". International Journal of Engineering and Technology Vol.2 (6), 397- 401.

[30] Shweta K Shahane, Prajakta V Tawale(2016)."Prediction On Crop Cultivation". International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 5, Issue 10.

[31] Liaw, A., Wiener, M. (2002). "Classification and regression by randomForest". R news, 2(3), 18-22.

[32] Sak, H., Senior, A., Beaufays, F. (2014). "Long short-term memory recurrent neural network architectures for large scale acoustic modeling". In Fifteenth annual conference of the international speech communication association.

[33] Mandic, D. P., Chambers, J. (2001). "Recurrent neural networks for prediction: learning algorithms, architectures and stability".

[34] Dahikar, S. S., Rode, S. V. (2014). "Agricultural crop yield prediction using artificial neural network approach". International journal of innovative research in electrical, electronics, instrumentation and control engineering, 2(1), 683-686.

[35] Mrs. N. Saranya, Ms. A. Mythili(2020),"Classification of Soil and Crop Suggestion using Machine Learning Techniques" International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org IJERTV9IS020315 (4.0 International License.) Vol. 9 Issue 02.