IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 11, 2022, accepted February 22, 2022, date of publication February 24, 2022, date of current version March 7, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154350

Crop Prediction Based on Characteristics of the Agricultural Environment Using Various Feature Selection Techniques and Classifiers

S. P. RAJA[1], BARBARA SAWICKA[2], ZORAN STAMENKOVIC[3], (Senior Member, IEEE), AND G. MARIAMMAL[4]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
[2]Department of Plant Production Technology and Commodities Science, University of Life Sciences in Lublin, 20-950 Lublin, Poland
[3]IHP—Leibniz-Institut für innovative Mikroelektronik, 15236 Frankfurt (Oder), Germany
[4]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Tamil Nadu 626126, India

Corresponding author: S. P. Raja (avemariaraja@gmail.com)

ABSTRACT Agriculture is a growing field of research. In particular, crop prediction in agriculture is critical and is chiefly contingent upon soil and environment conditions, including rainfall, humidity, and temperature. In the past, farmers were able to decide on the crop to be cultivated, monitor its growth, and determine when it could be harvested. Today, however, rapid changes in environmental conditions have made it difficult for the farming community to continue to do so. Consequently, in recent years, machine learning techniques have taken over the task of prediction, and this work has used several of these to determine crop yield. To ensure that a given machine learning (ML) model works at a high level of precision, it is imperative to employ efficient feature selection methods to preprocess the raw data into an easily computable Machine Learning friendly dataset. To reduce redundancies and make the ML model more accurate, only data features that have a significant degree of relevance in determining the final output of the model must be employed. Thus, optimal feature selection arises to ensure that only the most relevant features are accepted as a part of the model. Conglomerating every single feature from raw data without checking for their role in the process of making the model will unnecessarily complicate our model. Furthermore, additional features which contribute little to the ML model will increase its time and space complexity and affect the accuracy of the model's output. The results depict that an ensemble technique offers better prediction accuracy than the existing classification technique.

INDEX TERMS Agriculture, classification, crop prediction, feature selection.

I. INTRODUCTION

Crop prediction in agriculture is a complicated process [1] and multiple models have been proposed and tested to this end. The problem calls for the use of assorted datasets, given that crop cultivation depends on biotic and abiotic factors [2]. Biotic factors include those elements of the environment that occur as a result of the impact of living organisms (microorganisms, plants, animals, parasites, predators, pests), directly or indirectly, on other living organisms. This group also includes anthropogenic factors (fertilization, plant protection, irrigation, air pollution, water pollution and soils, etc.). These factors may contribute to the occurrence of many changes in the yield of crops, cause internal defects, shape defects and changes in the chemical composition of the plant yield. The shaping of the environment as well as the growth and quality of plants is influenced by abiotic and biotic factors. Abiotic factors can be divided into physical, chemical, and other. The recognized physical factors include: mechanical vibrations (vibration, noise), radiation (e.g., ionizing, electromagnetic, ultraviolet, infrared); climatic conditions (atmospheric pressure, temperature, humidity, air movements, sunlight); soil type, topography, soil rockiness, atmosphere, and water chemistry, especially salinity. The chemical factors include: priority environmental poisons, such as sulfur dioxide and derivatives, PAHs; nitrogen oxides and derivatives, fluorine, and its compounds, lead and its compounds, cadmium and its compounds, nitrogen fertilizers, pesticides, carbon

The associate editor coordinating the review of this manuscript and approving it for publication was Yongming Li.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

monoxide. The others are: mercury, arsenic, dioxins and furans, asbestos, and aflatoxins [3]. Abiotic factors also include bedrock, relief, climate, and water conditions - all of which affect its properties. Soil-forming factors have a diversified effect on the formation of soils and their agricultural value [4].

Predicting crops yields is neither simple nor easy. The methodology for predicting the area under cultivation is, according to Myers *et al.* [5] and Muriithi [6], a set of statistical and mathematical techniques useful in an evolving and improving optimization process. It also has important uses in design, development, and formulation new as well as improving existing products. Presentation or performance of statistical analysis requires the possession of numerical data. Based on them, conclusions are drawn as to various phenomena and further, on this basis, binding economic decisions can be made. According to Muriithi [6], the better you describe certain phenomena in terms of numbers, the more you can say about them, and with increasing data accuracy you can also obtain more accurate information and make more accurate decisions.

The biggest problem in the temperate climate zone is assessment of agroclimatic factors in terms of shaping the yield of winter plant species, mainly cereals. The key factor influencing wintering yield, which provides access to days with a temperature over of 5° C, their number and frequency, and the number of days in the wintering period with temperatures above 0°C and 5°C. A number of these can be estimated on the basis of public statistics and yield regression statistics in years. Developed models for checking the situation that assess whether they want to be a probation of state policy in the field of intervention in the cereal market. Efficient forecasting of productivity requires forecasting of agrometeorological factors. Aspects related to the variability of these factors may pose a particular problem [7]. Many researchers have dealt with this issue with varying degrees of success [8]–[10].

Grabowska *et al.* [9] predicted narrow-leaf lupine yields for 2050-2060 using weather models and three climate change scenarios for Central Europe: E-GISS model, HadCM3 and GFDL. The fit of the models was assessed by means of the determination coefficient R2, corrected coefficient of determination R2adj, standard error of estimation and the coefficient of determination R2pred calculated using the Cross Validation procedure. The selected equation was used to forecast lupine yield under the conditions of doubling the CO2 content in the atmosphere. These authors stated that the influence of meteorological factors on the yield of narrow-leaved lupine varied depending on the location of the station. The temperature (maximum, average, minimum) at the beginning of the growing season, as well as rainfall during the flowering - technical maturity period, most often had a significant influence on the yield. It has been shown that the predicted climate changes will have a positive effect on the lupine yield. The simulated profitability was higher than that observed in 1990-2008, and HadCM3 was the most favorable scenario.

Dąbrowska-Zielińska *et al.* [8] assessed the usefulness of plant biophysical parameters, calculated from the ranges of reflected electromagnetic radiation recorded by the new generation satellites Sentinel-2 and Proba-V, for forecasting crop yields in Poland. In 2016-2018, ground measurements were carried out in arable fields in the area included in the global crop monitoring network GEO Joint Experiment of Crop Assessment and Monitoring JECAM. Classification of crops was performed using optical and radar images Sentinel-1 and RadarSat-2. The PROtotypical model of Biomass and Evapotranspiration PRO was used to simulate the growth of winter wheat cultivation, to forecast its biomass size. Got high accuracy of 94% of the size of biomass modeled with real biomass.

Li *et al.* [10] found that accurate, high-resolution yield maps are needed to identify spatial patterns of yield variability, to identify key factors influencing yield variability, and to provide detailed management information in precision farming. Varietal differences may significantly affect the forecasting of potato tuber yields with the use of remote sensing technologies. These authors argue that improving potato crop forecasting with remote sensing of unmanned aerial vehicles (UAVs) by incorporating varietal information into machine learning methods has the best chance at present.

There are different challenges in this research area. Currently, crop prediction [11] models generate actual results that are satisfactory, though they could perform better. This paper attempts to propose an enhanced crop prediction model that addresses these issues. The prediction process [12] depends on the two fundamental techniques of feature selection [FS] and classification. Prior to the application of FS techniques, sampling techniques are applied to balance an imbalanced dataset.

A. WIRELESS TECHNOLOGY USED IN AGRICULTURE

ZigBee is a wireless technology used for short-range communications which is one of the most common standards for smart applications. The chief merits of the ZigBee technology arise from its low-cost and low-power functionalities. This makes ZigBee ideal for use on monitoring and data gathering devices where a primary concern is to ensure the longevity of batteries.

ZigBee has massive application in precision farming where the Internet of Things is used for SMART field management by precisely monitoring factors affecting the cultivated crops to facilitate increased and better agricultural output. In such a system, various factors which affect cultivation such as temperature, soil quality, pH, salinity, humidity, etc. are closely monitored to optimize the yield. For example, the nutrient quality of the soil may be accessed to optimize the use of fertilizers such that only areas with poor nutrition quality would be sprayed with fertilizers. Not only does this curb overuse of fertilizers but also reduces the time and money spent on excess fertilizing. Another example of this could be to improve the production of crops that require a constant amount of standing water – such as rice. Sensors can be laid out on the field to monitor the level of water. If the water level falls below the recommended threshold, the system would notify the farmer who may decide what action must be taken. In some cases, the sensor can also be programmed to automatically adjust the water levels by communicating with a control device that regulates the water supply. This also reduces manual labour that may be required to manage the crop.

Z-Wave is a network communication protocol created by the Danish company Zensys. Z-wave is a mesh network that uses low-energy radio waves for communication and is primarily used in SMART home and residential applications. It runs in 868.42 MHz (Europe) and 908.40 MHz (US) due to which it has a larger base range for communication and can communicate through barriers such as concrete, walls, etc. Z-Wave provides a medium for the transmission of small data packets with a throughput of 40kbit/sec which is predominantly used in monitor, transmission, and control applications, unlike Wi-Fi which is chiefly used for high-speed data transfer.

The Z-Wave Alliance, established in 2005, is a conglomeration of Z-Wave affiliated companies which developed appliances operating in Z-Wave in various home, industrial and business activities. The Z-Wave products feature interoperability at the application layer due to which any product, irrespective of its manufacturer, can communicate and effectively co-operate with other Z-Wave products. Every Z-Wave product must pass an established conformance test to prove its interoperability with Z-Wave standards. The Z-Wave Alliance has also laid out strong security standards for devices seeking to receive its certification.

Some advantages of the Z-Wave network are, firstly, its large range as compared to ZigBee. Z-Wave is also less susceptible to disturbances than ZigBee as it does not operate on the 2.4 GHz frequency band which ZigBee and Wi-Fi use. Z-Wave like ZigBee supports low power consumption devices and promotes battery longevity. The devices connected with Z-Wave may enter sleep mode whenever they are not in use to conserve power. Furthermore, all Z-Wave products are thoroughly tested and ensure robust interoperability. This is ensured by the Z-Wave Alliance as all Z-Wave products must obtain a certificate to operate on Z-Wave. The security in Z-Wave is also enhanced by the inclusion of another encrypted security layer. Z-Wave like ZigBee operates in form of a mesh network which allows for an extended range of operations by the introduction of intermediate devices which enables every Z-Wave device to connect to the network without directly connecting with, or being in the range of, the coordinator device. Plus, every Z-Wave device can intercommunicate at the coordinator as well as the intermediate node levels thus ensuring proper communication and smooth working without the involvement of a central device.

Some disadvantages of Z-Wave are, firstly, its low coverage. Thus, to cover a larger area, more devices would be required thus increasing the total cost of implementation. The speed of data transfer in Z-Wave is less (around 100kbps) which restricts it to low data transfer activities such as monitoring and control. Z-Wave can only support up to 232 nodes while ZigBee can potentially support 65000+ nodes. Although security standards in Z-Wave have been enhanced considerably, it still is vulnerable to attacks by a skilled hacker.

Z-Wave like ZigBee can be used for multiple monitoring and control systems in agriculture. The interoperability of the Z-Wave technology can be used to create interconnected agricultural systems which effectively communicate with each other to perform tasks. For example, in a greenhouse, a smart thermostat can be used to monitor the temperature. Whenever the temperature reaches higher than what is considered to be safe, the ventilation system (vents, exhaust fans, etc.) can be signaled to operate. Thus, reducing the temperature of the greenhouse. This way these appliances need not be operated the whole time and can be optimally used to save the cost of electricity.

Apart from this, Z-Wave is used extensively in home automation because of its increased security and its ability to penetrate through walls. Z-Wave can be used in SMART locks for the doors of the house which can be opened on being sent a signal from the user's phone. Z-Wave is also used to make SMART sensors which add an additional level of security to the homes. If a motion was detected in the house when the family was away, they would immediately receive a message on their phones. Sensors can also detect fire and smoke and turn on the sprinklers to contain the damage.

LoRa (Long Range) is a digital wireless communication network used in IoT. It was developed by Cyleo, a French company, which was later acquired by Semtech. Transmission in LoRa occurs over license-free communication bands of width 868MHz (Europe), 915MHz (North America), and 433(MHz Asia). With the use of license-free spectral ranges, the cost for the network provider as well as the end-users is considerably lowered. The key feature of LoRa is its ability to allow low-power communications over a long range. LoRa signals can extend up to 10 miles in open, barrierless areas and up to 3 miles in cities.

The LoRa technology governs the physical layer of transmission while the upper layers of transmission are governed by LoRaWAN (Long Range Wide Area Network). LoRaWAN is an LPWAN networking protocol used for connecting IoT devices to the internet and facilitates bi-directional communication and end-to-end security.

The LoRa Alliance was founded in 2015 by a group of companies to ensure better utilization of LoRaWAN and ensure interoperability of LoRa devices and networks. The LoRa Alliance is a non-profit association dedicated to the promotion and betterment of the LoRaWAN network. Just like Z-Wave Alliance, the LoRa Alliance too has its certification program to ensure interoperability and better provision of services to users. It aims to deliver sustainable and effective IoT applications by developing and promoting the LoRaWAN system.

The chief merits of LoRa are its long-range and low-power delivery. This makes it ideal for use in sensors and control mechanism systems. It has a low bit rate communication thus, it conserves the battery life of connected devices because of its low power requirement. It enables multi-year battery usage. Secondly, it is open-licensed, thus reducing the price of usage. LoRaWAN is the most suitable network for outdoor usage of IoT. LoRa permits inexpensive connectivity for devices in rural and remote areas. It is also of great use in mining and natural resources management operations. LoRa also enables advanced security by implementing a 2-level cryptographic security system. All data transmitted over LoRa is encrypted twice, once by the nodes and once by the LoRaWAN protocol. LoRa is also an open technology with an open and transparent standard. LoRa is backed by tech giants like CISCO and IBM, who are members of the LoRa alliance. The LoRa technology is elementary in nature due to its simplistic implementation and fast deployment.

LoRa however, cannot be used for transmission of large payloads of data. It is also not ideal for continuous monitoring applications. Because of its open frequency spectrum, LoRa may be vulnerable to transmission noise and disturbances. LoRa has been used for monitoring soil moisture content and optimizing irrigation. In a vineyard, all the irrigation valves are fitted with soil moisture sensors. The sensor measures soil moisture content at regular intervals and sends the received data to the LoRa gateway within its range. The Gateway can support up to 1000 sensors in a six-mile radius. The Gateway is connected to an internet router which transmits all the data to a vineyard management application (cloud-based or server-based). Depending upon the requirement, the irrigation valves can be regulated. This has allowed LoRa based farms to save up to 50% more water. Apart from this monitoring climatic conditions such as temperature and humidity can also be done using LoRa. LoRa has played a key role in bringing agriculture and IoT together and in the establishment of SMART farms.

In addition to agricultural monitoring, LoRa is also used in the installation of solar panels. LoRa enables the monitoring of miles-long solar panel networks using low power consumption devices. LoRa can also be used to detect water and gas consumption and be used to make flow adjustments. Furthermore, it can also be used to detect leaks. LoRa is also used in SMART buildings and energy metering. It enables monitoring of energy consumption of all floors of a building and is a step towards building a SMART City.

ZigBee, WiFi, Bluetooth and LoRa are used in agriculture to collect the real time data for prediction process. In this work, real time static agriculture dataset of previous year is used for the prediction process. So, ML techniques are used in this work.

This work uses the Random Over-Sampling Examples (ROSE), Synthetic Minority Over-sampling Technique (SMOTE), and Majority Weighted Minority Over - sampling Technique (MWMOTE) to help balance the given dataset. Feature selection is used to find salient features from the given dataset, resulting in better performance and classification techniques that help identify the target class. Wrapper feature selection techniques such as the Boruta, Recursive Feature Elimination (RFE), and Modified Recursive Feature Elimination (MRFE) are used in this work to discover the dataset's salient features. Several supervised classification techniques, such as the Naïve Bayes (NB), Decision Tree (DT), k Nearest Neighbor (kNN), Support Vector Machine (SVM), Bagging, and Random Forest (RF), are trained with the selected features to predict a suitable outcome from the dataset.

II. RELATED WORK

A. BASED ON SOIL CONDITIONS

Duro *et al.* [13] proposed pixel-based and object-based picture examination approaches for wide land cover classes, applying the three machine learning classifiers like DT, RF and SVM. Honawad *et al.* [14] proposed a digital image analysis approach to approximate the properties of physical soil.

The approach attempts to supplant conventional laboratory approaches in order to eliminate drawbacks such as manual involvement, time consumption, human error, and uncertain predictions. The signal processing method improved the quality of the original image through the use of filters and by computing the features in the enhanced images. The proposed algorithm uses color quantization and texture-based feature extraction by applying the Gabor filter and Laws' mask. Matching is achieved by applying statistical measurements like the mean, standard deviation, skewness, and kurtosis. You *et al.* [15] posited an adaptable and precise technique to anticipate yields by employing openly accessible remote sensing data.

The methodology enhances existing procedures in three different ways. To begin with, a remote detecting network is applied to propose a working methodology. Next, a novel dimensionality reduction procedure is presented that uses a convolutional neural network (CNN) alongside long-term memory. Finally, a Gaussian process is used to investigate and examine the spatio-transient structure of the data and enhance its accuracy. Anantha *et al.* [16] implemented a recommendation system using an associate ensemble model with majority voting. The random tree, Chi-square Automatic Interaction Detection (CHAID), kNN, and Naive Bayes (NB) are used as learners to help determine the most appropriate crop, taking into consideration soil parameters, with the results showing high accuracy and potency. The classified image generated by these techniques consists of ground truth-applied mathematics information Further, it incorporates such data as the parameters of the square measure in terms of the weather and crop yield, as well as state and district-wise crop produce. All of the above are employed to predict specific crop yields in a given set of circumstances. Rale *et al.* [17] developed a forecasting model which uses the default settings along with RF regression for crop yield production.

B. BASED ON ENVIRONMENTAL CONDITIONS

Jones *et al.* [18] modified the Decision Support System for Agrotechnology Transfer (DSSAT) crop model, using a decision support system algorithm. However, it is increasingly difficult to sustain DSSAT crop models, given the different sets of code in operation for different crops. The new design uses a multi-modular approach, comprising a cropping template as well as soil and weather modules. Further, there is a module that monitors light and water in the crops, soil, and environment.

Fernando *et al.* [19] studied data on annual coconut production from 1971 to 2001 in a particular region and assessed its economic impact. The research revealed that the loss sustained by the economy in crop shortage terms was around US $50 million. Ji *et al.* [20] advanced an estimation technique to predict rice yields. The study attempted to determine the effectiveness of artificial neural networks (ANN) in predicting rice yield in mountainous regions. It assessed the efficacy of the ANN, relative to biological parametric variations, and compared the efficiency of multiple bilinear regression models with the ANN model. Boryan *et al.* [21] proposed a decision tree-based technique to depict openly accessible state-level crop cover groups, in accordance with guidelines laid down by the Cropland Data Layer (CDL) and National Agricultural Statistics Service (NASS), and utilizing ground truth collected during the June Agricultural Survey. The proposed work outlines the NASS CDL program. It presents information dealing with handling strategies, order and approval, precision evaluation, and CDL item particulars, and product cost estimation procedure. Hansen and Loveland [22] proposed the use of Landsat to acquire satellite imagery that facilitates remote sensing of the environment. Current strategies for monitoring land cover changes across massive swathes of land commonly utilize Landsat information. Bolton and Friedl [23] created a precise model to forecast maize and soybean yield in the Central United States. Part of their examination included testing the capacity of the MODIS (Moderate Resolution Imaging Spectroradiometer) to catch between-yearly fluctuations in yields. Their outcomes demonstrate that the MODIS two-band Enhanced Vegetation Index outperforms the generally utilized Normalized Difference Vegetation Index in respect to anticipating maize yields. Taking into consideration data using vegetation phenology obtained from the MODIS has fundamentally enhanced the model execution internally as well as cross-wise, over the years.

Dempewolf *et al.* [24] designed and developed a practical wheat yield prediction model for the Punjab Province of Pakistan. Shannon and Motha [25] examined the agricultural lands of North America, Central America, and the Caribbean following various weather and climate-related natural disasters. The latest climate and weather data is needed to help farmers manage agricultural risks. The study discusses climatic uncertainties in agriculture such as drought, flood, typhoons, extreme heat, and freezing temperatures. A decision support system is used to prepare adequately for hazard management prior to the occurrence of a disaster. The Agro Climate Research Centre and Agro Meteorological Department play a critical role in agriculture-based risk management activities. Manjula and Djodiltachoumy [26] analyzed crop yield prediction data supported by association rules for the chosen region, that is, the district of Madras in an Asian nation.

Eswari and Vinitha [27] employed the Bayesian network classification supervised learning model in their proposed approach Environmental characteristics such as temperature and rainfall are analyzed alongside crop information to classify crops like rice, coconut, areca nut, black pepper, and dry ginger. Bayesian network classification is employed to explore the dataset.

C. SURVEY OF MACHINE LEARNING TECHNIQUES FOR CROP PREDICTION

Shivnath and Santanu [28] devised a machine learning approach to examine soil fertility and plant nutrient management. The backpropagation network (BPN) used is trained with inputs on crop growth characteristics, nutrient reserves in the soil, and external applications for crop production. The ML system follows the 3 steps of sampling (different soils with similar properties and completely different parameters), backpropagation, and weight change.

Paul *et al.* [29] designed a system that uses data processing techniques to foretell the class of the soil datasets analyzed in terms of crop yields. The process of prognosticating crop yields is formalized as a sorting rule, using NB and kNN clustering. Pudumalar *et al.* [30] devised an exactness agriculture approach, which is a smart farming technique that uses information on soil property, soil type, and crop yields to help farmers determine the most appropriate cultivable crops based on soil parameters.

A new ensemble model using the random tree, CHAID, kNN, and NB is proposed to recommend crops for a specific land area. Bodake *et al.* [31] developed a soil-based fertilizer guidance system that facilitates topical soil examination to help farmers cultivate the right crop. The tool is intended to be made available in the local language so farmers experience no difficulty in comprehension. Heupel *et al.* [32] proposed an unsupervised fuzzy classification approach that suggests crop types with produce harvested in early spring. The classification results are expected to improve with time. Liu *et al.* [33] investigated the probability of implementing multi-temporal Sentinel-2 satellite images to discern heavy metal-induced stress (i.e., Cd stress) in rice crops in four study areas in Zhuzhou City of Hunan Province in China. Priya *et al.* [34] advocated a crop yield prediction approach using the RF rule. Real-time information from Tamil Nadu state in India was

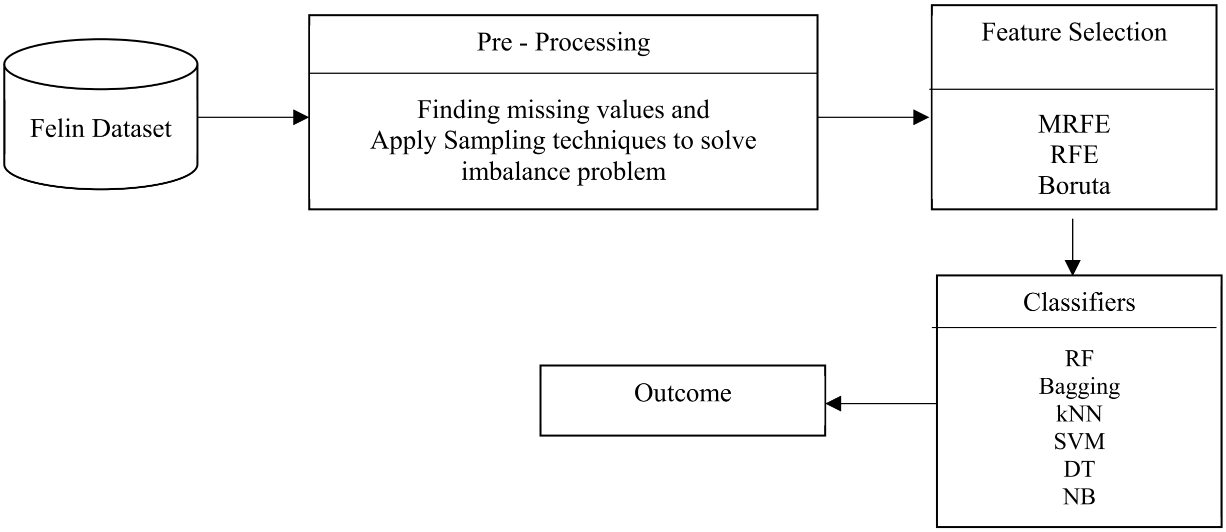

FIGURE 1. Outline of the proposed work.

used to develop the models, which were tested on several samples. The predictions generated help farmers forecast crop yields prior to cultivation.

Archana *et al.* [35] proposed an ontology-based recommendation system for crop quality and fertilizer use, successfully bridging the gap between the farming community and technological applications. The system predicts a relevant crop, taking into consideration the geographical area and soil type, and offers guidance on appropriate fertilizer use. The recommendation system uses the RF rule and k-means clump rule. Brogi *et al.* [36] proposed a superintended classification methodology to classify the Essential Commodities Act (ECA) and map regions on the basis of similar soil characteristics. Ali Al-Naji *et al.* [39] proposed method which is referred as a non-contact vision system based on a standard video camera to solve the irrigation-based problems in the agriculture. The authors have used the feed-forward back propagation neural network to analyze/irrigate the soil which is captured at various times, distances and illumination levels.

D. MOTIVATION AND JUSTIFICATION

Farming performs a vital function in everyday life. Crop prediction in farming, which is a challenge, is based on feature selection and classification. The literature survey above has revealed that crop prediction is best undertaken by feature selection techniques Recursive feature elimination (RFE) is a wrapper feature selection method that searches through a subset of features in the training dataset for the most important ones, eliminating the rest until the desired target is obtained. The RFE technique predicts classification accuracy well. It is, however, limited by the fact that it demands dataset updating during the feature elimination process. Such updating in the RFE is a difficult, time-consuming process. Motivated by these factors, this work proposes a new framework for selecting features from a crop, following which classification is undertaken to predict the crop While existing studies have resorted to a single prediction method, our work uses several classification techniques for crop prediction.

Analysing most of the research papers, the Feature selection techniques like Recursive Feature Elimination, Boruta and Modified Recursive Feature Elimination techniques woks efficiently than other techniques. As well as, the classification techniques k nearest neighbor, decision tree, naïve bayes, support vector machine, random forest, and bagging gives better prediction rate. So, these techniques are taken for the prediction process. Though, all feature selection techniques and classification techniques are existing, the dataset used in this work is real time felin dataset. The felin dataset contains the yield of potato tubers, their yield of dry matter and starch. These are the 7-year averages expressed in (dt/ha) and their coefficients of variation expressed as percentages. Such crops were obtained in the town of Felin, where the meteorological data come from. The outline of the proposed work is given in Fig. 1.

E. OUTLINE OF THE PROPOSED WORK

The rest of the paper is organized as follows. Section 2 discusses the methodology and Section 3 the feature selection techniques. Section 4 describes the classification techniques and Section 5 the experimental design. Section 6 concludes the paper.

III. PREPROCESSING

Sampling techniques are applied during preprocessing to balance the dataset and maximize the prediction performance [37]. The sampling techniques used include the ROSE, SMOTE and MWMOTE. ROSE is used for binary classification in the presence of rare classes and SMOTE for better classifier performance in the ROC space, while MWMOTE handles imbalanced dataset issues in crop prediction.

IV. FEATURE SELECTION TECHNIQUES

There are three commonly used feature selection techniques - filter, wrapper, and embedded. This work uses the wrapper techniques to select salient features.

A. BORUTA

Boruta is a random forest-based classification algorithm [38] that involves the voting of versatile unbiased indistinct classifiers in decision trees. The importance of a characteristic is estimated by calculating the loss of classification exactness caused by the random permutation of attributes within objects. The average and standard deviation of the loss of accuracy are calculated, and the average loss is divided by the standard deviation to obtain the Z score to measure average fluctuations in mean accuracy loss among crops.

A 'shadow' attribute is made for each tree by randomly rearranging the values of the initial attributes across objects. The importance of every attribute is determined by analyzing all the attributes in the system. Given the random nature of the fluctuations, the shadow attributes are used as a reference to point to the most important ones. As is to be expected, the degree of accuracy depends greatly on the shadow attributes. Consequently, the values will be re-shuffled constantly to obtain optimal results.

The Boruta algorithm comprises the following steps:

1. The data system, which is extended by affixing copies of all the shadow attributes, is always prolonged by 5 shadow attributes.
2. The added attributes are shuffled with the original attribute to remove any correlation with the response.
3. The Z score is computed by running a random forest algorithm on the widespread information system.
4. The Maximum Z Score Attributes (MZSA) are calculated and any attribute with a value higher than the MZSA is assigned a "hit".
5. For attributes with undetermined importance, a two-sided test of equality with the MZSA is carried out.
6. Attributes with importance significantly lower than the MZSA are identified as 'unimportant' and permanently eliminated from the information system.
7. Attributes with importance significantly higher than the MZSA are marked 'important'.
8. Shadow attributes are thus eliminated from the information system.
9. The process is repeated until all attributes are marked with a level of importance.

Prior to these steps, however, the algorithm starts with 3 start-up rounds with a simple criterion for importance. The 3 rounds help deal with the tremendous Z score fluctuations when there are large numbers of attributes to be dealt with. In the 3 start-up rounds, the attributes are compared to the 5th, 3rd, and 2nd best shadows. Rejections occur at the end of each initial round and confirmations, on the contrary, in every round.

The time complexity of the algorithm is approximately O (P · N), where P and N are the number of attributes and number of objects, respectively.

B. RECURSIVE FEATURE ELIMINATION (RFE)

The RFE technique is a wrapper feature selection technique that starts with the entire dataset. The ranking method crucial to the RFE technique orders the dataset from the best to the worst, based on which salient features are selected. At each iteration, it eliminates the least important features from the dataset and updates the dataset, continuing the process until the most important ones are selected. RFE is a Wrapper-type feature selection and elimination technique that employs the greedy algorithm. The RFE algorithm recursively identifies and eliminates the least relevant features from the dataset until a sophisticated level of optimization is achieved. In the Wrapper method, the feature selection process is carried out based on a core machine learning algorithm which is fit into the dataset.

In the first step, the model if fit to the dataset i.e., it is generalized. Next, the least significant features are picked from the model and eliminated until only the desired and most important features remain. For a simpler understanding, the number of features along with the performance of the model can be plotted. As relevant features are added, the performance of the model increases. Once all the relevant features are added, the performance of the model will start decreasing upon the addition of redundant features which will be characterized by a drop in the performance level on the graph. Thus, an optimal level of performance can be achieved by selecting the right number and type of features.

It is not known in advance how many features a model must keep. Therefore, to determine the optimal number of features, the RFE algorithm is cross-validated. Recursive Feature Elimination Cross-Validation (RFECV) works just like RFE but, in addition to RFE, it cross-validates the features, automatically selecting the features which give the best performance. All models cannot be paired with the RFE since the RFE starts by considering the entire set of predictors. In models where the number of predictors is more than the number of samples. Furthermore, some models benefit more from RFE than others.

The main advantage the RFE has over other methods is that it categorically verifies every feature's role in processing the output of the model and eliminates features only based on their performance. Thus, producing better results in comparison to filter methods. RFE is also better suited for small sample problems. By using multiple parameters like soil texture, pH, wetness, topography, gypsum content, etc. machine learning can be used to assess the land suitability which helps plan suitable use of the land for agriculture. Here multiple features may be considered to determine the suitability of land but it is not known in advance which features play a key role in determining the final output and which features bring about added redundancies. Thus, the RFE algorithm can be used here to eliminate insignificant features to help improve the accuracy of the model. ML algorithms were used to forecast a short-term soil moisture content in potato crop farming. RFE was one of the algorithms used here to select the most significant features which affected the soil moisture from a set of initial features. Such data may assist in agronomical decision-making.

Recursive Feature Elimination is often combined with the Random Forest algorithm to tackle the presence of correlated predictors which inhibit the accuracy of the Random Forest algorithm. This has shown positive results in smaller data sets. Monitoring pasture quality in farmlands is essential to ensure efficient pasture management. Hyperspectral imaging can be used to determine the biological properties of vegetation in pasture areas. Airborne hyperspectral imaging was used for predicting crude protein and metabolizable energy in farms. The data measure was developed into regression models which used Random Forest. The accuracy of the model showed a significant improvement when RFE was used in conjunction with Random Forest

C. MODIFIED RFE

The MRFE, a wrapper feature selection technique, removes non-salient features from the dataset. Initially, the MRFE technique permutes the dataset by shuffling it, following which it combines the shuffled and original datasets. The permutation dataset reduces computation time and eliminates the need for dataset updating. Thereafter, using a RF classifier, it ranks the features in order from the best to the worst. Based on the ranking result, it selects salient features for the prediction process.

V. CLASSIFICATION TECHNIQUES

A. NAÏVE BAYES (NB)

The Naive Bayes algorithm, derived from Bayes' theorem, is widely used in miscellaneous classification tasks. The three Naive Bayes algorithms are the multinomial, Bernoulli, and Gaussian. The three Naive Bayes algorithms are the multinomial, Bernoulli, and Gaussian. It is shown in Equation 1.

$$P(y \mid X) = \frac{P(X \mid y).P(y)}{P(X)} \tag{1}$$

where $P(y \mid X)$ = Posterior probability

$P(X \mid y)$ = Likelihood

$P(X)$ = Evidence

$P(y)$ = Prior probability.

The Naïve Bayes Algorithm is a supervised machine learning algorithm that is mainly used for classification problems. It works under the assumption that the probability of occurrence of any feature is independent of the occurrence of other features and every feature contributes equally to the final outcome. It is based on the Bayes theorem for calculating the probability of events given the occurrence of another event. The Bayes theorem aims to calculate the probability of an event occurring given that another event is true. Naïve Bayes algorithm is a probabilistic classifier technique i.e., it predicts the outcome based on probability.

Given a labeled dataset and a target variable, the Naïve Bayes Algorithm would calculate the result based on probability. First, the entire dataset is preprocessed and organized into a frequency table by noting down the events and their frequencies. Then, a likelihood table is generated using the frequency table. Finally, the Bayes theorem is applied to calculate the posterior probability.

The advantages of the Naïve Bayes Algorithm are, first, it can be used for binary as well as multi-class classification of data. Secondly, it is fast and easier to implement than the other ML algorithms. It also does not require a lot of training data. It can work with both discrete and continuous data. It's highly scalable and not sensitive to irrelevant features. When the assumption of independence is true, Naïve Bayes Algorithm performs better than other algorithms. The main disadvantage of the Naïve Bayes Algorithm is that it doesn't work well with correlated variables since it works on the assumption of independence and in real-time practice, there aren't many variables that do not correlate with each other.

In agriculture, the Naïve Bayes method can be used to make lucrative food crop recommendations. The Naïve Bayes Algorithm is used to classify weather data, agricultural products, and selling prices to recommend types of food crops to farmers. This recommendation of food crops would be extremely helpful to farmers especially in an era of climate change. Better choice of food crops would mean better income for farmers at the same time reducing the possibility of crop failures.

B. DECISION TREES

A decision tree is a flowchart-like tree structure generally used in supervised machine learning for classification and prediction. A DT can be turned into a set of rules where each path, heading from the root node to every leaf node, is a rule. In a decision tree, every internal node represents a test/condition or an attribute, every branch is a result of the test, and every leaf node has a class ascribed to it that is reachable if the attribute fulfils the condition of the branch leading to it. A famous example of a decision tree is the C4.5 by Ross Quinlan. There are, broadly, two types of decision trees, categorical and continuous variable, based on target attribute types. A decision tree starts with a root node that is compared with other attributes/features in the dataset for a perfect split. A perfect split implies that the number of outputs of one class are on one side of the tree and those of the other class on the other. In this way, every node gets split until it reaches a perfect split, the outcome of which becomes the leaf node of a tree. The real challenge in constructing a DT is attribute selection. That is, given the large number of attributes available, it is difficult to select the ones to be used as root nodes or internal nodes. To this end, there are two techniques that can be applied, Information Gain and Gini Index:

Information Gain (T, X) = Entropy (T) – Entropy (T, X), where T refers to the current state and X to the selected attribute;

$$\begin{aligned}\text{Gini index} &= 1 - \boldsymbol{\Sigma}(\text{p})\hat{}2 \\ &= 1 - [(\text{p}+)\hat{}2 + (\text{p}-)\hat{}2], \end{aligned} \tag{2}$$

where p^{+} represents the probability of Yes/Good and p- the probability of No/Bad.

C. SUPPORT VECTOR MACHINE (SVM)

In AI, support vector machines (SVMs) are supervised learning models that dissect information for order and relapse examination. Given a bunch of models, each set apart as possessing a place with one of two classes, the SVM assembles a model that designates new guides to one type of classification or another, rendering it a non-probabilistic double-straight classifier (although strategies such as Platt scaling, for example, exist to utilize the SVM in a probabilistic order setting). The SVM works to widen the gap between two classifications. New models are planned that fit in and have a place with a class that is dependent on which side of the gap they fall. SVMs help tackle an array of certifiable issues. They are useful in text and hypertext form, as their application decreases the demand for named preparing examples in both conventional inductive and transductive settings. A few techniques for shallow semantic parsing depend on support vector machines. SVMs have widespread uses in the natural sciences. They have been employed to group proteins, with up to 90% of the mixtures ordered accurately. Change tests reliant on SVM loads have been proposed as a system for the translation of SVM models. Support-vector machine loads have furthermore been employed to decipher SVM models in the past. Post-hoc understanding of support-vector machine models to specify highlights appropriated by the model to make expectations is a usually new space of exploration with exceptional importance in the natural sciences.

Support Vector Machine or SVM is a supervised ML technique that is used to solve regression and classification problems but is more suited to classification. SVM works well on small data sets but is more robust and efficient with large data sets. Given a dataset with n features, SVM initiates with plotting all points in the dataset in an n-dimensional space, and each point is assigned a coordinate according to the value of its features. Hereon, the classification process is conducted by determining a suitable hyperplane which to the furthest extent, differentiates the points into two distinct classes. Support vectors are essentially the points that are located close to the hyperplane and determine its position and orientation. The distance between the support vectors and the hyperplane is called the margin and to generate the most accurate hyperplane, the margin needs to be maximized as far as possible.

The advantages of the SVM algorithm are, firstly it is very effective in analyzing high dimensional datasets. It is of great use in cases where the number of dimensions is greater than the number of samples. SVM utilizes the support vectors for training and therefore consumes less memory.

Some disadvantages of the SVM algorithm are, firstly it is not suitable for very large datasets as the time required to train the model increases. It also gives inaccuracies when the target classes overlap with each other. Furthermore, the SVM algorithm cannot account for probability. SVM is used to classify agricultural data to allow for better decision-making. In a comparative study of classification techniques used for agricultural data, SVM was able to outperform Naïve Bayes and Artificial Neural Network methods.

D. K-NEAREST NEIGHBOR (KNN)

One of the most commonly used supervised and non-parametric machine learning techniques is the kNN, used in classification and regression problems. Supervised algorithms are the ones with labelled data. Labeled data refers to input data that is already tagged with the correct output in supervised learning. Supervised learning algorithms take the data and attempt to make models that predict the output data, given relevant inputs. There is, however, no actual learning in the k-NN algorithm, which follows the ‘‘lazy learning’’ principle, where all the work happens at the time a prediction is required.

The algorithm depends on distances between points, which can be ascertained using one of a few methods. A key aspect for consideration is that the distance is always required to be either zero or positive. This is done by squaring the distance or raising it to a certain power or taking the absolute values. Methods to find distances include the following:

Manhattan Distance

This distance is easier to calculate than the others.

$$|x_2 - x_1| + |y_2 - y_1| \tag{3}$$

(i) Euclidean Distance

This is the distance between two points, used in regular geometry.

$$((x_2 - x_1)^2 + (y_2 - y_1)^2)^{1/2} \tag{4}$$

(ii) Hamming Distance

This method finds distances by depending on common neighbors. $|x_1 - y_1| + |x_2 - y_2|$ If x_1 and y_1 are of the same type, their difference is 0, else it is 1.

(iii) Minkowski Distance

Similar to the Euclidean distance, an ‘‘n’’ value is needed here, $((x_2 - x_1)^p + (y_2 - y_1)^p)^{1/p}$ where xi and yi are the x and y coordinates of a point on an xy plane.

The k-nearest neighbors (KNN) method is a supervised machine learning algorithm used to solve classification and regression problems. kNN works on the premise of similar entities existing in close proximity. Related data fields would therefore occur nearby. This helps us in mapping similarities between datasets and a given query.

Before we implement the kNN method, all the labeled data must be pre-processed. Firstly, all the data must be normalized. Next, feature selection must be employed to delete the irrelevant features as kNN doesn't work well when too many features are present. Missing values cannot be tolerated, thus in the case of missing values that particular row must be deleted. Hereon, we can move towards the implementation of the kNN algorithm.

First, data is loaded into the model. kNN being a supervised learning method requires data to be loaded in labeled form. Next, K is declared according to the desired number of neighbours. Then, for every element in the dataset, the "distance" or "relation" with the query input is calculated by the machine learning algorithm. The distance between the element and the query input is then added to an ordered collection and is subsequently sorted in increasing order of the distances. Lastly, the first K items of the collection are selected and the output, depending upon the model being a regression or a classification problem, is returned by taking a mean, in case of the former, and taking the mode in case of the latter.

Choosing k is an important factor as it heavily influences the result of our ML model. If the value of K is too low, the model suffers from instability and the results become increasingly inaccurate. Conversely, an extremely high value of k will start furnishing an increased number of errors in the model. Therefore, the value of k must be balanced between the two extremums. In the case of a model where a vote is required to get the output, K should be taken as an odd number to ensure a deciding game. The chief advantages of the kNN algorithm are, firstly, it is simple and relatively easy to implement. Next, the algorithm can serve multiple purposes, right from classification and regression to searching problems. Furthermore, the algorithm can be improved by adding additional training data. The main disadvantage of kNN is that the speed of the algorithm goes on decreasing as our dataset becomes larger and larger as the cost of computation keeps increasing. Therefore, it is not suitable in cases where immediate results are required. Secondly, we must accurately determine the value of k to get appropriate results. This process can be difficult sometimes. Also, the kNN algorithm requires a large amount of memory to store large sets of data.

The kNN method can be used for predicting crop yield using a set of known parameters, namely, rainfall, temperature, humidity, and soil moisture. The value of crop yield is calculated by using the values of the nearest neighbors. kNN has yielded suitable accuracy in predicting crop yield. This model can be further enhanced by adding additional features and more data from all the seasons. kNN has also been applied for predictive analysis of paddy production and has shown better and faster results compared to the SVM algorithm.

E. BAGGING

Bagging, also known as bootstrap aggregation, is used with decision trees, where it significantly raises the stability of models in terms of advancing accuracy and diminishing variance so as to eliminate overfitting. In ensemble machine learning, bagging takes numerous weak models specializing in distinct sections of the feature space and aggregates them to pick the best prediction. An ensemble set of learners is developed and built utilizing the learning algorithm, but with each learner instructed on a different set of data. Such a process is termed bootstrap aggregating or bagging. Initially, numerous subsets of the data are created, with each being a subset of the initial data. So, a subset containing n'' values have an original dataset comprising n different instances. n'' of these are grabbed at random with replacements from the initial data. An n'', randomly grabbed and put in the bag, is chosen from the entire data collection. It is picked again at random and bagged, implying a certain degree of repetition, resulting in some recurrence. Such recurrence, however, creates no problems and is to be anticipated. So then, m groups or bags are created altogether, with each holding n'' different data instances, randomly picked, with replacements. Thus, n is the number of training instances in the original data, n'' the number of instances in individual bags, and m the number of bags. We nearly always want n'' to be less than n, usually by about 60%. Therefore, each of these bags has, as a rule of thumb, about 60% as many training instances as the original data.

Each of the data groups is used to practice a different model. There are m different models, each one practiced on different data, producing an ensemble of different learning algorithms, along with an ensemble of models to be queried identically. Each model is queried with the equivalent x and all of their output collected. The y output of specific models is taken with their mean to generate the y for the ensemble.

For example, assuming that there are L bootstrap samples, each of size b, then $\left\{z_1^1, z_2^1, \ldots, z_B^1\right\}, \left\{z_1^2, z_2^2, \ldots, z_B^2\right\}, \ldots, \left\{z_1^L, z_2^L, \ldots, z_B^L\right\}$ Whereas, assuming $z_b^l \equiv b$-th observation of the l-th bootstrap sample, L number of nearly independent weak learners can be fitted in $w_1(.), w_2(.), \ldots, w_L(.)$.

F. RANDOM FOREST (RF)

The random forest (RF) is one of the most successful supervised machine learning algorithms. The RF algorithm embodies the essence of ensemble learning in that it links multiple classifiers to resolve a complicated problem, thereby enhancing the performance of the model. In this method, the "forest" that is built is a set of decision trees. Characteristics in the RF are randomly picked in each decision split. The correlation between trees is diminished by randomly picking features that promote prediction and result in greater efficiency.

Random Forest is an ML classification algorithm that works by dividing the dataset into subsets or decision trees and aggregating the outputs of all the trees to produce the final output. Random Forest comes under the Bagging category of ensemble learning techniques. The Row and Feature samples from the main dataset are randomly selected and fed into the decision trees in the Random Forest Technique. The analyst chooses the number of decision trees for the model. Each decision tree works on the data and predicts a result based upon its calculation. Random Forest doesn't take the result from any one of the decision trees but combines the outputs from all the decision trees. Random Forest takes the majority of the result (in case the result is in a Boolean form) or the mean/median of the result (in case the result is in numerical form). Thus, a higher number of decision trees gives a more accurate result and circumvents the problem of overfitting.

The Random Forest technique provides several advantages. Firstly, Random Forest is simple and relatively easy to understand and is therefore extremely popular. It is also capable of performing both classification and regression tasks. It is also suitable for handling large sets of data with high dimensionality and most importantly, it makes the model much more precise and resolves the overfitting issue.

Random Forest cannot be used in case of extrapolation of data as it could produce inaccurate results. Although Random Forest can be used for both regression and classification, it is better suited for classification tasks. Also, it does not produce proper results when dealing with sparse data. Random Forest also needs more time for implementation and requires larger data and greater resources. In the presence of correlated predictors, Random Forest is known to produce inexact results.

Random Forest [40] can be used to predict pest attacks in cotton plants. Various factors were very considered and the Correlation filter selection method was used to select the most important features. Random Forest was then used to determine the number of trees to get a low error rate and important parameters were sighted out and used for clustering to determine the outcome. Optimized usage of water for farms is essential for reducing wastage as well as enhancing productivity. The use of a precision irrigation system for furnishing the optimal water supply needed by plants or crops will lead to better output. The amount of water required by plants can be expressed in terms of pH. The Random Forest algorithm is used to determine the pH level which in turn helps determine the water supply required by a piece of land. The concept of random forest is given in Fig. 2.

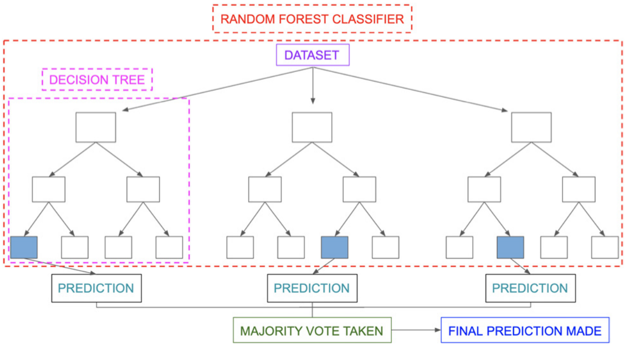

FIGURE 2. **Concept of the random forest.**

Given that each bagged tree is identically disseminated, the expectation of an average of B trees is the equivalent of the expectation of each. Since this accounts for the bias of bagged trees being the same as that of individual trees, a change may only be affected through variance reduction. This contrasts with advancing, where the trees are grown adaptively to exclude bias, and hence are not identically distributed. An average of B identically distributed random variables has a variance of σ^2. If the variables are completely identically distributed, but with positive pairwise correlation ρ, the variance of the average is given as

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{5}$$

It is observed that as B increases, the value of the second term shifts negligibly while that of the first term remains unchanged. Consequently, the size of the correlation of the bagged trees limits the benefits of averaging. The RF focuses on bagging variance minimization by cutting the correlation between the trees without increasing the variance excessively. The tree-growing process makes this procedure possible through picking input variables at random.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

This paper has used an agricultural dataset that incorporates soil and environmental characteristics, both of which are not publicly available. Thus, data manually collected with care from the farming community was used for the purpose of this research.

In this work, the performance of the feature selection and classification methods was assessed using the metrics of accuracy (ACC), specificity (S), recall (R), precision (P), F1 score, mean absolute error (MAE), log loss (LL), and area under the curve (AUC). The results are shown in the tables (Table 1 to Table 5).

Table 1 shows that the random forest algorithm, which consists in constructing many decision trees and generating a class that is the dominant of the all classes (classification) or the predicted mean (regression) of individual decision trees offers the highest accuracy, followed by the k-nearest neighbor and bagging classifiers. The Naïve Bayes Classifier has an accuracy of 70.64, with Kappa equal to 70.12, Precision of 78.80 and Specificity of 92.23. The next classifier that is Decision Tree Classifier has an accuracy of 73.22, Kappa value of 72.85, Precision of 83.24 and Specificity of 92.62. In case of Support Vector Machine Classifier: Accuracy is 77.50, Kappa Value of 75.01, Precision of 83.24 and Specificity of 93.87. The k Nearest Neighbor Classifier, accuracy of 83.24 is evaluated, Kappa value of 80.60, Precision of 87.00 and Specificity of 94.28. In case of Bagging Classifiers, the accuracy is 84.00, Kappa Value is 82.01, Precision is 89.11 and Specificity of 94.63. The last Classifier whose performance was evaluated is Random Forest Classifier which had the highest accuracy amongst all classifiers evaluated with a value of 87.43, Highest Kappa Value of 85.16, Highest Precision Value of 90.34 and Highest Specificity among all Classifiers with value of 95.67. From the table we can infer that Random Forest Classifier has the highest performance based on Felin Dataset.

Table 2 depicts that the random forest method, when used alongside sampling techniques, predicts crops very well. In Table 2, We are finding the most suitable sampling technique to Balance the Dataset. In this table we have analyzed 4 types of Sampling Technique namely, Without Sampling, SMOTE, MWMOTE, ROSE. These Sampling Techniques were further divided into 6 Types of Classifiers, namely, Naïve Bayes, Decision Tree, Support Vector Machine, k Nearest Neighbor, Bagging, Random Forest. We analyzed 9 Performance Metrics, namely, Accuracy, Kappa, Precision, Recall, Specificity, F1 Score, AUC, MAE, Log Loss. Analyzing the data in the table now we see that Without Sampling and with Naïve Bias Classifier we get an accuracy of 70.64, Kappa Value of 70.12, Performance Value of 78.80, Recall Value of 75.32, Specificity Value of 90.23, F1 Score Value of 77.02, AUC Value of 73.69, MAE value of 0.8, Log Loss Value of 0.14. In case of Without Sampling with Decision Tree Classifier we get an accuracy of 73.22, Kappa Value of 72.85, Performance Value of 80.16, Recall Value of 77.94, Specificity Value of 90.62, F1 Score Value of 79.03, AUC Value of 75.97, MAE value of 0.65, Log Loss Value of 0.09.In case of Without Sampling with Support Vector Machine Classifier we get an accuracy of 77.50, Kappa Value of 75.01, Performance Value of 83.24, Recall Value of 80.98, Specificity Value of 91.87, F1 Score Value of 82.09, AUC Value of 80.02, MAE value of 0.5, Log Loss Value of 0.06. In case of Without Sampling with k Nearest Neighbor Classifier we get an accuracy of 83.24, Kappa Value of 80.60, Performance Value of 87.00, Recall Value of 85.32, Specificity Value of 92.28, F1 Score Value of 86.15, AUC Value of 87.97, MAE value of 0.43, Log Loss Value of 0.05. In case of Without Sampling with Bagging Classifier we get an accuracy of 84.00, Kappa Value of 82.01, Performance Value of 89.11, Recall Value of 88.53, Specificity Value of 92.63, F1 Score Value of 88.81, AUC Value of 89.41, MAE value of 0.3, Log Loss Value of 0.04. In case of Without Sampling with Random Forest Classifier we get an accuracy of 87.43, Kappa Value of 85.16, Performance Value of 90.34, Recall Value of 89.12, Specificity Value of 93.67, F1 Score Value of 89.72, AUC Value of 92.39, MAE value of 0.3, Log Loss Value of 0.04. This completes all the data without sampling. In case of SMOTE with the same classifiers when the data was recorded we found this: In the case of SMOTE Sampling Technique with Naïve Bayes Classifier, Accuracy value is 74.76, Kappa Value is 73.63, Precision Value is 78.93, Recall Value is 77.18, Specificity Value is 91.41, F1 Score Value is 78.04, AUC Value is 76.93, MAE Value is 0.6, Log Loss Value is 0.09. In the case of SMOTE Sampling Technique with Decision Tree Classifier, Accuracy value is 75.65, Kappa Value is 74.97, Precision Value is 79.37, Recall Value is 78.64, Specificity Value is 92.36, F1 Score Value is 79.001, AUC Value is 77.42, MAE Value is 0.5, Log Loss Value is 0.07. In the case of SMOTE Sampling Technique with Support Vector Machine Classifier, Accuracy value is 79.81, Kappa Value is 78.08, Precision Value is 82.79, Recall Value is 81.68, Specificity Value is 92.47, F1 Score Value is 82.23, AUC Value is 81.02, MAE Value is 0.4, Log Loss Value is 0.04. In the case of SMOTE Sampling Technique with k Nearest Neighbor Classifier, Accuracy value is 85.72, Kappa Value is 83.62, Precision Value is 86.42, Recall Value is 84.79, Specificity Value is 94.15, F1 Score Value is 85.59, AUC Value is 87.63, MAE Value is 0.33, Log Loss Value is 0.03. In the case of SMOTE Sampling Technique with Bagging Classifier, Accuracy value is 87.39, Kappa Value is 85.71, Precision Value is 88.58, Recall Value is 87.95, Specificity Value is 94.71, F1 Score Value is 88.26, AUC Value is 89.23, MAE Value is 0.2, Log Loss Value is 0.03. In the case of SMOTE Sampling Technique with Random Forest Classifier, Accuracy value is 92.42, Kappa Value is 90.00, Precision Value is 94.47, Recall Value is 92.80, Specificity Value is 95.17, F1 Score Value is 93.62, AUC Value is 94.94, MAE Value is 0.2, Log Loss Value is 0.02. In the case of MWMOTE Sampling Technique with Naïve Base Classifier, Accuracy value is 75.64, Kappa Value is 75.01, Precision Value is 80.41, Recall Value is 79.00, Specificity Value is 92.90, F1 Score Value is 79.69, AUC Value is 77.86, MAE Value is 0.51, Log Loss Value is 0.08. In the case of MWMOTE Sampling Technique with Decision Tree Classifier, Accuracy value is 76.98, Kappa Value is 75.81, Precision Value is 80.74, Recall Value is 79.21, Specificity Value is 93.13, F1 Score Value is 79.96, AUC Value is 78.31, MAE Value is 0.41, Log Loss Value is 0.07. In the case of MWMOTE Sampling Technique with Support Vector Machine Classifier, Accuracy value is 81.66, Kappa Value is 79.56, Precision Value is 83.83, Recall Value is 82.23, Specificity Value is 93.81, F1 Score Value is 83.02, AUC Value is 83.80, MAE Value is 0.38, Log Loss Value is 0.04. In the case of MWMOTE Sampling Technique with k Nearest Neighbor Classifier, Accuracy value is 87.13, Kappa Value is 86.88, Precision Value is 88.81, Recall Value is 86.83, Specificity Value is 94.39, F1 Score Value is 87.80, AUC Value is 89.38, MAE Value is 0.3, Log Loss Value is 0.03. In the case of MWMOTE Sampling Technique with Bagging Classifier, Accuracy value is 89.12, Kappa Value is 87.49, Precision Value is 90.44, Recall Value is 89.15, Specificity Value is 95.45, F1 Score Value is 89.79, AUC Value is 91.97, MAE Value is 0.2, Log Loss Value is 0.02. In the case of MWMOTE Sampling Technique with Random Forest Classifier, Accuracy value is 93.29, Kappa Value is 91.04, Precision Value is 95.86, Recall Value is 94.95, Specificity Value is 96.60, F1 Score Value is 95.40, AUC Value is 95.89, MAE Value is 0.2, Log Loss Value is 0.02.In the case of ROSE Sampling Technique with Naïve Base Classifier, Accuracy value is 73.98, Kappa Value is 72.01, Precision Value is 77.43, Recall Value is 76.21, Specificity Value is 90.65, F1 Score Value is 76.81, AUC Value is 75.99, MAE Value is 0.75, Log Loss Value is 0.1. In the case of ROSE Sampling Technique with Decision Tree Classifier, Accuracy value is 75.20, Kappa Value is 73.16, Precision Value is 77.38, Recall Value is 76.95, Specificity Value is 91.60, F1 Score Value is 77.16, AUC Value is 77.12, MAE Value is 0.63, Log Loss Value is 0.08.In the case of ROSE Sampling Technique with Support Vector Machine Classifier, Accuracy value is 79.18, Kappa Value is 77.05, Precision Value is 80.27, Recall Value is 78.25, Specificity Value is 93.30, F1 Score Value is 79.24, AUC Value is 81.95, MAE Value is 0.47, Log Loss Value is 0.05. In the case of ROSE Sampling Technique with k Nearest Neighbor Classifier, Accuracy value is 84.00, Kappa Value is 82.63, Precision Value is 85.05, Recall Value is 83.16, Specificity Value is 93.51, F1 Score Value is 84.09, AUC Value is 86.00, MAE Value is 0.4, Log Loss Value is 0.05. In the case of ROSE Sampling Technique with Bagging Classifier, Accuracy value is 85.42, Kappa Value is 83.74, Precision Value is 87.34, Recall Value is 86.87, Specificity Value is 94.25, F1 Score Value is 87.10, AUC Value is 87.31, MAE Value is 0.25, Log Loss Value is 0.04. In the case of ROSE Sampling Technique with Random Forest Classifier, Accuracy value is 90.90, Kappa Value is 88.60, Precision Value is 93.80, Recall Value is 91.89, Specificity Value is 94.73, F1 Score Value is 92.83, AUC Value is 92.62, MAE Value is 0.2, Log Loss Value is 0.03. This summarizes the data collected in Table 2 and we can therefore conclude that.

In Table 3, we are identifying the best feature selection techniques with various classifiers using the felin dataset. We have considered 3 Feature Selection namely, MRFE, RFE and Boruta. Each of these have following classifiers namely, Naïve Bayes, Decision Tree, Support Vector Mechanism, k Nearest Neighbor, Bagging, Random Forest. On analyzing the data case by case we find: In case of Feature Selection of MRFE and 6 Selected Attributes, The Naïve Bayes classifier gives an accuracy of 85.64, Kappa Value of 83.11, Precision Value of 87.14, Recall Value of 86.53, Specificity Value of 93.31, F1 Score of 86.83and AUC of 87.62 value. In case of Feature Selection of MRFE and 6 Selected Attributes, The Decision Tree classifier gives an accuracy of 87.98, Kappa Value of 85.52, Precision Value of 88.92, Recall Value of 87.40, Specificity Value of 94.04, F1 Score of 88.15 and AUC of 89.23 value. In case of Feature Selection of MRFE and 6 Selected Attributes, The Support Vector Machine classifier gives an accuracy of 90.66, Kappa Value of 88.93, Precision Value of 91.82, Recall Value of 89.10, Specificity Value of 94.20, F1 Score of 90.43 and AUC of 92.31 value. In case of Feature Selection of MRFE and 6 Selected Attributes, k Nearest Neighbor classifier gives an accuracy of 92.13, Kappa Value of 89.25, Precision Value of 92.60, Recall Value of 91.32, Specificity Value of 95.40, F1 Score of 91.95 and AUC of 94.99 value. In case of Feature Selection of MRFE and 6 Selected Attributes, The Bagging classifier gives an accuracy of 95.12, Kappa Value of 93.04, Precision Value of 96.50, Recall Value of 95.91, Specificity Value of 96.18, F1 Score of 96.20 and AUC of 97.19 value. In case of Feature Selection of MRFE and 6 Selected Attributes, The Random Forest classifier gives an accuracy of 97.29, Kappa Value of 95.17, Precision Value of 98.94, Recall Value of 97.54, Specificity Value of 98.00, F1 Score of 98.23 and AUC of 99.23 value. In case of Feature Selection of RFE and 8 Selected Attributes, The Naïve Bayes classifier gives an accuracy of 84.87, Kappa Value of 82.92, Precision Value of 86.98, Recall Value of 85.29, Specificity Value of 93.18, F1 Score of 86.12 and AUC of 86.76 value. In case of Feature Selection of RFE and 8 Selected Attributes, The Decision Tree classifier gives an accuracy of 86.17, Kappa Value of 84.67, Precision Value of 88.57, Recall Value of 86.77, Specificity Value of 93.98, F1 Score of 87.66 and AUC of 88.85 value. In case of Feature Selection of RFE and 8 Selected Attributes, The Support Vector Machine classifier gives an accuracy of 88.83, Kappa Value of 86.78, Precision Value of 89.81, Recall Value of 88.37, Specificity Value of 94.14, F1 Score of 89.08 and AUC of 90.93 value. In case of Feature Selection of RFE and 8 Selected Attributes, k Nearest Neighbor classifier gives an accuracy of 90.67, Kappa Value of 88.94, Precision Value of 92.11, Recall Value of 89.40, Specificity Value of 94.91, F1 Score of 90.73 and AUC of 92.61 value. In case of Feature Selection of RFE and 8 Selected Attributes, The Bagging classifier gives an accuracy of 94.86, Kappa Value of 92.61, Precision Value of 95.87, Recall Value of 94.03, Specificity Value of 95.77, F1 Score of 94.94 and AUC of 96.08 value. In case of Feature Selection of RFE and 8 Selected Attributes, The Random Forest classifier gives an accuracy of 96.17, Kappa Value of 95.09, Precision Value of 97.74, Recall Value of 95.46, Specificity Value of 97.58, F1 Score of 96.58 and AUC of 98.57 value. In case of Feature Selection of Boruta and 9 Selected Attributes, The Naïve Bayes classifier gives an accuracy of 83.93, Kappa Value of 82.07, Precision Value of 85.93, Recall Value of 84.21, Specificity Value of 92.98, F1 Score of 85.06 and AUC of 85.33 value. In case of Feature Selection of Boruta and 9 Selected Attributes, The Decision Tree classifier gives an accuracy of 85.71, Kappa Value of 83.15, Precision Value of 86.37, Recall Value of 85.02, Specificity Value of 93.50, F1 Score of 85.68 and AUC of 87.77 value. In case of Feature Selection of Boruta and 9 Selected Attributes, The Support Vector Machine classifier gives an accuracy of 86.58, Kappa Value of 84.84, Precision Value of 87.79, Recall Value of 87.01, Specificity Value of 94.00, F1 Score of 87.39 and AUC of 88.56 value. In case of Feature Selection of Boruta and 9 Selected Attributes, k Nearest Neighbor classifier gives an accuracy of 87.95, Kappa Value of 85.16, Precision Value of 91.42, Recall Value of 89.00, Specificity Value of 94.88, F1 Score of 90.19 and AUC of 89.22 value. In case of Feature Selection of Boruta and 9 Selected Attributes, The Bagging classifier gives an accuracy of 94.09, Kappa Value of 92.30, Precision Value of 95.58, Recall Value of 92.81, Specificity Value of 95.60, F1 Score of 93.68 and AUC of 96.30 value. In case of Feature Selection of Boruta and 9 Selected Attributes, The Random Forest classifier gives an accuracy of 94.91, Kappa Value of 93.97, Precision Value of 96.47, Recall Value of 94.00, Specificity Value of 97.01, F1 Score of 95.21 and AUC of 96.07 value.

In Table 4, we evaluate the performance of the MRFE with the RF based on the Fold Validation Method. In this evaluation we take 9 Folds into consideration. For each fold we have 7 performance metrics. For Fold 10, Accuracy

TABLE 1. A performance evaluation of various classifiers based on the felin dataset.

Classifiers	Performance Metrics								
	Accuracy	Kappa	Precision	Recall	Specificity	F1 Score	AUC	MAE	Log Loss
Naïve Bayes	70.64	70.12	78.80	75.32	92.23	77.02	73.69	0.8	0.14
Decision Tree	73.22	72.85	80.16	77.94	92.62	79.03	75.97	0.65	0.09
Support Vector Machine	77.50	75.01	83.24	80.98	93.87	82.09	80.02	0.5	0.06
k Nearest Neighbor	83.24	80.60	87.00	85.32	94.28	86.15	87.97	0.43	0.05
Bagging	84.00	82.01	89.11	88.53	94.63	88.81	89.41	0.3	0.04
Random Forest	87.43	85.16	90.34	89.12	95.67	89.72	92.39	0.3	0.04

TABLE 2. Finding the most suitable sampling technique to balance the dataset.

Sampling Techniques	Classifiers	Performance Metrics								
		ACC	Kappa	P	R	Sp	F1	AUC	MAE	Log Loss
Without Sampling	NB	70.64	70.12	78.80	75.32	90.23	77.02	73.69	0.8	0.14
	DT	73.22	72.85	80.16	77.94	90.62	79.03	75.97	0.65	0.09
	SVM	77.50	75.01	83.24	80.98	91.87	82.09	80.02	0.5	0.06
	kNN	83.24	80.60	87.00	85.32	92.28	86.15	87.97	0.43	0.05
	Bagging	84.00	82.01	89.11	88.53	92.63	88.81	89.41	0.3	0.04
	RF	87.43	85.16	90.34	89.12	93.67	89.72	92.39	0.3	0.04
SMOTE	NB	74.76	73.63	78.93	77.18	91.41	78.04	76.93	0.6	0.09
	DT	75.65	74.97	79.37	78.64	92.36	79.001	77.42	0.5	0.07
	SVM	79.81	78.08	82.79	81.68	92.47	82.23	81.02	0.4	0.04
	kNN	85.72	83.62	86.42	84.79	94.15	85.59	87.63	0.33	0.03
	Bagging	87.39	85.71	88.58	87.95	94.71	88.26	89.23	0.2	0.03
	RF	92.42	90.00	94.47	92.80	95.17	93.62	94.94	0.2	0.02
MWMOTE	NB	75.64	75.01	80.41	79.00	92.90	79.69	77.86	0.51	0.08
	DT	76.98	75.81	80.74	79.21	93.13	79.96	78.31	0.41	0.07
	SVM	81.66	79.56	83.83	82.23	93.81	83.02	83.80	0.38	0.04
	kNN	87.13	86.88	88.81	86.83	94.39	87.80	89.38	0.3	0.03
	Bagging	89.12	87.49	90.44	89.15	95.45	89.79	91.97	0.2	0.02
	RF	93.29	91.04	95.86	94.95	96.60	95.40	95.89	0.2	0.02
ROSE	NB	73.98	72.01	77.43	76.21	90.65	76.81	75.99	0.75	0.1
	DT	75.20	73.16	77.38	76.95	91.60	77.16	77.12	0.63	0.08
	SVM	79.18	77.05	80.27	78.25	93.30	79.24	81.95	0.47	0.05
	kNN	84.00	82.63	85.05	83.16	93.51	84.09	86.00	0.4	0.05
	Bagging	85.42	83.74	87.34	86.87	94.25	87.10	87.31	0.25	0.04
	RF	90.90	88.60	93.80	91.89	94.73	92.83	92.62	0.2	0.03

is 97.29, Kappa Value is 95.17, Precision Value is 98.94, Recall Value is 97.54, Specificity Value is 98.00,F1 Score Value is 98.23,AUC Value is 99.23.For Fold 20, Accuracy is 96.93, Kappa Value is 94.49, Precision Value is 97.01, Recall Value is 96.79, Specificity Value is 89.76,F1 Score Value is 96.79,AUC Value is 98.04. For Fold 30, Accuracy is 94.63,Kappa Value is 92.18, Precision Value is 94.70,Recall Value is 94.48, Specificity Value is 87.45, F1 Score Value is 94.48,AUC Value is 96.46. For Fold 40, Accuracy is 95.95,Kappa Value is 93.2, Precision Value is 95.72,Recall Value is 95.50, Specificity Value is 88.47,F1 Score Value is 95.50,AUC Value is 97.17. For Fold 50, Accuracy is 95.43,Kappa Value is 92.68, Precision Value is 95.20,Recall Value is 94.98, Specificity Value is 87.95, F1 Score Value is 94.98,AUC Value is 97.10. For Fold 60, Accuracy is 94.13,Kappa Value is 91.38, Precision Value is 93.90,Recall Value is 93.68, Specificity Value is 86.65, F1 Score Value is 93.68,AUC Value is 96.44. For Fold 70, Accuracy is 94.83,Kappa Value is 91.98, Precision Value is 94.50,Recall Value is 94.28, Specificity Value is 87.25, F1 Score Value is 94.28,AUC Value is 96.97. For Fold 80, Accuracy is 95.4,Kappa Value is 92.55, Precision Value is 95.07,Recall Value is 94.85, Specificity Value is 87.82, F1 Score Value is 94.85,AUC Value is 97.89. For Fold 90, Accuracy is 93.93,Kappa Value is 92.22, Precision Value is 94.74,Recall Value is 94.52, Specificity Value is 87.49,F1 Score Value is 94.52,AUC Value is 95.79. From Table 3, the best feature selection technique was identified. In this table, the following attributes were confirmed as significant according to the BORUTA algorithm. They were: soil temperature at various depths (5, 10, 20, 50 and 100 cm); air humidity, precipitation; average, minimum and maximum air temperature. Five attributes were confirmed as irrelevant including: cloud cover, visibility, wind direction, snow cover.

According to recursive feature elimination (RFE), six most important variables were selected out of eight. These were the monthly soil temperatures at various depths (5, 10, 20, 50 and 100 cm) and the minimum temperature.

TABLE 3. Identifying the best feature selection techniques with various classifiers using the felin dataset.

FS	Selected Attributes		Classifiers	Performance Metrics						
	No. of Attributes	Selected Attributes		Accuracy	Kappa	Precision	Recall	Specificity	F1 Score	AUC
MRFE	15	6	NB	85.64	83.11	87.14	86.53	93.31	86.83	87.62
			DT	87.98	85.52	88.92	87.40	94.04	88.15	89.23
			SVM	90.66	88.93	91.82	89.10	94.20	90.43	92.31
			kNN	92.13	89.25	92.60	91.32	95.40	91.95	94.99
			Bagging	95.12	93.04	96.50	95.91	96.18	96.20	97.19
			RF	97.29	95.17	98.94	97.54	98.00	98.23	99.23
RFE		8	NB	84.87	82.92	86.98	85.29	93.18	86.12	86.76
			DT	86.17	84.67	88.57	86.77	93.98	87.66	88.85
			SVM	88.83	86.78	89.81	88.37	94.14	89.08	90.93
			kNN	90.67	88.94	92.11	89.40	94.91	90.73	92.61
			Bagging	94.86	92.61	95.87	94.03	95.77	94.94	96.08
			RF	96.17	95.09	97.74	95.46	97.58	96.58	98.57
Boruta		9	NB	83.93	82.07	85.93	84.21	92.98	85.06	85.33
			DT	85.71	83.15	86.37	85.02	93.50	85.68	87.77
			SVM	86.58	84.84	87.79	87.01	94.00	87.39	88.56
			kNN	87.95	85.16	91.42	89.00	94.88	90.19	89.22
			Bagging	94.09	92.30	94.58	92.81	95.60	93.68	96.30
			RF	94.91	93.97	96.47	94.00	97.01	95.21	96.07

TABLE 4. Performance evaluation of the MRFE with the RF based on the fold validation method.

Method	Folds	Performance Metrics						
		Accuracy	Kappa	Precision	Recall	Specificity	F1 Score	AUC
MRFE with RF	10	97.29	95.17	98.94	97.54	98.00	98.23	99.23
	20	96.93	94.49	97.01	96.79	89.76	96.79	98.04
	30	94.63	92.18	94.70	94.48	87.45	94.48	96.46
	40	95.95	93.2	95.72	95.50	88.47	95.50	97.17
	50	95.43	92.68	95.20	94.98	87.95	94.98	97.10
	60	94.13	91.38	93.90	93.68	86.65	93.68	96.44
	70	94.83	91.98	94.50	94.28	87.25	94.28	96.97
	80	95.4	92.55	95.07	94.85	87.82	94.85	97.89
	90	93.93	92.22	94.74	94.52	87.49	94.52	95.79

TABLE 5. Performance evaluation of the MRFE with the RF based on the data splitting validation method.

Method	Data Splitting	Performance Metrics						
		Accuracy	Kappa	Precision	Recall	Specificity	F1 Score	AUC
MRFE with RF	25-75	84.51	81.16	84.45	88.4	87.21	86.37	86.26
	30-70	89.17	85.53	86.54	89.95	89.60	88.21	91.01
	35-65	91.81	88.67	88.74	90.14	91.36	89.43	93.21
	40-60	94.67	90.38	90.92	92.63	93.85	91.76	96.79
	45-55	96.37	91.97	92	93.21	94.04	92.60	98.04
	50-50	94.71	91.12	91.37	93.77	94.74	92.55	96.97
	55-45	96.9	93.97	93.46	95.64	96.13	94.53	98.42
	60-40	96.52	92.45	95.38	94.12	95.72	94.74	97.51
	65-35	97	94.37	97.57	96.42	97.25	96.99	98.78
	70-30	97.29	95.17	98.94	97.54	98.00	98.23	99.23
	75-25	97.11	95.55	98.56	97.02	97.98	97.78	99.00

According to the modified elimination of recursive features (MRFE), 6 variables were selected: average soil temperature, average air temperature, minimum and maximum air temperature, rainfall, and air humidity. Performance metrics, as: Accuracy, Kappa, Precision Recall, Specificity, F1 Score, AUC were at a high level.

In Table 5, we evaluate the performance of the MRFE with the RF based on the Data Splitting Validation Method. In case of Data Splitting (25-75), we get the Accuracy value as 84.51, Kappa value 81.16, Precision value as 84.45, Recall value 88.4, Specificity Value 87.21, F1 Score 86.37, AUC 86.26.

In data splitting (30-70), we get an Accuracy value of 89.17, Kappa Value 85.53, Precision Value 86.54, Recall Value 89.95, Specificity 89.60, F1 Score 88.21, AUC 91.01. In Data Splitting(35-65), Accuracy value is 91.81, Kappa Value is 88.67, Precision Value is 88.74, Recall Value is 90.14, Specificity Value is 91.36, F1 Score value is 89.43, AUC Value is 93.21. In Data Splitting(40-60), Accuracy

Value is 94.67, Kappa Value is 90.38, Precision Value is 90.92, Recall Value is 92.63, Specificity Value is 93.85, F1 Score Value is 91.76, AUC Value is 96.79. In Data Splitting(45-55), Accuracy Value is 96.37, Kappa Value is 91.97, Precision Value is 92, Recall Value is 93.21, Specificity Value is 94.04, F1 Score Value is 92.60, AUC Value is 98.04. In Data Splitting(50-50), Accuracy Value is 94.71, Kappa Value is 91.12, Precision Value is 91.37, Recall Value is 93.77, Specificity Value is 94.74, F1 Score Value is 92.55, AUC Value is 96.97. In Data Splitting(55-45), Accuracy Value is 96.9, Kappa Value is 93.97, Precision Value is 93.46, Recall Value is 95.64, Specificity Value is 96.13, F1 Score Value is 94.53, AUC Value is 98.42. In Data Splitting(60-40), Accuracy Value is 96.52, Kappa Value is 92.45, Precision Value is 95.38, Recall Value is 94.12, Specificity Value is 95.72, F1 Score Value is 94.74, AUC Value is 97.51. In Data Splitting(65-35), Accuracy Value is 97, Kappa Value is 94.37, Precision Value is 97.57, Recall Value is 96.42, Specificity Value is 97.25, F1 Score Value is 96.99, AUC Value is 98.78. In Data Splitting(70-30), Accuracy Value is 97.29, Kappa Value is 95.17, Precision Value is 98.94, Recall Value is 97.54, Specificity Value is 98.00, F1 Score Value is 98.23, AUC Value is 99.23. In Data Splitting(75-25), Accuracy Value is 97.11,Kappa Value is 95.55, Precision Value is 98.56, Recall Value is 97.02, Specificity Value is 97.98, F1 Score Value is 97.78, AUC Value is 99.00.

Table 4 and Table 5 show the results of the random forest technique when used in conjunction with different fold validation and data splitting validation methods. Table 4 and 5 shows the performance evaluation of the MRFE, and RF methods based on the compartmentalized data validation method. As the ranges of characteristics increased, the values of the measures decreased.

VII. CONCLUSION

Predicting crops for cultivation in agriculture is a difficult task. This paper has used a range of feature selection and classification techniques to predict yield size of plant cultivations. The results depict that an ensemble technique offers better prediction accuracy than the existing classification technique. Forecasting the area of cereals, potatoes and other energy crops can be used to plan the structure of their sowing, both on the farm and country scale. The use of modern forecasting techniques can bring measurable financial benefits.

REFERENCES

[1] R. Jahan, "Applying naive Bayes classification technique for classification of improved agricultural land soils," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 6, no. 5, pp. 189–193, May 2018.

[2] B. B. Sawicka and B. Krochmal-Marczak, "Biotic components influencing the yield and quality of potato tubers," *Herbalism*, vol. 1, no. 3, pp. 125–136, 2017.

[3] B. Sawicka, A. H. Noaema, and A. Gáowacka, "The predicting the size of the potato acreage as a raw material for bioethanol production," in *Alternative Energy Sources*, B. Zdunek, M. Olszáwka, Eds. Lublin, Poland: Wydawnictwo Naukowe TYGIEL, 2016, pp. 158–172.

[4] B. Sawicka, A. H. Noaema, T. S. Hameed, and B. Krochmal-Marczak, "Biotic and abiotic factors influencing on the environment and growth of plants," (in Polish), in *Proc. Bioróżnorodność Środowiska Znaczenie, Problemy, Wyzwania. Materiały Konferencyjne*, Puławy, May 2017. [Online]. Available: https://bookcrossing.pl/ksiazka/321192

[5] R. H. Myers, D. C. Montgomery, G. G. Vining, C. M. Borror, and S. M. Kowalski, "Response surface methodology: A retrospective and literature survey," *J. Qual. Technol.*, vol. 36, no. 1, pp. 53–77, Jan. 2004.

[6] D. K. Muriithi, "Application of response surface methodology for optimization of potato tuber yield," *Amer. J. Theor. Appl. Statist.*, vol. 4, no. 4, pp. 300–304, 2015, doi: 10.11648/j.ajtas.20150404.20.

[7] M. Marenych, O. Verevska, A. Kalinichenko, and M. Dacko, "Assessment of the impact of weather conditions on the yield of winter wheat in Ukraine in terms of regional," *Assoc. Agricult. Agribusiness Econ. Ann. Sci.*, vol. 16, no. 2, pp. 183–188, 2014.

[8] J. R. Olędzki, "The report on the state of remotesensing in Poland in 2011–2014," (in Polish), *Remote Sens. Environ.*, vol. 53, no. 2, pp. 113–174, 2015.

[9] K. Grabowska, A. Dymerska, K. Poáarska, and J. Grabowski, "Predicting of blue lupine yields based on the selected climate change scenarios," *Acta Agroph.*, vol. 23, no. 3, pp. 363–380, 2016.

[10] D. Li, Y. Miao, S. K. Gupta, C. J. Rosen, F. Yuan, C. Wang, L. Wang, and Y. Huang, "Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning," *Remote Sens.*, vol. 13, no. 16, p. 3322, Aug. 2021, doi: 10.3390/rs13163322.

[11] N. Chanamarn, K. Tamee, and P. Sittidech, "Stacking technique for academic achievement prediction," in *Proc. Int. Workshop Smart Info-Media Syst.*, 2016, pp. 14–17.

[12] W. Paja, K. Pancerz, and P. Grochowalski, "Generational feature elimination and some other ranking feature selection methods," in *Advances in Feature Selection for Data and Pattern Recognition*, vol. 138. Cham, Switzerland: Springer, 2018, pp. 97–112.

[13] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sens. Environ.*, vol. 118, pp. 259–272, Mar. 2012.

[14] S. K. Honawad, S. S. Chinchali, K. Pawar, and P. Deshpande, "Soil classification and suitable crop prediction," in *Proc. Nat. Conf. Comput. Biol., Commun., Data Anal.* 2017, pp. 25–29.

[15] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4559–4565.

[16] D. A. Reddy, B. Dadore, and A. Watekar, "Crop recommendation system to maximize crop yield in ramtek region using machine learning," *Int. J. Sci. Res. Sci. Technol.*, vol. 6, no. 1, pp. 485–489, Feb. 2019.

[17] N. Rale, R. Solanki, D. Bein, J. Andro-Vasko, and W. Bein, "Prediction of Crop Cultivation," in *Proc. 19th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Las Vegas, NV, USA, 2019, pp. 227–232.

[18] J. Jones, G. Hoogenboom, C. Porter, K. Boote, W. Batchelor, L. Hunt, P. Wilkens, U. Singh, A. Gijsman, and J. Ritchie, "The DSSAT cropping system model," *Eur. J. Agronomy*, vol. 18, nos. 3–4, pp. 235–265, 2003.

[19] M. T. N. Fernando, L. Zubair, T. S. G. Peiris, C. S. Ranasinghe, and J. Ratnasiri, "Economic value of climate variability impact on coconut production in Sri Lanka," in *Proc. AIACC Working Papers*, vol. 45, 2007, pp. 1–7.

[20] B. Ji, Y. Sun, S. Yang, and J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions," *J. Agricult. Sci.*, vol. 145, no. 3, pp. 249–261, Jun. 2007.

[21] C. Boryan, Z. Yang, R. Mueller, and M. Craig, "Monitoring U.S. agriculture: The U.S. department of agriculture, national agricultural statistics service, cropland data layer program," *Geocarto Int.*, vol. 26, no. 5, pp. 341–358, 2011.

[22] M. C. Hansen and T. R. Loveland, "A review of large area monitoring of land cover change using Landsat data," *Remote Sens. Environ.*, vol. 122, pp. 66–74, Jul. 2012.

[23] D. K. Bolton and M. A. Friedl, "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics," *Agricult. Forest Meteorol.*, vol. 173, pp. 74–84, May 2013.

[24] J. Dempewolf, B. Adusei, I. Becker-Reshef, M. Hansen, P. Potapov, A. Khan, and B. Barker, "Wheat yield forecasting for Punjab province from vegetation index time series and historic crop statistics," *Remote Sens.*, vol. 6, no. 10, pp. 9653–9675, Oct. 2014.

[25] H. D. Shannon and P. M. Raymond, "Managing weather and climate risk to agriculture in North America," *Central Amer. Caribbean*, vol. 10, pp. 50–56, Dec. 2015.

[26] E. Manjula and S. Djodiltachoumy, "A model for prediction of crop yield," *Int. J. Comput. Intell. Inform.*, vol. 6, no. 4, pp. 298–305, 2017.

[27] K. E. Eswari and L. Vinitha, "Crop yield prediction in Tamil Nadu using Baysian network," *Int. J. Intell. Adv. Res. Eng. Comput.*, vol. 6, no. 2, pp. 1571–1576, 2018.

[28] G. Shivnath and S. Koley, "Machine learning for soil fertility and plant nutrient management using back propagation neural networks," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. 2, pp. 292–297, 2014.

[29] M. Paul, S. K. Vishwakarma, and A. Verma, "Analysis of soil behaviour and prediction of crop yield using data mining approach," in *Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN)*, Dec. 2015, pp. 766–771.

[30] S. Pudumalar, E. Ramanujam, R. R. Harine, C. Kavya, T. Kiruthika, and J. Nisha, "Crop recommendation system for precision agriculture," in *Proc. 8th Int. Conf. Adv. Comput. (ICoAC)*, 2017, pp. 32–36.

[31] K. Bodake, R. Ghate, H. Doshi, P. Jadhav, and B. Tarle, "Soil-based fertilizer recommendation system using the Internet of Things," *MVP J. Eng. Sci*, vol. 1, pp. 13–19, 2018.

[32] K. Heupel, D. Spengler, and S. Itzerott, "A progressive crop-type classification using multitemporal remote sensing data and phenological information," *J. Photogramm., Remote Sens. Geoinf. Sci.*, vol. 86, pp. 53–69, Apr. 2018.

[33] M. Liu, T. Wang, A. K. Skidmore, and X. Liu, "Heavy metal-induced stress in rice crops detected using multi-temporal Sentinel-2 satellite images," *Sci. Total Environ.*, vols. 637–638, pp. 18–29, Oct. 2018.

[34] P. Priya, U. Muthaiah, and M. M. Balamurugan, "Predicting yield of the crop using a machine learning algorithm," *Int. J. Eng. Sci. Res. Technol.*, vol. 7, pp. 1–7, Apr. 2018.

[35] A. Chougule, V. Kumar, and D. Mukhopadhyay, "Crop suitability and fertilizer recommendation using data mining techniques," in *Progress in Advanced Computing and Intelligent Engineering* (Advances in Intelligent Systems and Computing), vol. 714. Singapore: Springer, 2019, pp. 205–213.

[36] C. Brogi, J. A. Huisman, S. Pätzold, C. von Hebel, L. Weihermüller, M. S. Kaufmann, J. van der Kruk, and H. Vereecken, "Large-scale soil mapping using multi-configuration EMI and supervised image classification," *Geoderma*, vol. 335, pp. 133–148, Feb. 2019.

[37] G. Mariammal, A. Suruliandi, S. P. Raja, and E. Poongothai, "Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 5, pp. 1132–1142, Oct. 2021.

[38] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, pp. 18–22, Apr. 2002.

[39] A. Al-Naji, A. B. Fakhri, S. K. Gharghan, and J. Chahl, "Soil color analysis based on a RGB camera and an artificial neural network towards smart irrigation: A pilot study," *Heliyon*, vol. 7, no. 1, Jan. 2021, Art. no. e06078.

[40] K. Kashyap. (2019). *Machine Learning Decision Trees and Random Forest Classifiers*. [Online]. Available: https://medium.com/analytics-vidhya/machine-learning-decision-trees-and-random-forest-classifiers-81422887a544

S. P. RAJA was born in Sathankulam, Tuticorin, Tamil Nadu, India. He received the B.Tech. degree in information technology from the Dr. Sivanthi Aditanar College of Engineering, Tiruchendur, in 2007, and the M.E. degree in computer science and engineering and the Ph.D. degree in image processing from Manonmaniam Sundaranar University, Tirunelveli, in 2010 and 2016, respectively. He completed his schooling at the Sacred Heart Higher Secondary School, Sathankulam. He is currently working as an Associate Professor with the School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu. He has published 46 papers in international journals, 24 in international conferences, and 12 in national conferences. He is an Associate Editor of the *International Journal of Interactive Multimedia and Artificial Intelligence*, *Brazilian Archives of Biology and Technology*, *Journal of Circuits, Systems and Computers*, *Computing and Informatics*, *KSII Transactions on Internet and Information Systems*, the *International Journal of Wavelets, Multiresolution and Information Processing*, the *International Journal of Image and Graphics*, and the *International Journal of Biometrics*.

BARBARA SAWICKA has been a Full Professor with the Faculty of Agrobioengineering, University of Life Sciences in Lublin, Poland, since 1999. Her writings and creative works contain 998 publications, including 150 monographs and chapter of monographs, 380 original research papers, and 15 books. She holds two patents. Her teaching activity includes 11 Ph.D. students promoted, over 190 M.Sc. students promoted, and 250 promoted engineers. She has reviewed 40 Ph.D. dissertations, 25 postdoctoral works, 12 applications for the Title of Professor, and over 400 scientific papers. Her research interests include agronomy, agribusiness and rural development, bioengineering, climate change, commodity products, environmental protection, food sciences, and food safety. She has received scholarship from the Institute of Potato in Młochów, Institute of Potato in Jadwisin, Poland; Onderzoek- en Voorlichtingscentrum voor Land- en Tuinbouw, Roeselare, Belgium; the University of Agriculture Kaunas, Lithuania; Aleksandras Stulginskis University, Lithuania; Bingöl University, Turkey; and the Lithuanian Center for Agriculture and Forestry Sciences, Kėdainiai, Lithuania.

ZORAN STAMENKOVIC (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from the University of Niš, Serbia, in 1995. He is currently a Scientist at IHP GmbH, Frankfurt (Oder), Germany. He has published more than 150 scientific book chapters, theses, journal articles, and conference papers. He has given more than 25 invited talks in the field of design and test of integrated circuits and systems. He is the Lead Editor (and a coauthor) of the book *Silicon Systems for Wireless LAN*. His research interests include SOC design for wireless communications, fault-tolerant circuits and systems, and integrated circuit yield and reliability modeling. He serves as a Program Committee Member for many scientific conferences, among them are DDECS, IOLTS, EWDTS, DTIS, MIEL, and TELFOR. He was the General Chair of DDECS15 and the Program Chair of DDECS18 and DDECS20. He is a Regional Editor of the *Journal of Circuits, Systems, and Computers*.

G. MARIAMMAL received the B.E. degree in computer science and engineering from the Francis Xavier Engineering College, Tirunelveli, India, in 2011, and the M.E. degree in computer science and engineering from Manonmaniam Sundaranar University, Tirunelveli, in 2017, where she is currently pursuing the Ph.D. degree in computer science and engineering. Her research interests include machine learning, data analytics, and image processing.