JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 18, No. 3, September 2024, pp. 152-168

Ensemble Learning Based on Feature Selection and Distance Normalization for Enhancing Corn and Weed Classification

Faisal Dharma Adhinata
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
Faculty of Informatics, Institut Teknologi Telkom Purwokerto, Jawa Tengah, Indonesia
faisaldharma@mail.ugm.ac.id

Wahyono* and Raden Sumiharto
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
wahyo@ugm.ac.id, r_sumiharto@ugm.ac.id

Abstract
Weeds need to be removed from the immediate areas surrounding crops as they compete for soil nutrients. Farmers currently clear weeds manually, which is both tiring and imprecise. Therefore, researchers have developed artificial intelligence (AI) using deep learning or non-handcrafted methods to facilitate precise detection. However, these methods have yet to achieve real-time inference speeds. Consequently, this study adopts a handcrafted approach that employs visual leaf features for classification via ensemble learning. The objective is to employ feature selection and data normalization to create an accurate and efficient machine-learning model. The experimental findings obtained in this work demonstrate that Information Gain effectively reduces features by 50%, from 22 to 11, while maintaining accuracy. Chebyshev normalization emerges as the most suitable normalization technique, as it significantly enhances classification accuracy in ensemble learning. The accuracy obtained when using histogram gradient boosting is found to be 0.92 with an inference time of 5.955 ms per image. These outcomes illustrate that handcrafted features achieve higher accuracy than non-handcrafted methods, ultimately improving efficiency and enabling real-time implementation.

Category: Artificial Intelligence

Keywords: Handcrafted; Ensemble learning; Information gain; Chebyshev normalization; Weed

I. INTRODUCTION

Artificial intelligence (AI) is a critical tool in modern agriculture, as it is indispensable in accurately identifying weeds in corn fields. Weed management must be completed in a timely fashion since weeds are likely to significantly reduce crop yields by competing with crops for nutrients, light, and space [1]. The integration of AI in processing agricultural data through images is one way that weed and crop identification approaches are being modernized. AI is trained on such data to recognize patterns and make accurate and fast decisions regarding weed and crop identification [2, 3]. Further, the use of AI in corn and weed identification significantly reduces the application of chemical herbicides, lowers production costs, and promotes environmentally friendly and sustainable practices

Open Access http://dx.doi.org/10.5626/JCSE.2024.18.3.152 http://jcse.kiise.org

Received 29 March 2024; **Accepted** 9 September 2024
*Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

[4]. Overall, existing studies suggest that AI will find increasing use in agriculture research and development as well as real-world applications.

One application of AI for identification involves the use of deep learning or non-handcrafted features. However, using deep learning in the context of corn and weed classification poses significant challenges related to data processing speed [5]. Although deep learning techniques are known for their ability to produce models with very high accuracy [6], their architectural complexity and large computational requirements often make them impractical for field applications [7], particularly in areas with limited computing resource availability. This is a problem because a fast detection speed is required for weed control with real-time detection needs. Therefore, the present research focuses on developing lighter and faster methods using machine learning approaches; this is why the features are handcrafted. Using this approach, it is possible to achieve a balance between high detection accuracy and processing time efficiency that facilitates effective application in the field.

There are several options among machine learning methods. Ensemble machine learning has various advantages over traditional machine learning, which strengthens its predictive performance in weed detection [8-10]. Ensemble learning is a technique that involves combining the predictions of numerous models to generate more accurate and robust output. This approach often outperforms the capabilities of individual models. This method also effectively reduces variance and bias [11], which are the two primary sources of error in machine learning, by integrating the strengths of various models as well as compensating for their weaknesses. In an ensemble, it is possible to minimize the overfitting tendencies of complex models because errors from one model tend to be compensated for by other models in the ensemble [12]. Therefore, the current research will use an ensemble machine-learning approach.

One of the main difficulties in applying machine learning with handcrafted features is selecting the features that are most effective and efficient. This important task is due to the complexity of image data, including the considerable variation in shape, texture, and color that is seen in weeds. Feature selection is a crucial step of handcrafted approaches, as it allows for the enhancement of machine learning accuracy and the minimization of deficiencies in ensemble learning [13, 14]. The key aspect is determining and selecting the most relevant features that should guide the decision-making process. Such a process helps decrease the data's complexity and facilitates the training time.

This process also minimizes the risk of overfitting, thus allowing for noise or irrelevant features to be excluded, which may lead to a case in which the machine learning model cannot detect patterns in the feature set learning [15]. In ensemble learning, which is characterized by the merging of several models to ultimately make more precise predictions, it is essential to use feature selection to improve the quality and effectiveness of such combinations [16]. This process also enhances the performance of each model in the ensemble, as it only involves the use of high-quality features. As a result, models may focus on the most critical aspects of data to help recognize patterns accurately. Efficient feature selection contributes to the performance improvement that can be achieved by ensemble learning by ensuring that only the most relevant and effective information is used for decision-making. As a result, systems can make predictions with greater precision.

The features that are selected through feature selection still contain numbers with varying scales, because they use handcrafted color, shape, and texture features. The obtained features may be biased at certain feature scales [17], so it is necessary to apply data normalization. This process helps reduce distortion caused by features with a wide range, thus allowing the algorithm to focus more on important patterns rather than differences in scale. Normalization also improves the numerical stability of the algorithm and can speed up the training process [18]. One data normalization method is distance normalization. Using the distance normalization techniques on handcrafted features, specifically within machine learning applications such as corn and weed classification, offers distinct advantages over other normalization methods. For example, Wang et al. [19] used distance normalization before initiating the training stage using machine learning. Distance normalization has a very significant effect on accuracy, and it also reduces model complexity. Therefore, this research adopts distance normalization by evaluating several distance algorithms.

The main objective of this research is to apply feature selection and distance normalization on handcrafted features to ultimately build an accurate and efficient model using ensemble learning. It is hoped that this handcrafted approach can obtain a model that is not excessively complex, such that it is more efficient when carrying out the model inference process. Distance normalization can also increase the efficiency of the training stage. The contributions of this research can be broadly described as follows:

- It explores the use of distance algorithms for data normalization to improve the accuracy and efficiency of the training stage.
- It explores feature selection methods to search for relevant features.
- It evaluates the different ensemble learning methods in comparison with traditional machine learning methods.

The rest of this paper contains the following sections: Section II is a literature review of feature selection and normalization methods. Section III is a research methodology that explains data input in terms of feature selection and

data normalization. Section IV discusses the results and relevant insights from the research experiment. Finally, Section V concludes this work.

II. LITERATURE REVIEW

Machine learning and computer vision techniques have increasingly been implemented in precision agriculture with a special emphasis on weed detection. This represents a task of paramount importance, given that it can help improve crop management techniques, reduce herbicide application, and improve crop productivity. One of the approaches to consider is the utilization of manually designed features as well as generated and handcrafted data features from the images to retrieve important information regarding the visual properties of weeds regarding crops. However, in most cases, there are two crucial preprocessing steps that should be considered to achieve the best results with models that use manually designed features: data normalization and feature selection.

A. Feature Selection

The performance and significance of feature selection in different domains, such as machine learning, data mining, and pattern recognition, have represented some of the critical and widely discussed topics in recent years. The available literature suggests that, among the numerous methods that are available, Information Gain is superior in multiple contexts. In particular, Kilic et al. [20] underscore the method's effectiveness in typical feature selection techniques when compared to traditional methods such as gain ratio, chi-square, and ReliefF. Further, in their comparative study of correlation feature selection using the bat algorithm (CFS-BA), Lehavi and Kim [21] found that the model-building efficiency and accuracy of CFS-BA are significantly high, and that they can closely match the best random forest Information Gain.

This led to Information Gain being used for feature selection to predict the performance of tumor detection processes on microarray data [22]. The inclusion of the described analysis method, as part of grey wolf optimization and search, was shown to reduce the computational load and enhancement of prediction accuracy, which was critical to improving tumor detection. The Information Gain method has also been shown to enhance the accuracy of developed predictive models in the context of epileptic seizure classification [23]. Finally, in a study focusing on the development of household big data, Nuanmeesri [24] highlights that the feature selection method has superior efficacy when compared to other methods that are used to analyze errors in the data.

Overall, the results of the described studies indicate that the application of Information Gain for feature selection has a significant effect in multiple research fields, thereby enhancing model accuracy and reducing computational load. At the same time, when implementing the method, it is critical to test the accuracy of the model for each level of feature ranking, as is done in other information-based approaches, and it is also important to make feature selection decisions that are based on extracting attribute information from gain measurements while considering the proper interaction between features; therefore, it is necessary to use an iterative evaluation to achieve the optimal balance between model complexity and performance efficiency.

B. Data Normalization

Recent studies using pre-processing methods for data normalization have been able to prove the significance of this procedure in enhancing the performance of machine learning models. The normalization methods that were most commonly employed were min-max scaler, z-score, MaxAbsScaler, and RobustScaler. Each of these approaches has specific usages and benefits for tackling issues like outliers and dissimilar feature scales.

Min-max scaling to normalize all features to the same range has been reported to result in compression of the feature distribution in the presence of outliers [25]. Silva et al. [26] employ z-score normalization as a method to normalize the features to a mean of zero and a standard deviation of one, thus reducing the impact of outliers. Zhang et al. [27] use another normalization type, RobustScaler, to address the problems with outliers that could not be solved by min-max scaling. This method is more suitable for datasets in which non-uniform data dominates or extreme outlier datasets.

In this context, distance normalization is a new concern. Lafuente et al. [28], explained the promising potential of distance normalization in maximizing model performance. Distance normalization tunes the data by processing the data based on the point distances. The authors claim that, through this process, the scaling of a feature based on distance adjustment can better reflect the intrinsic nature of the data. Distance normalization highly accents the norm on distance, which is shown as an alternative distance ratio from the starting point. Their results in recent years indicate that these approaches to processing data structure are useful for catering to cost and benefit preferences [29]. The critical feature of distance normalization is that the relationships of distances from the normalized point remain. As a result, the model processes this information to learn more accurately and generalize the patterns.

In summary, the use of a different number of normalizations returns different solutions for data pre-processing. One of the possible approaches, distance normalization, shows promising potential for boosting machine learning's performance in obtaining distance information. The current study suggests that this is in some concerns the

superior model, particularly when the intrinsic search of the data structure is the primary stage. However, the question here is what distance is optimal for the weed problem when only three handcraft features are applied: color, shape, and texture. This is a crucial consideration to evaluate to obtain the best machine learning accuracy.

III. RESEARCH METHODOLOGY

The classification step demands data acquisition in the form of images of weed and corn seeds. The first pre-processing step for uniformity across all data points involves resizing the images. Secondly, a Gaussian blur is applied to the images to reduce noise in the image and the detail present. Finally, HSV color space is used, which is appropriate for image segmentation. Segmenting the data to identify the area of leaves involves color thresholding for green, followed by morphological operations to clean up the regions and then masking these regions. Handcrafted features are then extracted in the region. Feature selection experiments are conducted in this research to decide on the most informative attributes, and the data values are normalized to obtain uniformity in the feature scale. After that, the data that have been pre-processed and optimized for features are split into several training and test sets. The experiment concludes with a final step of measuring the system's performance with evaluation metrics applied to the ensemble learning algorithm used for classification. Fig. 1 depicts the flow of this research.

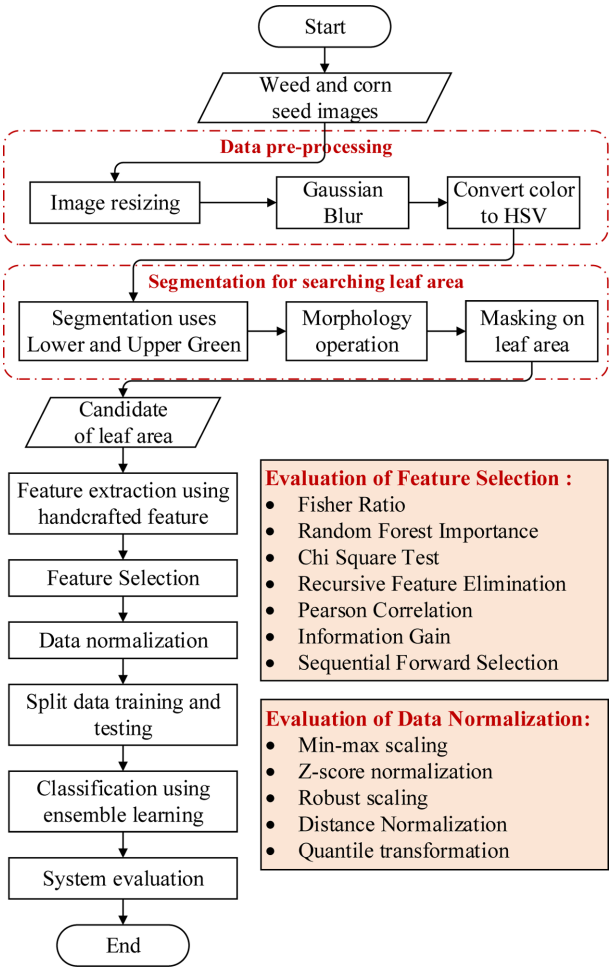

Fig. 1. The flow of corn-weed classification.

A. Data Acquisition

This research utilized a dataset that Jiang et al. [6] compiled from a corn plant nursery in 2016 using a Canon PowerShot SX600 HS camera. The images were taken from a vertical perspective above the plants. This dataset comprises five distinct classes: *Cirsium setosum*, *Chenopodium album*, bluegrass, sedge, and corn. Each image has a resolution of 800 × 600 pixels and is in full RGB color. In this research, the data configuration is such that 80% of the data is used for training and 20% is used for testing. Table 1 presents examples from the dataset and the proportion of each class.

B. Data Pre-processing

In this study, the initial data preprocessing stage involves resizing the dataset images to a uniform dimension of

Table 1. Example data

Example of data	Training data	Testing data
 *Cirsium setosum*	960	240
 *Chenopodium album*	960	240
 Bluegrass	960	240
 Sedge	958	240
 Corn	960	240

(a) (b) (c) (d) (e)

Fig. 2. Example segmentation result: (a) *Cirsium setosum*, (b) *Chenopodium album*, (c) bluegrass, (d) sedge, and (e) corn.

200 × 200 pixels. This standardization is critical for consistency in the scale of the leaf objects due to the use of handcrafted features. After resizing, a Gaussian blur is applied to diminish image noise. The final step in preprocessing involves transforming the color model from RGB to HSV, which is instrumental for the subsequent data segmentation process, particularly the green hues in the HSV color space.

C. Leaf Segmentation

The segmentation phase is conducted to distinguish the leaf areas from the surrounding soil or other objects—this study's feature extraction is only in the leaf regions, as it utilizes handcrafted features. The segmentation leverages a color spectrum ranging from light to dark green, specifically within the range from (20; 20; 50) to (80; 255; 240). After the isolation within the green spectrum, any existing holes in the leaf structures are sealed using morphological operations. The culmination of this segmentation process involves masking the original image, ultimately allowing the RGB color of the leaf areas to be utilized during the feature extraction phase. An illustrative example of the result of segmentation is provided in Fig. 2.

D. Feature Extraction

In machine learning for agricultural applications, particularly in corn and weed classification, it is crucial to extract handcrafted features to distinguish between weeds and crops based on their visual characteristics. This research uses handcrafted features that are categorized into three groups: color [30], shape [31, 32], and texture [32], each offering unique insights into the physical attributes of the plants that can be observed through images. The color features in this research are the mean and standard deviation of RGB color. The shape features are solidity, eccentricity, circularity, compactness, and Hu moments. The texture features are angular in the second moment, contrast, inverse different moments, entropy, and correlation.

Color features are essential for identifying the plant species [33], and they are calculated using the mean and standard deviation of RGB color channels. The mean color value measures the intensity of a particular color in an image or segmented region. The standard deviation reveals the color variance, which indicates the possible colors in the region of interest. These colors are important in corn and weed classification, as weeds have different colors than corn.

The shape features are related to the leaf geometry, which is described using solidity, eccentricity, circularity, compactness, and Hu moments. Solidity indicates the filling degree of a region, and it therefore indicates the extent of compression in the leaf structures. Eccentricity measures the elongation of a region, which can be used to differentiate corn from weeds. Circularity and compactness represent how close a region appears to be to a circle, which helps classify leaves with similar shapes. Hu moments are invariant descriptors that provide a global description of leaf shape such that they can discriminate geometries in the presence of distortions in orientation, scale, and position. These features all facilitate the identification of weeds and corn based on their geometry [34].

Texture features analyze the patterns and structural differences in the surface of objects [35]. These include angular second moment, contrast, inverse difference moment, entropy, and correlation. The angular second moment reveals the texture of an object as being smooth or rough. The contrast is the color or luminance level difference, which can make an object unique. The inverse difference moment describes the texture homogeneity, while entropy is the texture complexity manifesting roughness in the surface. Correlation reflects the linearity of pixel traces and indicates periodicity in the trace mark. Texture features are useful in corn and weed classification because they show the morphology of the surface of the leaf and the stem, which can help identify corn and weed.

Machine learning models can differentiate between corn and weeds because of the precise computation and analysis of these features. This accurate knowledge assists in the establishment of exact classification systems to enhance weed management and corn productivity because of the precise creation of the classification model. Therefore, the total number of features is 22, and

the analysis must be performed with a feature selection to acquire the most significant features for the optimal accuracy of the machine model.

F. Feature Section

Information Gain is used to select relevant handcrafted features. In corn and weed classification, the most important features are determined to include color, shape, and texture attributes, which can help in the differentiation of crop and weed instances. Selecting the most useful features helps reduce the model's complexity and improve its interpretability and performance. Information Gain is one of the most used feature selection methods.

Feature selection is done using entropy values. In this method, the features will be ranked based on their values [36]. The largest value of the ranking is the most relevant feature, and the relationship with the related dataset is strong. The method will rank the features based on the entropy value that is found using one of the classes before and after observing the same data. The first phase consists of finding the value of the entropy using Eq. (1):

$$Ent(S) = \sum_{i=1}^{m} -P_i . log_2 P_i, \tag{1}$$

where S is the sample (record or total), m is the maximum number of features, and P_i is the ratio of the number of samples in class i to the total sample. Then, the entropy value after weighing each feature is calculated using Eq. (2):

$$Ent_F(S) = \sum_{j=1}^{n} -P_j . Ent(S_j), \tag{2}$$

where n is the number of values in the class, P_j is the ratio of the total sample j to the total samples in the feature, and $Ent(S_j)$ is the entropy value for sample j. The Information Gain value can be obtained by reducing Eqs. (1) and (2), as shown in Eq. (3):

$$Gain(S, F) = Ent(S) - |Ent_F(S)|. \tag{3}$$

The present research incorporates a comprehensive approach to feature selection by evaluating various methods, alongside Information Gain, to identify the most relevant handcrafted features for corn and weed classification. We will conduct a comparative analysis using different feature selection methods, namely Fisher ratio, random forest importance, chi-square test, recursive feature elimination (RFE), Pearson correlation, and sequential forward selection (SFS).

1) Fisher ratio: This approach assesses the discriminatory power of each feature by using the Fisher criterion [37]. This criterion computes the variance ratio between classes to the variance within subclasses for each feature. A higher Fisher score suggests that a trait has better discriminatory power for discriminating between different classes. This capability is particularly useful when the features are both normally distributed.
2) Random forest importance: This is a technique that measures the importance of the features based on the Random Forest algorithm, one of the popular ensemble learning algorithms [38]. In random forest, one tree is formed using one bootstrap sample of the data. The importance of features is given by the decrease in accuracy in the out-of-bag samples that occurs after excluding a feature. It is robust enough to overfit and good at handling high-dimensional data.
3) Chi-square test: This is a statistical technique that is used to determine whether there is a significant association between two categorical variables [39]. It is calculated between each feature and the target within the feature selection process. A high value implies that the features are unlikely to be independent of class. It is mostly used on solely positive data.
4) Recursive feature elimination: In this backward selection process, the least important features are recursively discarded based on the significance given by the model's importance scores or coefficients [40]. This process has many applications, like in support vector machines, which rank features based on their effect on the model's performance.
5) Pearson correlation: This measures the linear correlation between two continuous variables, which—in feature selection—can be applied between each feature and the target variable [41]. Features with high absolute correlation values are considered to be more relevant. However, it only captures linear relationships and might miss nonlinear dependencies.
6) Sequential forward selection is a wrapper feature selection methodology wherein a vacant model is initially utilized. The features that yield the most substantial enhancement in model performance are added in order of importance until the addition ceases to improve performance substantially [42]. Although it is computationally demanding, this method has the potential to produce feature subsets that are highly optimized.

By evaluating these feature selection methods in conjunction with Information Gain, the research aims to systematically identify the most predictive handcrafted features for corn and weed classification, in the process enhancing the effectiveness of the classification model used in this domain. Each method brings a unique perspective on feature relevance, ultimately allowing for

a more nuanced understanding of feature importance in the context of agricultural image analysis.

F. Data Normalization

After using feature selection to select the most relevant features for a machine learning task, it is crucial to normalize these features to ensure that the model treats them equally during training and prediction. Normalization is a procedure that involves adjusting the scales of the features so that they have comparable ranges and distributions [43]. This process is essential since most features vary widely in magnitudes, units, or ranges. The reason for the significance of this step is that features with larger numerical values can dominate those with smaller ones, ultimately leading to inaccurate predictions. The results of different features might also be biased based on the scale of the features in consideration. Notably, distance normalization is a common type of data normalization [19]. This is because, in many cases, the distance between data points is used to make decisions based on the characteristics of data points. For example, in the case of corn and weed classification, handcrafted features could include measurements that are related to the color, shape, and texture of plants, among others. The process of normalization guarantees that the distance calculations weigh each feature in such a manner that ensures that one category of features does not dominate the rest due to its scale. In this research, the distance normalizations that are considered include Euclidean, Manhattan, Minkowski, and Chebyshev.

1) Euclidean normalization: This transforms the input vector such that its length (or its Euclidean norm) reaches unity. Mathematically, the Euclidean normalization of vector x is vector x', where each element x_i' is calculated using Eq. (4):

$$x_i' = \frac{x_i}{\sqrt{\sum_j x_j^2}}. \tag{4}$$

2) Manhattan normalization: This transforms the input vector so that the absolute number of its elements is one. Eq. (5) calculates each element x_i' of the vector x'.

$$x_i' = \frac{x_i}{\Sigma|x_j|}. \tag{5}$$

3) Minkowski normalization: This is a generalization of Manhattan and Euclidean normalization, where the Minkowski norm of a vector is defined by Eq. (6):

$$x_i' = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \tag{6}$$

where p is a parameter that determines the normalization order. $p = 1$ corresponds to Manhattan normalization, and $p = 2$ corresponds to Euclidean normalization. Meanwhile, Minkowski normalization uses $p = 3$.

4) Chebyshev normalization adjusts the input vector by measuring the largest distance from zero among all vector elements. In the context of normalization, this involves dividing each vector element by the largest absolute value of all its elements, as shown in Eq. (7):

$$x_i' = \frac{x_i}{max_{i=1}^{n}|x_j|}, \tag{7}$$

where $max_{i=1}^{n}|x_i|$ is the maximum of the absolute values of the elements in vector x_j.

This research will also evaluate several other normalization methods: min-max scaling, z-score normalization, robust scaling, and quantile transformation. Evaluating these methods will help determine the most effective approach for normalizing handcrafted features for corn and weed classification, in turn potentially improving the accuracy and robustness of models.

G. Ensemble Machine Learning

The classification of corn and weed species is a critical task in precision agriculture. Ensemble learning techniques, such as histogram gradient boosting (HistGBoost), LightGBM, and random forest, are currently the most used methods in tackling this difficulty because of their resilience and precision. HistGBoost enhances performance by efficiently handling continuous features, while LightGBM offers speed and effectiveness. Random forest contributes to the ensemble's diversity, offering multiple deep decision trees that prevent overfitting.

1) The HistGBoost classifier is an advanced ensemble machine learning algorithm that builds upon the principles of gradient boosting by using decision trees as the base learners [44]. Fig. 3 depicts the HistGBoost method. This method begins with continuous handcrafted features and custom-designed numerical input variables. These continuous features are then subjected to a discretization process in which they are converted into a finite number of intervals or bins, and where they are ultimately transformed into discretized features. This discretization is a form of feature engineering that simplifies the data without losing the essence of the information it conveys. Following this discretization, feature histograms are created for each feature. These histograms are graphical representations showing the frequency distribution of the data across the defined bins. In machine learning, particularly in the HistGBoost algorithm, these histograms allow for efficient calculation of the best splitting points in decision trees by quantifying the distribution of data

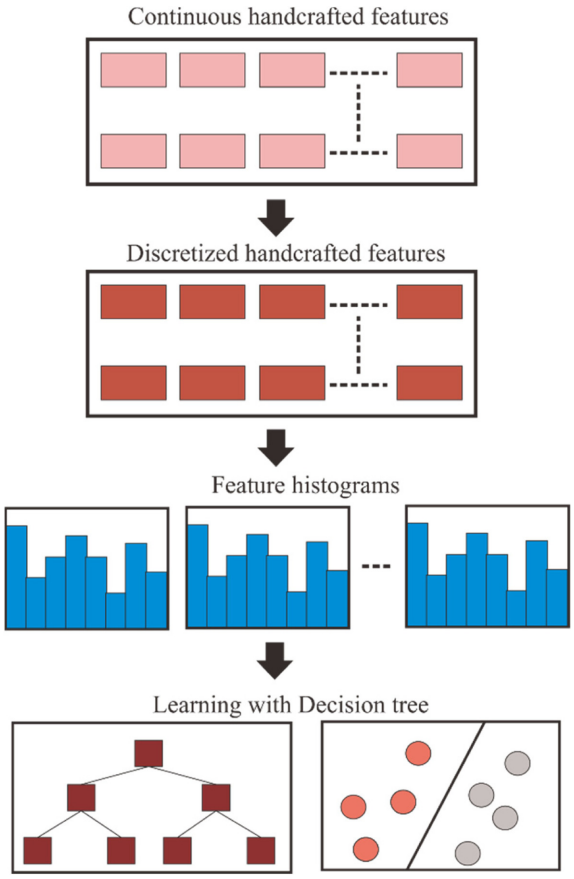

Fig. 3. Illustration of histogram gradient boosting.

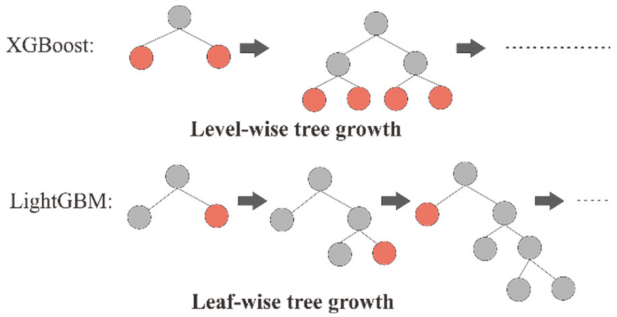

Fig. 4. Difference between XGBoost and LightGBM.

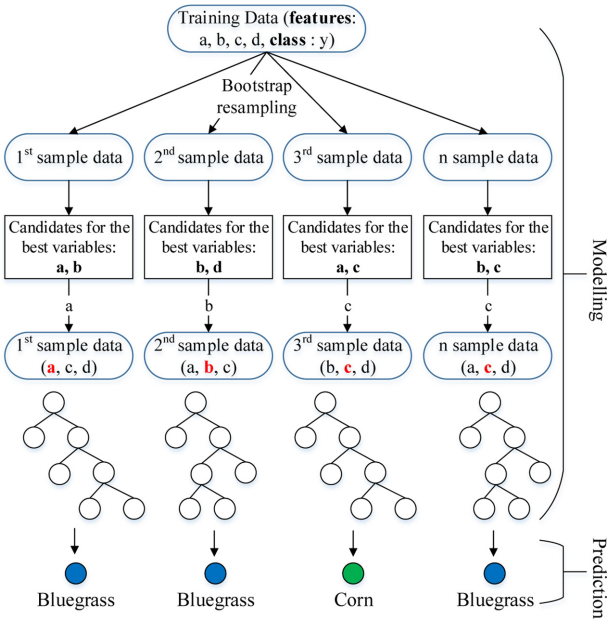

Fig. 5. Illustration of random forest.

points within each feature. In the final step, the decision tree is employed to train the data. A decision tree is a model consisting of branches and nodes, which represent decisions that are made based on the input attributes.

2) LightGBM, or light gradient boosting machine, is a distributed, rapid gradient boosting framework that is based on decision tree algorithms [45]. It stands out due to its exceptional efficacy and speed, which it achieves through two innovative approaches: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). GOSS allows LightGBM to focus more on informative instances while reducing less important ones, thus maintaining accuracy while accelerating computation. EFB efficiently combines mutually exclusive features, significantly decreasing dimensionality without compromising performance. Fig. 4 shows a comparison between XGBoost and LightGBM. XGBoost constructs trees using a level-wise strategy where all nodes at a given depth are evaluated and expanded before progressing to the next, ensuring that trees are balanced but at the cost of reduced efficiency, as it involves evaluating more potential splits, even if it is ultimately better for reducing loss.

3) By contrast, LightGBM utilizes a leaf-wise growth strategy that involves expanding only the most promising node at each iteration, regardless of its depth in the tree. This can lead to deeper but more unbalanced trees. This method is more computationally efficient as it allows the algorithm to converge faster by focusing on the splits that provide the most significant gains in terms of loss reduction.

4) Random forest: This is an ensemble learning method for classification and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class, which is the mode of the classes (classification) of the individual trees [46]. Fig. 5 illustrates the flow of the random forest. The procedure commences with the initial training dataset, which comprises several features (a, b, c, d) along with the label of the target class (y). The

random forest algorithm generates multiple subsets of this data through bootstrap resampling, allowing for the creation of different data samples with replacement, which means that some instances can appear more than once within the same subset or across multiple subsets. For each of these bootstrap samples, a decision tree is grown. However, to ensure that each of these trees is unique, the algorithm selects a random subset of candidate features for the best split at each node rather than considering all features. This step infuses randomness into the model and helps de-correlate the trees in the forest. For instance, the first tree considers features "a" and "b" for the best split, while the second tree might consider features "b" and "d." Each tree is constructed independently and fully grown to its maximum size without pruning, ultimately ending as a strong learner within the ensemble. When making predictions, each tree casts a vote for a class, and the final prediction for a new sample is determined as a result of a majority vote from all trees in the forest. The diagram shows individual predictions of different trees, which are represented by blue and green dots for classes such as bluegrass and corn.

H. System Evaluation

In the research of corn and weed classification, a comprehensive evaluation will be used to focus on the accuracy and inference time of the model. Accuracy metrics will be measured using the sklearn library in Python, which is a standard tool for such tasks in machine learning. The model performance will be evaluated using various normalization techniques, including min-max scaling, z-score normalization, robust scaling, quantile transformation, and Euclidean normalization. Feature selection will also be executed using different methods, such as Fisher ratio, random forest importance, chi-square test, RFE, Pearson correlation, and SFS.

The evaluation will examine multiple facets of this research. First, the impact of the feature selection level will be explored by using datasets with varying numbers of selected features, starting from 75% of the total available data as determined by the optimal selection method. Distance normalization algorithms, including Euclidean, Manhattan, Minkowski, and Chebyshev, will be experimented with to determine their effect on accuracy. Ensemble model configurations and prediction speeds will be inspected to optimum accuracy and inference time. This research will use multi-step evaluation to classify corn and weed. A combination of normalization techniques, feature selection, and learning parameters will be evaluated. The aim is to enhance both accuracy and inference time performance to contribute to precision farming applications.

IV. RESULTS AND DISCUSSION

A. System Configuration

For corn and weed classification, this research utilizes a workstation featuring a 6-core CPU and 8 GB of RAM within a Core i5 processor for robust data processing and modelling. This research also utilizes Python 3 programming language. To elaborate, this research uses the following libraries: the numpy library is used for array manipulation and numerical operations, the sklearn or sci-kit-learn library is used for machine learning tasks, and the time library is used to determine how long it will take the model to train and infer. Light gum is also used to implement the LightGBM algorithm. These libraries are vital for this research, as they enable the development and evaluation of a classification model.

B. Result of Feature Selection using Information Gain

The outcome of the feature selection using Information Gain has provided valuable insights for corn and weed classification. The data has been feature-selected using Information Gain, which uses the 10 best features based on their ranking [47]. Based on the scores obtained in this way, the top 10 features are found to be the most informative in distinguishing between corn and weed classes. Fig. 6 shows the results of feature selection using Information Gain.

Correlation has the highest Information Gain score, indicating its strong discriminative power in the classification task. It is followed in order by GreenVar and GreenMean, which suggest that variations and average values in the green color channel are crucial for classification. RedMean and BlueMean are also identified as necessary, thus highlighting the significance of mean color values in red and blue channels. Additional features that made it into the top 10 are entropy, which assesses the randomness and, consequently, the textural information

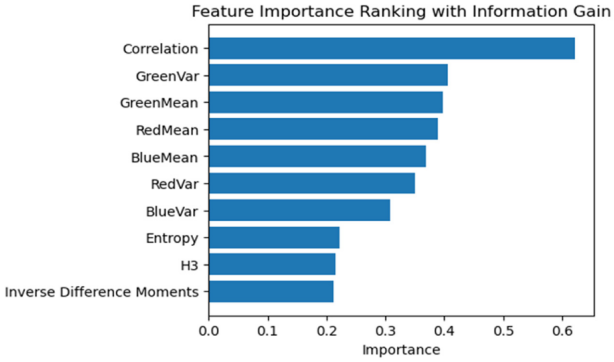

Fig. 6. Results of feature selection.

in the images, and RedVar and BlueVar, which indicate the fluctuations in the red and blue channels, respectively. The Hu moments, specifically the H3 and inverse difference moments, are also found to be statistically significant, thus providing information on shape and texture that is essential in distinguishing between different plant species. The scores reflect the relative importance of these features, with the chosen top 10 offering the most valuable information for the classifier. These features are expected to be instrumental in training accurate and robust machine learning models for the task at hand, with the ultimate aim of achieving high performance in separating corn plants from weeds using handcrafted features.

C. Effect of Data Normalization on Ensemble Learning

The results from the corn and weed classification study compare accuracy across three different ensemble learning models: random forest, HistGBoost, and LightGBM. These models were evaluated using various data normalization techniques to determine their impact on classification accuracy. Table 2 lists the effects of data normalization on ensemble learning accuracy.

The accuracy of classification studies employing ensemble learning algorithms, such as random forest, HistGBoost, and LightGBM, exhibits substantial variation based on the data normalization technique used. The accuracy results without data normalization using the random forest method are 0.84, HistGBoost of 0.873, and LightGBM of 0.871. Using min-max scaling did not significantly improve random forest accuracy but did result in minimal improvements for LightGBM. Using z-score normalization showed a slight increase in accuracy, namely in random forest of 0.842 and HistGBoost of 0.874, but in LightGBM, it decreased to 0.867. Robust scaling provides consistent results for random forest along with mild improvements for LightGBM. The quantile transformation method provides almost the same results without normalization, indicating that this technique does not significantly change the accuracy of the models tested.

Taken together, these results indicate that implementing data normalization using Euclidean normalization produces the best accuracy in all models. Random forest improved to 0.887, HistGBoost peaked at 0.905, and LightGBM significantly improved to 0.897. This confirms that selecting an appropriate normalization method is critical in maximizing the model's potential in classification tasks. Therefore, the following experiment uses Euclidean normalization for pre-processing.

D. Effect of Feature Selection on Ensemble Learning

The Fisher ratio, random forest importance, chi-square test, RFE, Pearson correlation, SFS, and Information Gain feature selection methods were evaluated in testing this feature selection method. The number of features to be taken was 10. The data normalization used was Euclidean normalization. Table 3 lists the results of the influence of the various feature selection methods used.

The results demonstrate the efficacy of the three examined machine learning models, random forest, HistGBoost, and LightGBM, in accurately categorizing corn and weed species. The evaluation involved the use of different feature selection strategies. All three models were found to achieve respectable accuracy without any feature selection, with LightGBM leading slightly at 0.863 and random forest at 0.816. The Fisher score, which assesses features' discrimination potential, resulted in a decrease in accuracy for random forest and LightGBM. Using random forest importance and chi-square as feature selection methods yielded modest accuracy improvements over the use of the Fisher score but did not surpass the performance of the models without feature selection. RFE provided a slight boost to HistGBoost and LightGBM.

At the same time, the Pearson correlation caused a significant reduction in accuracy across all models, implying that there is a weak linear relationship between features and the target variable. SFS and Information Gain emerged as the most effective feature selection

Table 2. Results of data normalization effects

Normalization methods	Accuracy		
	Random forest	HistGBoost	LightGBM
Without normalization	0.840	0.873	0.871
Min-max scaling	0.840	0.873	0.872
Z-score normalization	0.842	0.874	0.867
Robust scaling	0.842	0.872	0.869
Quantile transformation	0.841	0.874	0.872
Euclidean normalization	**0.887**	**0.905**	**0.897**

The bold font indicates the best performance of normalization method in each test.

Table 3. Results of feature selection effects

Feature selection methods	Accuracy		
	Random forest	HistGBoost	LightGBM
Without feature selection	0.816	0.861	0.863
Fisher score	0.784	0.83	0.815
Random forest importance	0.805	0.823	0.830
Chi-square	0.815	0.830	0.834
Recursive feature elimination	0.807	0.840	0.835
Pearson correlation	0.559	0.588	0.590
Sequential forward selection	0.879	0.891	0.883
Information Gain	**0.887**	**0.905**	**0.897**

The bold font indicates the best performance of feature selection method in each test

methods, as they significantly improved model accuracy. Information Gain produced the highest accuracy scores across all models, with a notable increase for HistGBoost to 0.905, thus positioning it as the most beneficial feature selection technique in this research. The method's effectiveness in capturing the most informative features for the classification task directly contributes to an enhancement in the predictive performance of the model. Therefore, the following experiment will explore the use of information gain for feature selection.

E. Effect of Leveling in Information Gain

The use of Information Gain for feature selection requires evaluation at each level. The output from this feature selection process is an Information Gain value that can be ranked. In this study, we will evaluate the use of 16 features or select the top 75% of features based on their ranking. The sequence of 16 features based on the Information Gain ranking is shown in Fig. 7. This experiment also includes an evaluation that uses traditional machine learning algorithms, such as support vector machine (SVM), k-nearest neighbor (KNN), and decision tree. Fig. 8 displays the accuracy results obtained from the Information Gain levelling experiment.

All models showed improved accuracy with an increase in the amount of information obtained, suggesting that having more informative features leads to higher classification performance. All models exhibit relatively low accuracy at the earlier levels of 1 and 2, implying that the features used may not have been sufficiently informative for effective discrimination between corn and weed. However, while progressing to higher levels, particularly from levels 4 to 13, there is a noticeable increase in accuracy for all models, with the most significant improvements observed in the ensemble methods: random forest, HistGBoosting, and LightGBM. Interesting levelling occurs at level 11, where the

Feature	Score
Correlation	0.622154
GreenVar	0.406227
GreenMean	0.397837
RedMean	0.388274
BlueMean	0.368635
RedVar	0.350753
BlueVar	0.307487
Entropy	0.221878
H3	0.214131
Inverse Difference Moments	0.212101
H1	0.174241
H4	0.165605
H5	0.162312
H6	0.161100
H7	0.158295
Contrast	0.157347

Fig. 7. Information Gain ranking results.

HistGBoosting model reaches an accuracy of 0.912 and LightGBM attains 0.911. The 11 most relevant features are found to be correlation, GreenVar, GreenMean, RedMean, BlueMean, RedVar, BlueVar, entropy, H3, inverse difference moments, and H1. They all have comparatively high scores, indicating robust performance in the classification tasks. In addition, the decision tree and random forest models demonstrate a substantial improvement at this stage, with respective accuracies of 0.827 and 0.895. The SVM and KNN models exhibit incremental enhancements but fail to achieve the elevated levels of accuracy that are demonstrated by ensemble approaches. This could be attributable to the capacity of ensemble models to identify and incorporate complex patterns in the data, which is further improved by utilizing information gained to select the most relevant features.

The data suggests that leveraging Information Gain to select handcrafted features can be particularly beneficial for ensemble methods in corn and weed classification, as

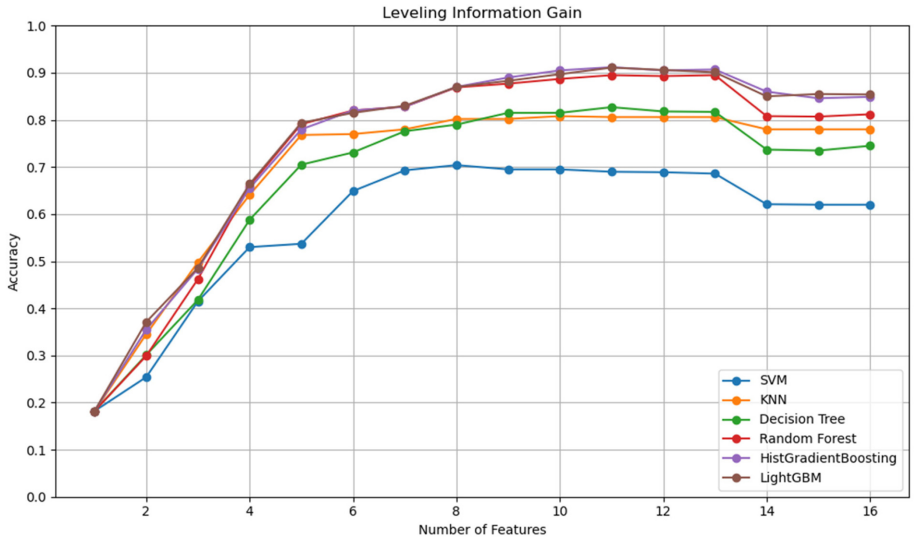

Fig. 8. Result of levelling Information Gain.

these methods appear to maximize feature relevance at higher information gain levels. Therefore, the following experiment will use 11 handcrafted features.

F. Exploration of Distance Algorithm

This step will evaluate the use of several distance algorithms, namely Euclidean, Manhattan, Minkowski, and Chebyshev. The features employed consist of 11 handcrafted features resulting from the feature selection using Information Gain. This experiment will also assess the performance of traditional machine learning algorithms, specifically SVM, KNN, and decision tree. Fig. 9 displays the experimental results obtained while utilizing variations of the distance algorithm for data normalization.

The experiment comparing various distance algorithms for data normalization in the classification of maize and weed has revealed that combining the Chebyshev distance algorithm and the HistGBoost model achieves the highest accuracy: This combination yields an accuracy score of 0.918. When partnered with the HistGBoost model, the Euclidean distance technique closely follows, with an accuracy score of 0.912. The LightGBM model, which again utilizes the Euclidean distance algorithm, also follows closely with an accuracy of 0.911.

The Chebyshev distance contributes effectively to the model's ability to classify the data, which could be attributable to its emphasis on the maximum difference between features, which might capture critical information for this specific classification task. Table 4 presents an example of distance normalization with the Chebyshev algorithm. Before the normalization process, the number scales on the features vary. For example, the GreenVar

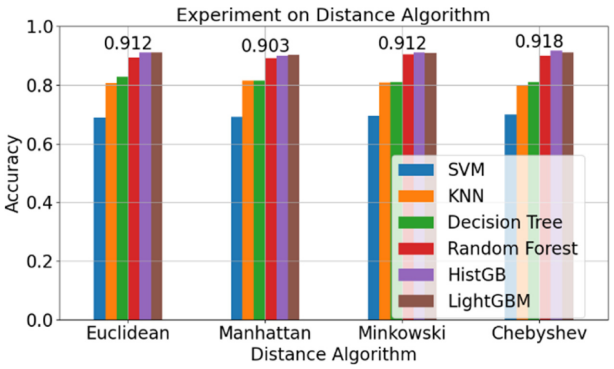

Fig. 9. Results of distance algorithm experiments.

Table 4. Example of normalization using Chebyshev normalization

	Normalization	
	Before	**After**
Correlation	0.741	0.026
GreenVar	28.798	1
GreenMean	5.051	0.175
RedMean	4.074	0.141
BlueMean	3.358	0.117
RedVar	23.647	0.821
BlueVar	20.061	0.697
Entropy	0.645	0.022
H3	0.033	0.001
Inverse difference moments	0.969	0.034
H1	0.48	0.017

and Redvar features produce large numbers of scales. The differences are in correlation, entropy, H3, inverse difference moment, and H1 features, where the value is less than 1. The normalization results change the value from 0 to 1 so that the same scale is used between features. The results suggest that the Chebyshev distance, when used for normalization in the HistGBoost, can achieve enhanced predictive performance for distinguishing between corn and weed species.

Traditional algorithms generally show lower accuracy compared to ensemble methods like random forest, HistGBoosting, and LightGBM. However, they still exhibit variability with different normalization techniques, thus underscoring the importance of feature scaling even when using these more basic models. The ensemble methods, leveraging their strength in aggregating multiple models or trees, tend to outperform these traditional single-model approaches, particularly when paired with the right normalization technique. Therefore, the parameter exploration experiment here will only evaluate ensemble learning.

G. Exploration of Parameter of Ensemble Learning

This experiment uses a number of ensemble learning methods, including random forest, HistGBoost, and LightGBM. Each method is configured with the values of relevant parameters. The n_estimators parameter in random forest is relevant to this method since it is responsible for the number of trees the model has in the forest. The max_iter parameter is associated with HistGBoost since it refers to the number of boosting stages that the model undertakes; this parameter influences the extent to which the model can learn the data. The parameter n_estimators is relevant to LightGBM as it refers to the number of boosted trees that are constructed. This measure is crucial to LightGBM since its capacity to address the issue is associated with gradient boosting, which the technique applies. However, the method also emphasizes other aspects, such as inference time. The values of parameters of these ensemble methods relevant to the experiment range from 60 to 100. This range is selected to assess whether there are benefits of the performance of the model when more trees or iterations are added to the complexity of data. The values are measured in increments of 10. The effect of the use of relevant parameters of ensemble learning on performance is presented in Table 5.

Analyzing Table 5 for the ensemble learning parameter experiment in corn and weed classification using handcrafted features, it is evident that there are models, random forest, HistGBoost, and LightGBM, with varying parameters for the number of n_estimators for random forest and LightGBM as well as max_iter for HistGBoost.

Random forest shows a slight fluctuation in accuracy, with its highest value being 0.901, which is achieved at several points when the n_estimators are set to 70, 80, and 100. This indicates that there is a plateau in performance improvement beyond a certain number of trees, suggesting that adding more estimators does not necessarily lead to higher accuracy. HistGBoost consistently delivers robust performance, with accuracy scores exceeding 0.918 across all iterations. Notably, the model achieves its peak accuracy at 0.92 with max_iter set to 90. There is an optimal number of iterations for this dataset and feature set beyond which performance does not improve and may even slightly decrease. LightGBM, similar to random forest, does not exhibit a clear trend with increasing n_estimators. Its accuracy hovers just above 0.91 with minor variations; interestingly, it does not consistently increase with more estimators. The highest accuracy for LightGBM is 0.912, with n_estimators set to 70.

Table 5. Results of the parameter experiment

Number of estimators	Accuracy		
	Random forest	HistGBoost	LightGBM
60	0.895	0.919	0.907
70	0.901	0.919	0.912
80	0.901	0.919	0.911
90	0.900	0.920	0.911
100	0.901	0.918	0.910

The best accuracy is obtained with HistGBoost at 0.92, indicating its effectiveness for the given dataset. Each model has its optimal parameters for the number of trees or iterations, and there is no direct correlation between the number of estimators and accuracy. These nuances highlight the significance of parameter adjustment in ensemble learning. This experimentation highlights that, while the use of ensemble methods generally improves prediction accuracy, careful tuning is critical to extracting the best performance for specific case studies, such as classifying corn and weed species.

H. Discussion

Handcrafted and non-handcrafted features are frequently used in case study classifications, such as corn and weed classification. Previous research has leaned toward deep learning and its modified counterparts for optimal accuracy. Each feature is handcrafted, manually crafted based on domain expertise, or non-handcrafted, i.e., automatically extracted by deep learning models. Each approach has its own set of advantages and limitations. Table 6 presents a comparative analysis of handcrafted versus non-handcrafted features grounded in deep learning techniques, while highlighting how each contributes to

Table 6. Comparison between using handcrafted and non-handcrafted features

Methods	Accuracy	Inference time (ms)
ResNet-101 [6]	0.965	337.5
GCN-ResNet-101 [6]	0.978	432.5
Proposed		
HistGBoost	0.920	5.955
LightGBM	0.912	1.998
Random forest	0.901	3.988

the performance and effectiveness of corn and weed classification. We evaluate inference time using our hardware specifications.

As can be seen in Table 6, deep learning methods like ResNet-101 and GCN-ResNet-101 achieve higher accuracy rates in corn and weed classification of 0.965 and 0.978, respectively, but ensemble learning methods offer significantly faster inference times. For instance, HistGBoost achieves an accuracy of 0.92, which is competitive with the deep learning models but with a substantially lower inference time of only 5.955 ms. LightGBM has a slightly lower accuracy of 0.912 but an even faster inference time of 1.998 ms. Meanwhile, random forest achieves an inference time of 3.988 ms while maintaining an accuracy of 0.901. These results prove that using handcrafted features with 11 relevant features can be used to obtain results that approach the accuracy results of the non-handcrafted approach. Still, the model inference speed is very different and much faster using handcrafted features.

This trade-off between accuracy and inference speed is fundamental in real-time implementation. Although deep learning models are frequently accurate, they usually require high computational resources and time for processing, which may not be optimal in a real-time context. By contrast, ensemble learning approaches balance inference time and accuracy by employing feature selection with information gain and Chebyshev normalization on handcrafted features. Consequently, these approaches are highly suitable for real-time conditions that require fast processing. The results obtained through this research indicate that ensemble models perform much better when effective normalization techniques are included, alongside accurate feature engineering and selection. The limitation of this research is in selecting handcrafted features. This selection must take place only after a great deal of experimentation, as feature selection and data normalization inherently alter the performance of machine learning models depending on manually handcrafted features, an instance of which is their quality and relevance. There is therefore a need for a significant investment of time in the experiment to find the most effective features.

Moreover, the appropriateness of handcrafted features is quite specific to the exact case study or dataset in question. Features are handcrafted with several relevant general patterns in mind, as they do not contain overall crucial system information that would justify their applicability in all novel situations. The resulting models may be quite good within the selected domain. However, they need more generality to perform well across various datasets. This contrasts with deep learning models, which can adapt from data independently and demonstrate much better generalizability across novel contexts and datasets in many cases

V. CONCLUSION

The accurate classification of corn and weeds in agricultural areas requires the use of accurate and real-time performance methods. Application in the field requires accurate results, and the method that is used must also be capable of fast classification. Deep learning produces high accuracy, but the speed of model inference is very slow. Therefore, this research proposes using handcrafted features by carrying out feature selection and data normalization. The best results were obtained when using a combination of Information Gain for feature selection and Chebyshev distance for data normalization. The resulting accuracy was 0.92, which is not very different from the results presented in previous studies. However, the model inference time is 72 times faster using HistGBoost, namely 5.955 ms. When using the LightGBM method, the inference time was 1.998 ms for processing per image, or 217 times faster than the fastest time presented in previous research. This achievement means the method we propose is capable of running in real-time. Through these results, handcrafted methods can be implemented in real-time for weed detection in the field. Future research can use other handcrafted features to achieve more optimal accuracy results. However, beyond accuracy, it is also necessary to evaluate the inference time to facilitate real-time application.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

ACKNOWLEDGEMENTS

This research was supported by a research grant from Rekognisi Tugas Akhir (RTA) Universitas Gadjah Mada year 2024 with Notification Letter No. 4971/UN1.P1/PT.01.01/2024.

REFERENCES

1. A. Monteiro and S. Santos, "Sustainable approach to weed management: the role of precision weed management," *Agronomy*, vol. 12, no. 1, article no. 118, 2022. https://doi.org/10.3390/agronomy12010118
2. W. H. Su, "Crop plant signaling for real-time plant identification in smart farm: a systematic review and new concept in artificial intelligence for automated weed control," *Artificial Intelligence in Agriculture*, vol. 4, pp. 262-271, 2020. https://doi.org/10.1016/j.aiia.2020.11.001
3. A. Subeesh and C. R. Mehta, "Automation and digitization of agriculture using artificial intelligence and Internet of Things," *Artificial Intelligence in Agriculture*, vol. 5, pp. 278-291, 2021. https://doi.org/10.1016/j.aiia.2021.11.004
4. A. Nasiri, M. Omid, A. Taheri-Garavand, and A. Jafari, "Deep learning-based precision agriculture through weed recognition in sugar beet fields," *Sustainable Computing: Informatics and Systems*, vol. 35, article no. 100759, 2022. https://doi.org/10.1016/j.suscom.2022.100759
5. F. Garibaldi-Marquez, G. Flores, D. A. Mercado-Ravell, A. Ramirez-Pedraza, and L. M. Valentin-Coronado, "Weed classification from natural corn field-multi-plant images based on shallow and deep learning," *Sensors*, vol. 22, no. 8, article no. 3021, 2022. https://doi.org/10.3390/s22083021
6. H. Jiang, C. Zhang, Y. Qiao, Z. Zhang, W. Zhang, and C. Song, "CNN feature based graph convolutional network for weed and crop recognition in smart farming," *Computers and Electronics in Agriculture*, vol. 174, article no. 105450, 2020. https://doi.org/10.1016/j.compag.2020.105450
7. K. Hu, G. Coleman, S. Zeng, Z. Wang, and M. Walsh, "Graph weeds net: a graph-based deep learning method for weed recognition," *Computers and Electronics in Agriculture*, vol. 174, article no. 105520, 2020. https://doi.org/10.1016/j.compag.2020.105520
8. M. Alam, M. S. Alam, M. Roman, M. Tufail, M. U. Khan, and M. T. Khan, "Real-time machine-learning based crop/weed detection and classification for variable-rate spraying in precision agriculture," in *Proceedings of 2020 7th International Conference on Electrical and Electronics Engineering (ICEEE)*, Antalya, Turkey, 2020, pp. 273-280. https://doi.org/10.1109/ICEEE49618.2020.9102505
9. H. Pathak, C. Igathinathane, K. Howatt, and Z. Zhang, "Machine learning and handcrafted image processing methods for classifying common weeds in corn field," *Smart Agricultural Technology*, vol. 5, article no. 100249, 2023. https://doi.org/10.1016/j.atech.2023.100249
10. N. Islam, M. N. Rashid, S. Wibowo, C. Y. Xu, A. Morshed, S. A. Wasimi, S. Moore, and S. M. Rahman, "Early weed detection using image processing and machine learning techniques in an Australian chilli farm," *Agriculture*, vol. 11, no. 5, article no. 387, 2021. https://doi.org/10.3390/agriculture11050387
11. F. Alserhani and A. Aljared, "Evaluating ensemble learning mechanisms for predicting advanced cyber attacks," *Applied Sciences*, vol. 13, no. 24, article no. 13310, 2023. https://doi.org/10.3390/app132413310
12. K. Li, P. Rahnama, R. Novella, and B. Somers, "Combining flamelet-generated manifold and machine learning models in simulation of a non-premixed diffusion flame," *Energy and AI*, vol. 14, article no. 100266, 2023. https://doi.org/10.1016/j.egyai.2023.100266
13. N. R. Abid-Althaqafi and H. A. Alsalamah, "The effect of feature selection on the accuracy of X-Platform User credibility detection with supervised machine learning," *Electronics*, vol. 13, no. 1, article no. 205, 2024. https://doi.org/10.3390/electronics13010205
14. A. Nazir and R. A. Khan, "A novel combinatorial optimization based feature selection method for network intrusion detection," *Computers and Security*, vol. 102, article no. 102164, 2021. https://doi.org/10.1016/j.cose.2020.102164
15. T. Yin, H. Chen, T. Li, Z. Yuan, and C. Luo, "Robust feature selection using label enhancement and β-precision fuzzy rough sets for multilabel fuzzy decision system," *Fuzzy Sets and Systems*, vol. 461, article no. 108462, 2023. https://doi.org/10.1016/j.fss.2022.12.018
16. A. Khoder and F. Dornaika, "Ensemble learning via feature selection and multiple transformed subsets: Application to image classification," *Applied Soft Computing*, vol. 113, article no. 108006, 2021. https://doi.org/10.1016/j.asoc.2021.108006
17. N. Bento, J. Rebelo, M. Barandas, A. V. Carreiro, A. Campagner, F. Cabitza, and H. Gamboa, "Comparing handcrafted features and deep neural representations for domain generalization in human activity recognition," *Sensors*, vol. 22, no. 19, article no. 7324, 2022. https://doi.org/10.3390/s22197324
18. S. Chen, Z. Liu, W. Zhang, and J. Yang, "A hard-constraint wide-body physics-informed neural network model for solving multiple cases in forward problems for partial differential equations," *Applied Sciences*, vol. 14, no. 1, article no. 189, 2024. https://doi.org/10.3390/app14010189
19. T. Wang, G. Song, W. Ni, and Q. Zeng, "MIFD-Net : a hand gesture recognition model based on feature fusion of MLP and CNN," in *Third International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI 2022) (Proceedings of SPIE 12509)*. Bellingham, WA: International Society for Optics and Photonics, 2023, p. 1250905. https://doi.org/10.1117/12.2655882
20. U. Kilic, E. S. Essiz, and M. K. Keles, "Binary anarchic society optimization for feature selection," *Romanian Journal of Information Science and Technology*, vol. 26, no. 3-4, pp. 351-364, 2023. https://doi.org/10.59277/ROMJIST.2023.3-4.08
21. A. Lehavi and S. Kim, "Feature reduction method comparison towards explainability and efficiency in cybersecurity intrusion detection systems," in *Proceedings of the 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, Nassau, Bahamas, 2022, pp. 1326-1333. https://doi.org/10.1109/ICMLA55696.2022.00211
22. K. D. Sree Devi, P. Karthikeyan, U. Moorthy, K. Deeba, V. Maheshwari, and S. M. Allayear, "Tumor detection on microarray data using grey wolf optimization with gain information," *Mathematical Problems in Engineering*, vol. 2022, article no. 4092404, 2022. https://doi.org/10.1155/2022/4092404
23. M. A. Raibag, J. V. Franklin, and R. Sarkar, "An investigation on epileptic seizure classification using machine learning and multiple feature selection strategies," in *Proceedings of 2022 3rd International Conference for Emerging*

Technology (INCET), Belgaum, India, 2022, pp. 1-6. https://doi.org/10.1109/INCET54531.2022.9824799
24. S. Nuanmeesri, "Feature selection for analyzing data errors toward development of household big data at the sub-district level using multi-layer perceptron neural network," *International Journal of Interactive Mobile Technologies*, vol. 16, no. 5, pp. 121-138, 2022. https://doi.org/10.3991/ijim.v16i05.22523
25. S. Kandanaarachchi, M. A. Munoz, R. J. Hyndman, and K. Smith-Miles, "On normalization and algorithm selection for unsupervised outlier detection," *Data Mining and Knowledge Discovery*, vol. 34, no. 2, pp. 309-354, 2020. https://doi.org/10.1007/s10618-019-00661-z
26. L. Silva, J. Hermsdorf, V. Guedes, F. Teixeira, J. Fernandes, B. Bispo, and J. P. Teixeira, "Outliers treatment to improve the recognition of voice pathologies," *Procedia Computer Science*, vol. 164, pp. 678-685, 2019. https://doi.org/10.1016/j.procs.2019.12.235
27. B. Zhang, S. Hu, and M. Li, "Comparative study of multiple machine learning algorithms for risk level prediction in goaf," *Heliyon*, vol. 9, no. 8, article no. e19092, 2023. https://doi.org/10.1016/j.heliyon.2023.e19092
28. E. Lafuente, J. C. Leiva, J. Moreno-Gomez, and L. Szerb, "A non-parametric analysis of competitiveness efficiency: the relevance of firm size and the configuration of competitive pillars," *BRQ Business Research Quarterly*, vol. 23, no. 3, pp. 203-216, 2020. https://doi.org/10.1177/2340944420941440
29. A. Aytekin, "Comparative analysis of normalization techniques in the context of MCDM problems," *Decision Making: Applications in Management and Engineering*, vol. 4, no. 2, pp. 1-25, 2021. https://doi.org/10.31181/dmame210402001a
30. F. Mekhalfa and F. Yacef, "Supervised learning for crop/weed classification based on color and texture features," 2021, [Online]. Available: http://arxiv.org/abs/2106.10581.
31. P. Bosilj, T. Duckett, and G. Cielniak, "Connected attribute morphology for unified vegetation segmentation and classification in precision agriculture," *Computers in Industry*, vol. 98, pp. 226-240, 2018. https://doi.org/10.1016/j.compind.2018.02.003
32. A. Muneer and S. M. Fati, "Efficient and automated herbs classification approach based on shape and texture features using deep learning," *IEEE Access*, vol. 8, pp. 196747-196764, 2020. https://doi.org/10.1109/ACCESS.2020.3034033
33. Y. Zheng, Q. Zhu, M. Huang, Y. Guo, and J. Qin, "Maize and weed classification using color indices with support vector data description in outdoor fields," *Computers and Electronics in Agriculture*, vol. 141, pp. 215-222, 2017. https://doi.org/10.1016/j.compag.2017.07.028
34. F. Lin, D. Zhang, Y. Huang, X. Wang, and X. Chen, "Detection of corn and weed species by the combination of spectral, shape and textural features," *Sustainability*, vol. 9, no. 8, article no. 1335, 2017. https://doi.org/10.3390/su9081335
35. A. Humeau-Heurtier, "Texture feature extraction methods: a survey," *IEEE Access*, vol. 7, pp. 8975-9000, 2019. https://doi.org/10.1109/ACCESS.2018.2890743
36. M. A. Thanoon, M. J. M. Zedan, and A. N. Hameed, "Feature selection based on wrapper and information gain," in *Proceedings of 2019 1st Al-Noor International Conference for Science and Technology (NICST)*, Sulimanyiah, Iraq, 2019, pp. 32-37. https://doi.org/10.1109/NICST49484.2019.9043805
37. A. Mortazavi and M. H. Moattar, "Robust Feature selection from microarray data based on cooperative game theory and qualitative mutual information," *Advances in Bioinformatics*, vol. 2016, article no. 1058305, 2016. https://doi.org/10.1155/2016/1058305
38. D. Huang, Z. Liu, and D. Wu, "Research on ensemble learning-based feature selection method for time-series prediction," *Applied Sciences*, vol. 14, no. 1, article no. 40, 2024. https://doi.org/10.3390/app14010040
39. M. L. McHugh, "The chi-square test of independence," *Biochemia Medica*, vol. 23, no. 2, pp. 143-149, 2013. https://doi.org/10.11613/BM.2013.018
40. K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353-363, 2015. https://doi.org/10.1016/j.snb.2015.02.025
41. Q. Yang, Q. Kang, Q. Huang, Z. Cui, Y. Bai, and H. Wei, "Linear correlation analysis of ammunition storage environment based on Pearson correlation analysis," *Journal of Physics: Conference Series*, vol. 1948, article no. 012064, 2021. https://doi.org/10.1088/1742-6596/1948/1/012064
42. J. B. Awotunde, F. E. Ayo, R. Panigrahi, A. Garg, A. K. Bhoi, and P. Barsocchi, "A multi-level random forest model-based intrusion detection using fuzzy inference system for Internet of Things networks," *International Journal of Computational Intelligence Systems*, vol. 16, article no. 31, 2023. https://doi.org/10.1007/s44196-023-00205-w
43. D. Singh and B. Singh, "Feature wise normalization: an effective way of normalizing data," *Pattern Recognition*, vol. 122, article no. 108307, 2022. https://doi.org/10.1016/j.patcog.2021.108307
44. H. Nhat-Duc and T. Van-Duc, "Comparison of histogram-based gradient boosting classification machine, random forest, and deep convolutional neural network for pavement raveling severity classification," *Automation in Construction*, vol. 148, article no. 104767, 2023. https://doi.org/10.1016/j.autcon.2023.104767
45. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu, "LightGBM: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146-3154, 2017.
46. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001. https://doi.org/10.1023/A:1010933404324
47. A. Bakhshipour, "Cascading feature filtering and boosting algorithm for plant type classification based on image features," *IEEE Access*, vol. 9, pp. 82021-82030, 2021. https://doi.org/10.1109/ACCESS.2021.3086269

Faisal Dharma Adhinata https://orcid.org/0000-0002-2624-173X

Faisal Dharma Adhinata earned his Master of Computer Science (M.Cs.) degree in computer science from Universitas Gadjah Mada, Indonesia. He is pursuing a Doctoral Degree in Computer Science at the Universitas Gadjah Mada, Indonesia. In addition, he serves as a lecturer in the Department of Software Engineering at Institut Teknologi Telkom Purwokerto, Indonesia. His research areas include Artificial Intelligence, image processing, and computer vision.

Wahyono https://orcid.org/0000-0002-2639-8411

Wahyono received the B.Sc. degree in Computer Science from Universitas Gadjah Mada, Indonesia, and the Ph.D. degree from The University of Ulsan, South Korea. He is an Associate Professor at the Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia. His research interests include machine learning, computer vision, and pattern recognition. He actively participates as a member of the societies, a reviewer in reputable international journals, and an editor in several journals.

Raden Sumiharto https://orcid.org/0000-0003-4902-0697

Raden Sumiharto received the B.Sc. degree in Physics from Universitas Gadjah Mada, Indonesia, the master's degree from the Universitas Gadjah Mada, Indonesia and the Ph.D. degree from the Universitas Gadjah Mada, Indonesia. He is a lecturer in the Department of Computer Sciences and Electronics, Universitas Gadjah Mada. His research interests include parallel processing, image processing, sensor network, and computer vision.