**IRFC – SMART FARMING: CROP RECOMMENDATION SYSTEM USING IMPROVED RANDOM FOREST CLASSIFICATION ALGORITHM**

**G. Buvaanyaa1, Dr. M. Gobi2**

*1 PhD, Research Scholar, PG & Research Department of Computer Science, Chikkanna Government Arts College, Tiruppur, Tamil Nadu, India.*

*buvaanyaa220497@yahoo.com*

*2 Associate Professor, PG & Research Department of Computer Science, Chikkanna Government Arts College, Tiruppur, Tamil Nadu, India.*

*gobimohan2786@gmail.com*

**Abstract:** A rapidly expanding field that is vital to the development of agricultural methods is crop prediction. Within the framework of machine learning theory, Random Forest (RF) was created. This article examines crop recommendation performance using the Random Forest (RF) machine-learning technique. To accurately forecast which crop would thrive in a particular area, the suggested approach uses a variety of features, including soil and climate data. Performance indicators including MAE, R2, Recall, Precision, and F1-Score were used to demonstrate the effectiveness of this RF model. It revealed that the anticipated result of the RF representations matches the expected data, proving the efficacy of the model. This proposed strategy has significant ramifications for farmers, agricultural stakeholders, and policymakers since it makes data-driven decisions and resource allocation easier, which raises crop yields.

**Keywords:** Machine Learning, Feature Selection, Classification, Random Forest, Crop prediction.

**I. INTRODUCTION**

Machine-Learning [1]-[2] is the science of enabling computer to study without explicit programming. To produce predictions or judgments, machine-learning algorithms [3] are skilled on vast volumes of information. As a significant global industry, agriculture depends on farmers producing sustainable and profitable crops. Crop production can be significantly impacted by poor crop selection, which can subordinate productivity and possibly result in losses for farmers. Farmers may find it difficult for the selected crops to flourish and reach their maximum yield potential if they neglect important aspects like soil conditions, market demand, and climate appropriateness. Inadequate climate adaptation can cause inappropriate crops to grow poorly, become more susceptible to pests and illnesses, and produce less overall. The financial burden on farmers may be exacerbated if commodities that don't meet market demand have trouble finding customers or obtaining favourable pricing [4]. To overcome these obstacles and make wise choices that will optimize crop yield and guarantee agricultural sustainability over the long run, farmers can use machine learning-based crop recommendation systems.

Inspired by Amit and Geman's prior work, Leo Breiman introduced Random Forests. Random Forests are an outgrowth of Breiman's bagging concept and were created as a rival to boosting, despite the fact that this is not immediately clear from the description in [5]. Random Forests can be used for either a categorical response variable, referred to in [6] as “classification”, or a continuous response, referred to as “regression”. Likewise, the predictor variables may be continuous or categorical [7].

Random Forests are attractive from a computational perspective because they can

* naturally handle both regression and (multiclass) classification;
* they train and predict very quickly;
* It may be used directly to high-dimensional problems;
* it has an integrated estimate of generalization error;
* it only requires one or two tuning parameters;
* it is simple to execute in parallel.

The other characteristics that Random Forests offer,

* like measures of variable importance,
* differential class weighting,
* missing value imputation,
* visualization,
* outlier detection,
* unsupervised learning,
* make them attractive from a statistical standpoint.

Figure 1 shows a construction for crop forecasting in agricultural yields. This framework includes all agricultural yield information. The data-set is pre-processed using data-cleaning techniques. The dataset can be classified using machine learning techniques once extraneous characteristics have been removed using a feature selection strategy. This aids in identifying the kind of dataset that works with machine learning algorithms [8]. For a given field, this aids in determining the kind and timing of the crop forecast.

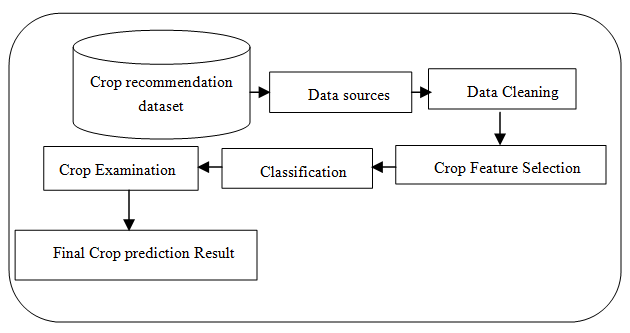


Fig.1. Diagram of the crop forecasting data-mining process

The aim of the suggested structure is towards offer an improved and more effective feature selection and classification method that is effective in managing scale and estimating density, both of which have an impact on crop forecasting outcomes. This article present the methodologies, namely data pre-processing, Enhanced Recursive Feature Selection (ERFS) and Improved Random Forest Classification (IRFC) used to determine crop forecasting result and instance effectiveness.

The major contribution of the recommended method is the follow:

* The objective of upward an effective categorization technique is to remove unrelated characteristics and review the crops by the IRFC categorization.
* The planned IRFC classifier is a classifier case to evaluate the forecast class. By means of the assist of its characteristics, the classification of random selection functions achieve 22 classes.

**II. LITERATURE REVIEW**

G. Buvaanyaa and M. Gobi [9] present a crop recommendation system using hybrid classification algorithm. Comparing this hybrid classification algorithm with other classification algorithms such as Naive Bayes, the Random Forest algorithms reveal that it has high performance metrics. According to selvi et al. [10], multispectral drone photos were taken during the growing season. Machine learning methods, such as SVR(support vector regression) and RFR(random forest regression), were used to aggregate cultivar information from a number of low indices. An original FS method called MRFE (modified recursive function elimination) was presented by Mariammal et al. [11]. Salient features are chosen and ranked using a classification algorithm in the suggested MRFE scheme. According to the survey results, the MRFE approaches choose the mainly accurate trait, but the bagging strategy correctly forecasts a suitable crop. Shafiulla Sheriff et al. have demonstrated a critical component for the continued existence and growth of the Indian financial system [12]. India was a significant manufacturer of a wide range of farming goods. Soil is an essential component of crop cultivation.

An analytical study was provided by Shahhosseini et al. [13] to show how crop combination model and machine-learning might develop corn output projections in the US Corn-Belt. They explained that in order to enhance performance forecasts, their recommended machine learning model need more hydrological input. Based on genotype and ecological data (soil & weather), Khaki & Wang [14] produced a deep neural network (NN)-based method to forecast maize hybrid yield, yields control, and yields variation. Through an RMSE of 12% of the mean performance and 5% of the standard variation for the test data-set utilize predict meteorological information, their representation was shown to be highly accurate.

Research on the forecast of potato tuber production was also carry out by Abbas et al. [15]. Based on relative soil and crop data gathered by a neighborhood survey, they assessed the yield of potato tubers (Solanum Tuberosum) by four machine-learning algorithms: SVR(support vector regression), (KNN)k-nearest neighbor, elastic network, and linear regression. A center for the study of agricultural yield in African nations was developed by Kaneko et al. [16]. Using district-level maize projections from satellite image data in six African nation —Ethiopia, Kenya, Malawi, Nigeria, Tanzania, and Zambia — they are developing a deep learning architecture. The most significant evaluation for producers and consumers is crop productivity, where the consequences of climate change are most noticeable.

**III. PROPOSED METHODOLOGY**

Every experiment utilizing the suggested methods on a sizable benchmark dataset was victorious. The study method takes into description an efficient categorization process that is developed in this part and makes use of the Improved Random Forest classification (IRFC) process, data pre-processing, and the Enhanced Recursive Feature Selection (ERFS) algorithm.

**3.1 DATA COLLECTION AND PRE-PROCESSING**

The method for gathering and evaluating data from a lot of source, like the Kaggle website, is data collecting. To acquire a system approximate dataset. The follow qualities should be there present in this dataset. Crop forecasts will take into account the following factors: Temperature, humidity, rainfall, crop data, Soil pH and NPK are the first five factors. Data pre-processing includes data-cleaning [17]-[18]. It is necessary to pre-process the data prior to training the model. Reading the gathered data is the primary step in data pre-processing, which is followed by data cleaning. During data cleaning, several unnecessary features were eliminated from the data sets; these attribute are not engaged into description for making crop predictions. Consequently, ERFS should eliminate unnecessary attributes and datasets with missing values in order to increase accuracy. It is necessary to either eliminate these missing numbers or substitute undesired NaN (Not a Number) values [19].

**3.2 ENHANCED RECURSIVE FEATURE SELECTION (ERFS) ALGORITHM**

The Recursive feature Elimination (RFE) method served as the foundation for the Enhanced Recursive Feature Selection (ERFS) strategy. The suggested ERFS recursively eliminate the character at each stage and reclassifies the enduring features by regressing Support Vector Coefficient (SVC) on the equilibrium attribute. Only weak attributes will be eliminated using ERFS. An average trait can also be useful when paired with other attributes. Therefore, the straightforward elimination of weak or redundant features can also be supported by the categorization presentation. In direct to ascertain the effects of the potentially weak feature; ERFS first assesses the categorization representation in conditions of the value of the corresponding weight vector.

Consider a data set Dt = {Dtn = | n = 1,2,3,….,p} shown in equation (1) adapted to p quantity of data R nf; ∈Dnn is the numeral of attribute feati. To define Dtnq as the qth attribute feat, consequence of the data set Dn, each data Dtn of the set Dt able to be symbolize by a vector.

Suppose a feat set represent all features in equation (2),

Algorithm 1's feature selection procedure eliminates the weakest properties after the necessary number of features has been reached. By eliminating all potential dependencies and colinearity from the model, an ERFS seeks to recursively recall a limited amount of characteristics every loop. According to the importance dimensions of coefficients or attributes of the model, the features are classified in order of importance [20]. Figure 2 below shows the selected features in the data set and Figure 3 showing the correlation matrix of the selected features. The below equation is used to calculate mc, or the totality numeral of hyper planes, in the case where the data set contain further than two classes.

Equation (3) refers the decision function.

**Algorithm 1**

**Enhanced Recursive Feature Selection (ERFS) Algorithm**

**Input:** Training Set, Set of features , Number of features to select.

**Output:** Selected Set of features providing highest accuracy

**Begin the process**

**For** each n in N, do

Perform ERFS and select n features → F

Train with RF using F

Compute accuracy of model with out of bag predictions

**End** for

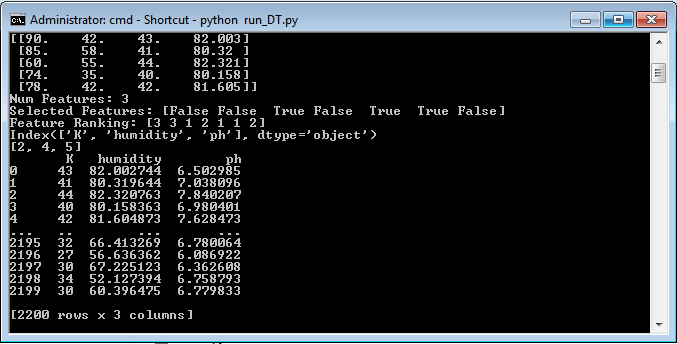


Fig.2. Selected Features

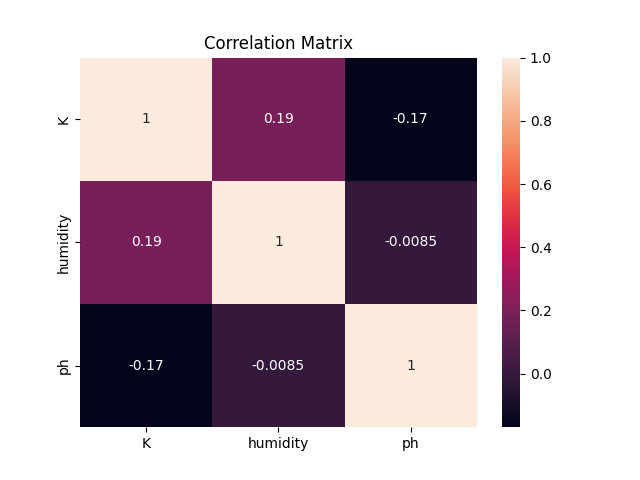


Fig.3. Selected Correlation Feature matrix

**3.3 IMPROVED RANDOM FOREST CLASSIFICATION (IRFC)**

This study presents a novel approach to Random Forest classification (IRFC), which is an extension of the probability classification model and decision tree. Initializing the amount of tree and target variables is the first phase in this classification process. The final feature selection serves as the test feature and each pre-processed characteristic serves as the practice feature. These are the inputs for categorization. This method first transforms the training and testing functions to obtain the feature set for the feature size. The IRFC method's ability to handle high-dimensional datasets effectively is one of its advantages. The IRFC Algorithm 2 predicts outcomes more accurately than the support vector machine method.

To determine the prediction class in equation (4), an IRFC classification case is organized using a range of First Order Logical Decision (FOLD) techniques. Its functions enable the random selection functions to be organized into 22 classes.

Where made are the attribute; C is the quantity of class; Indicates an unique attribute of the qualified dataset. The classification accuracy is consistent with Equation (5) after the feature selection process and the IRFC.

**Algorithm 2:**

**IRFC Algorithm**

**Input:** Taining Dataset T, Set of features

**Output:** Accuracy Prediction

**Begin Process**

**Step 1:** Initialization Pd = 0 // Pd is prediction

**Step 2:** **For** *k* = 1: *FB*

*TR (k)* = A sample from feature

*Pd(i)* = Learn(*TR(k)*, *F*)

*PD* = *PD* ∪ {*pd*(*i*)} using eqn. (8)

PA ← PD

**End for**

**Function** Learn(*TR*, *F*)

At every Node:

*F* ← a small subset of *F*

Calculate selection unique feature using eqn, (7)

return prediction tree

**end function**

**IV. RESULTS AND DISCUSSION**

By using Python version 3.8 simulation, through experimental evaluation it is found that the recommended IRFC Algorithm perform fine when run on a 3.21GHz x64 – based Intel I5 series pc with 8 GB memory and windows 10 OS.

**4.1 MODEL EVALUATION**

The planned IRFC method was used to compute dataset Mean Absolute Error, R-Square, Recall, precision, F1-Score and accuracy measures. While we comparing the dataset by splitting 60% as train set and 40% as test set and then also we made a split of 70% training and 30% testing. Again we are splitting the dataset into 80% as train and 20% as test and splitting the same data set to 90% training and 10% as testing. By doing this we will comes to know that which splitting is best for our model and which splitting gives best accuracy level for our prediction. Lastly we will compare all of the split-up dataset results and the comparisons of Mean Absolute Error, R-Square, Precision, Recall, F1-Score and accuracy measures of training and testing is described in table 1, and figure 4 showing the 60% training set and 40% testing set performance measures comparison chart. Figure 5 showing the 70% training set and 30% testing set performance measures comparison chart. Figure 6 showing the 80% training set and 20% testing set performance measures comparison chart. Figure 7 showing the 90% train set and 10% test set performance measures comparison chart.

**Table 1: Performance measures of the crop yield model using RF Classifier.**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Random Forest Classification Algorithm** | | | | | | | | | | |
|  | **R2 (%)** | | **MAE** | | **Precision (%)** | | **Recall (%)** | | **F1 Score (%)** | |
| **Splitting Size** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| 60% - Train  40% - Test | 91.88 | 88.71 | 0.013 | 0.138 | 90.12 | 89.30 | 90.30 | 88.48 | 90.94 | 90.12 |
| 70% - Train  30% - Test | 81.10 | 78.31 | 0.015 | 0.158 | 88.5 | 87.69 | 88.69 | 86.89 | 89.30 | 88.50 |
| 80% - Train  20% - Test | 95.11 | 91.83 | 0.012 | 0.128 | 91.25 | 90.42 | 91.42 | 89.59 | 92.07 | 91.25 |
| 90% - Train  10% - Test | 96.27 | 92.95 | 0.012 | 0.128 | 90.00 | 89.18 | 89.27 | 88.36 | 90.81 | 90.15 |

**4.2 PERFORMANCE MEASURES**

* **R-Square (R2):** The coefficient of determination, also referred to as the R2 score, is used to evaluate the performance of a classification model. It is the extent to which the independent variable or factors in the input predict the change of the yield dependent attribute. The effectiveness of the model's simulation of observed findings is evaluated based on the ratio of the sum deviation of the results the model describes, which is displayed in Equation (6).

(6)

* **Mean Absolute Error (MAE):** It is the mean variation between the original and predict values shown in Equation (7). In essence, it shows how the actual result matches our predictions. One drawback, though, is that it provides no indication of the error's direction—that is, whether our data is being over or under-predicted.

(7)

* **Precision:** Accuracy in predictive modeling is the proportion of accurately anticipated optimistic observations to all anticipated optimistic observations shown in Equation (8). In organize to ensure the true precision and dependability of the positive predictions the representation generates—qualities required for decision-making and, consequently, error reduction in many other domains—it shows how well the model reduces false positives.

(8)

* **Recall:** In predictive modeling, recall is the proportion of actual optimistic cases that the model properly recognized shown in Equation (9). It is useful in fields like fraud detection and medical diagnostics where finding all positives is crucial since it demonstrates how well the model finds all pertinent instances of a given class.

(9)

* **F1-Score:** Equation (10), which calculates the harmonic mean of accuracy and recall, provides a comprehensive assessment of a model's presentation. Avoiding both false positives and false negatives is particularly helpful with this.

(10)

* **Accuracy:** Accuracy in predictive modeling refers to the degree to which the model's forecasts and actual results shown in Equation (11). It measures how reliable and accurate the model is, which is important for making judgments and predictions in a range of situations.

(11)

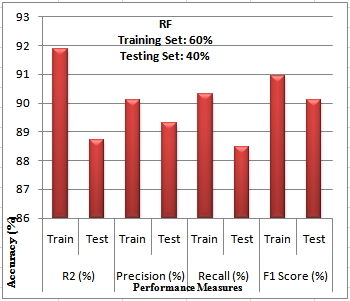


Fig.4. Comparison chart for the performance measures 60% - 40%

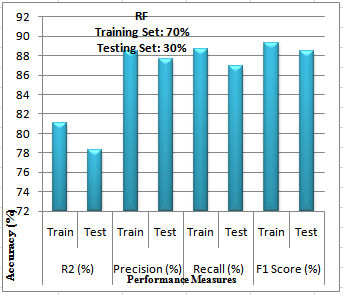


Fig.5. Comparison chart for the performance measures 70% - 30%

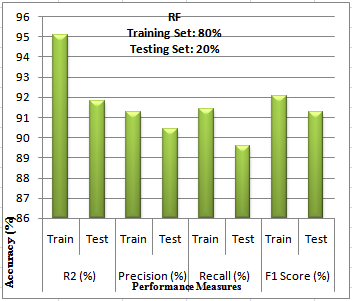


Fig.6. Comparison chart for the performance measures 80% - 20%

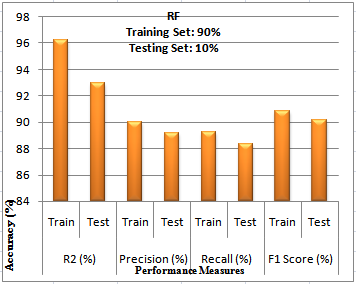


Fig.7. Comparison chart for the performance measures 90% - 10%

The above comparison chart of splitting up of dataset has shown the results. For this performance measures the accuracy has been achieved through the classification model. Below figure 8 shows the confusion matrix of the best splitting as training set 70% and testing set as 30%.

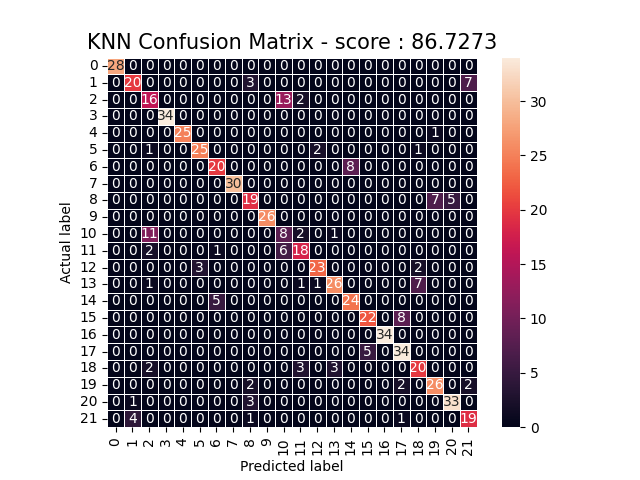


Fig.8. Confusion Matrix

**V. CONCLUSION**

This document proposes a Random Forest classification algorithm for predicting agricultural productivity. This categorization technique allows for the intelligent training of various crops' soil properties to forecast the quality of their output. Therefore, compared to training each crop type separately, using multiple data divisions, the training efficiency and prediction accuracy for estimating the yield of numerous crops simultaneously using a Random Forest classification method are increased. The best crop prediction accuracy in this instance is achieved by splitting the train-set into 70% and the test-set into 30%. So, it can be said that this Random Forest classifier, which performs better than the other classification approach, able to be very supportive in real-time agricultural application.

**REFERENCES**

|  |  |
| --- | --- |
| [1] | Yogesh Gandge, Sandhya, “A Study on Various Data Mining Techniques for Crop Yield Prediction”, 2017 International Conference on Electrical Electronics, Communication, Computer and Optimization Techniques (ICEECCOT). |
| [2] | R. Medar, V. S. Rajpurohit and S. Shweta, "Crop Yield Prediction using Machine Learning Techniques," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-5. |
| [3] | A. Nigam, S. Garg, A. Agrawal and P. Agrawal, "Crop Yield Predict ion Using Machine Learning Algorithms," 2019 Fifth International Conference on Image Information Processing (ICIIP), 2019, pp. 125-130. |
| [4] | J. Sun, Z. Lai, L. Di, Z. Sun, J. Tao, Y. Shen, Multilevel deep learning network for county-level corn yield estimation in the u.s. corn belt, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13 (2020) 5048–5060, doi: 10.1109/JSTARS.2020.3019046. |
| [5] | Lontsi Saadio Cedric, Wilfried Yves Hamilton Adoni, Rubby Aworka, Jérémie Thouakesseh Zoueu, “Crops yield prediction based on machine learning models: Case of West African countries”, Smart Agricultural Technology, Published by Elsevier, March 2022, https://doi.org/10.1016/j.atech.2022.100049. |
| [6] | Dr. S. Selvi, Dr. M. Gobi, “Improving Cloud Data Security using Hyper Elliptical Curve Cryptography & Stenography”, Int. J. Scientific Research & Development, Vol. 5, Issue 4, 2017. |
| [7] | S. Selvi, M. Gobi, “Hyper Elliptic Curve Cryptography in Multi Cloud – Security using DNA (genetic) techniques”, Int. Con. On computing Methodologies and Communication (ICCMC), July 2017. |
| [8] | Dr. M. Gobi, B. Arunapriya, “A Survey on public–key and Identify–Based Encryption Scheme with Equality Testing over Encrypted Data in Cloud Computing”, Journal of algebraic Statistics, Vol.13,No.2,2022, p.2129-2134. |
| [9] | G. Buvaanyaa, M. Gobi, “Crop Recommendation System Using Hybrid Classification Algorithm”, African Journal of Biological sciences”, Volume 6, Issue 14, Aug 2024, doi: 10.48047/AFJBS.6.14.2024.8913-8918. |
| [10] | S. Selvi, M. Gobi, “Hyper Elliptic Curve based Homomorphic Encryption Scheme for Cloud Data Security”, Int. Con. On Intelligent Data Communication Technologies and Internet of Things (ICICI), 2018. |
| [11] | G. Mariammal, A. Suruliandi, S. P. Raja, and E. Poongothai, “Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modied recursive feature elimination technique with various classiers,'' IEEE Trans. Computat. Social Syst., vol. 8, no. 5, pp. 1132-1142, Oct. 2021. |
| [12] | Shafiulla Shariff, R B Shwetha, O G Ramya, H Pushpa and K R Pooja, "Crop Recommendation using Machine Learning Techniques", Volume&Issue-ICEI-2022, vol. 10, no. 11. |
| [13] | M. Shahhosseini, G. Hu, I. Huber, S. Archontoulis, Coupling machine learning and crop modeling improves crop yield prediction in the US corn belt, Sci. Rep. 11 (2021), doi: 10.1038/s41598-020-80820-1. |
| [14] | S. Khaki, L. Wang, Crop yield prediction using deep neural networks, Front. Plant Sci. 10 (2019), doi: 10.3389/fpls.2019.00621. |
| [15] | F. Abbas, H. Afzaal, A.A. Farooque, S. Tang, Crop yield prediction through proximal sensing and machine learning algorithms, Agronomy 10 (7) (2020), doi: 10.3390/agronomy10071046. |
| [16] | A. Kaneko , T. Kennedy , L. Mei , C. Sintek , M. Burke , S. Ermon , D. Lobell , Deep learning for crop yield prediction in Africa, 2019. |
| [17] | Dr. M. Gobi, R. Sridevi, “ECC Encryption and LSB Data Embedding Technique for Message Security”, Int. J. for Res. In Technological Studies, Vol.2, Issue 7, June 2015. |
| [18] | G. Siva Brindha, Dr. M. Gobi, “CryptoData MR: Enhancing the Data Protection using Cryptographic Hash and Encryption/Decryption Through Map Reduce Programming Model”, International Conference on Innovative Computing and Communcations (ICICC) – 2023. |
| [19] | M. Gobi, G. Buvaanyaa, “An Efficient Naïve Bayes Imputation Method for Missing Values”, IRJMETS, Vol.2, Issue 7, July-2020. |
| [20] | R. S. Vindan, M. Gobi, “Pinning based Energy aware Computation Offloading for Mobile Cloud Computing”, First Int. Con. Computational Science and Technology (ICCST), 2022. |
| [21] | Dr. S. Selvi, Dr. M. Gobi, “An Efficient Data Security Model using Hyper Elliptic Curve Cryptography and Stenography”, Int. J. Research and Development in Technology, Vol. 7, Issue 6, June – 2017. |
| [22] | G. Siva Brindha, Dr. M. Gobi, “Improving the Map Reduce Performance using Symmetric Key Algorithm”, International Journal of Science and Research (IJSR), Vol. 10, Issue 4, April 2021. |
|  |  |
|  |  |