Weed Detection in Soybean Crops using Regression Analysis and Deep Learning

MSc Research Project
Data Analytics

Rohit Vad
Student ID: X18190936

School of Computing
National College of Ireland

Supervisor: Mr. Hicham Rifai

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rohit Vad		
Student ID:	X18190936		
Programme:	M.Sc. Data Analytics	**Year:**	2019-2020
Module:	Research Project		
Supervisor:	Hicham Rifai		
Submission Due Date:	28/09/2020		
Project Title:	Weed Detection in Soybean Crops Using Regression Analysis and Deep Learning		
Word Count:	**8097**	**Page Count 20**	

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Rohit Vad

Date: 27/09/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Weed Detection in Soybean Crops using Regression Analysis and Deep Learning

Rohit Vad
X18190936

Abstract
The troublesome and unwanted effects of weeds are often are a cause for an injury or degradation to the health of the yield. Weeds constantly consume nutrients, water, and sunlight, thereby decreasing the quality and the quantity of the crop that is cultivated. Using deep learning and regression analysis, this paper focuses on the principle of image classification and extends it to detecting weeds in soybean crops and classifying the images into broadleaf weeds, soil, grass or soybean. This application provides great help to the agricultural sector by aiding the farmers select the specific herbicide for spraying depending on the type of weeds and thereby stunting it. Deep learning techniques such as Convolution Neural Network (CNN), Artificial Neural Network (ANN) and traditional machine learning algorithm Logistic Regression were implemented for the purpose of detecting and classifying weeds in images. The performances of these models were compared and evaluated based on accuracy of each model. The Convolutional Neural Network model significantly outperforms the other two with an accuracy of 97%.

1 Introduction

There are constraints of different types when it comes to growing crops, as well as various socio-economic factors. Of these, weeds are the biggest biotic constraint that affects the agricultural produce the most in developed as well as developing countries (Grand challenges in weed management). The damage or loss of yield due to the adverse effects of weeds are unprecedented. The cost of controlling weeds in Australia was estimated at AUD 3.3 billion per year in 2016 Llewellyn et al. (2016) while a staggering cost of USD 11 billion per year was the estimated cost for Indian agriculture to tackle the problem of weeds Gharde et al. (2018). Thus, the demand for new systems or the need for automation in existing systems to efficiently determine and classify weeds and differentiate them into different categories is justified.

A recent study involving a data from 2019 indicated that the total global production of soybean was 341.8 million metric tonnes. With the growing applications of soybean in the food, biodiesel and other industries, it is therefore essential that new techniques are developed to tackle the issue of weeds hampering the yield quality and quantity to secure a higher production of soybean, while also understanding the importance in a socio-economic aspect. (International: World soybean production, 2019).

Weeds are troublesome and unwanted plants that grow on the agricultural land and often competes and dominates the sunlight, nutrients in the soil, space and water. This competition in turn results in a higher production cost, substantially higher damage to the crops and also results in an infected yield Rizzardi and Fleck (2004).

Broadleaf and grass are the prime categories of weeds. Their treatment also varies as per the kind of herbicides used for a specific class of weed that enhances the results if the treatment is specific for a certain species of weed (Herrera et. al, 2014). Often, herbicides and other chemicals are utilized to handle the problem posed by weeds. There are, however, many damaging effects of these chemicals on the crops along with the environment Dankhara, Patel and Doshi (2019).

Rizzardi and Fleck (2004) believe that it is paramount to know the adverse effects of weed infestation so as to come up with effective strategies to fight against it. The shift towards increasing the production while minimizing the use of herbicides, insecticides and other chemicals gave birth to the idea of precision agriculture. Wrigley et al. (2016) Precision agriculture includes a way of managing a farm and its resources by collecting the data of the farmland and the physical factors that aids in the growth of crops producing a better yield and a higher quantity.

Therefore, it is of utmost importance to minimize the herbicide use and automate the available technology for a simpler and a more efficient management and elimination of weeds at an early stage, while also making use of the data of the farmland as well as the physical factors that affect the crop growth, thereby promoting precision agriculture for obtaining better results in terms of quality as well as quantity of the yield.

To maximize the production, efficiency and the profits of the farmland and simultaneously reducing the adverse effects to the ecosystem, precision farming makes use of high-end technology to fulfil all these desired goals. Precision farming combines information and production-based farming methodologies Bucci et al. (2019). A further impetus to precision agriculture has been experienced owing to a higher use of technology which includes leveraging use of various machine learning and deep learning methodologies. This has substantially brought about an improvement in all the activities and tasks that require the use of computer vision and use the above mentioned techniques for identifying, differentiating and separating of weeds from crops in farmlands Dankhara, Patel and Doshi (2019).

1.1 Research Question

How well can logistic regression, convolutional neural network and artificial neural network help in the detection of weeds in soybean crops?

2 Related Work

This section focuses on the previous work done or studies carried out for detection of weeds in crops and differentiating them. To better understand the process and the related work, it is important to first understand image processing and image classification.

2.1 Image Processing

Image processing involves manipulation images using a computer. The growing importance of image processing for various technological and non-technological aspects and applications has made it a cornerstone in the field of computer science. The process of image processing leverages either digital or analogue image processing to obtain useful information from images which involves performing certain operations such as enhancement, restoration, analysis, and compression of images.

Images can be considered as a continuous function of two variables (x, y), that involves the use of sampling to be processed digitally, which need to be converted into a matrix form in order to process them digitally. The applications of image processing are vast, and have become indispensable particularly in the medical field, with x-ray images, gamma rays' images in nuclear medicine, infrared microwaves, and processing radio waves images. Digital images are intensively used in the electromagnetic spectrum. The use of ultrasound imaging is a common practice for sonographies etc (Image Processing - an overview | ScienceDirect Topics, 2020), (Anbarjafari, 2020).

2.2 Image Classification

Image classification basically involves extracting information from images. The notion that and image contains a single feature that is unique or multiple features of which every feature belongs to another unique class. The classification of images is done after leveraging many different decision-making approaches along with theoretic approaches for identification and classification of images. Depending on the need of the classification, images can be classified in a supervised or an unsupervised manner.

During the initial phases, each class category is created based on the unique characteristics of the images. Images are classified and categorized in the further phases with the help of these obtained characteristics (What is image classification? ArcGIS Help | Documentation, 2020), (Image Analysis - Classification, 2020).

2.3 Analysis of studies based on traditional machine learning methods

Agriculture and activities relating to it have benefitted by leveraging image classification to enhance various aspects of farming. Multiple technologies and techniques are utilized to categorize different types of crops, defects in the crops, seedlings etc. based on their images. The eventual aim is to automate the process of crop and weed classification with the aid of various technologies.

Recent study by Hung Xu and Sukkarieh (2014) involved the use of Sparse Autoencoders in order to try and efficiently classify weeds. There were three species of weed that were mainly under the scope viz. Tropical Soda Apple, Serrated Tussock and Water Hyacinth. The implemented model achieved different precision scores depending on the pixel sizes. A precision score of 72.2% was attained when the Pixel size was 256, while a score of 92.9% was obtained when the Pixel size was 384, and finally, a precision reading of 94.3% for a pixel size of 512 was obtained. There was, however, a drawback in the data collection phase as the drone used to capture the data could not cover large areas due to a short flight.

Ahmed et al. (2012) leveraged Support Vector Machines (SVM) to identify and differentiate between 6 classes of weeds. An accuracy of 97.4% was attained on a relatively small dataset comprising of 224 images, while Herrera et al. (2014) implemented a Fuzzy decision-making model making use of shape descriptors on an even smaller dataset that contained only 66 images and produced an output accuracy of 92.9%.

In the earlier part of the previous decade, Support Vector Machines was a popular approach due to a higher generalization capacity and output accuracy that was obtained in most studies. Thus, continuing the trend, Tellaeche et al. (2011) implemented the SVM model for the purpose of classification of weeds as did Saha et al. (2016). The SVM model was also preferred by Siddiqi et al. (2014) for the classification of weeds. This model was applied on a dataset that comprised of 1200 images in all and produced an output accuracy of 98.1%.

An alternative approach to substitute human labour in the form of an intelligent robot model was proposed by Beibei and Matson (2015) while also leveraging the traditional machine learning algorithms for weed discrimination. Rice and weeds could be classified by the intelligent robot with the help of a sensor that was attached, which enabled an easy transition towards automating the process. Algorithms such as Decision Tree, Naïve Bayes and Support Vector Machine were implemented, with a correlation matrix used to measure the performance of each algorithm. The algorithms were trained using feature extraction with a view to obtain a higher and a better precision for discrimination between the rice crops and weeds. The ends or tips of the leaves, rice and that of weeds were separated with the help of Harris Corner Algorithm. Out of the 1261 Harris points, 561 rice tips were accounted while the other 700 were weed tips. With an output accuracy of 98.2%, the Decision Tree algorithm was better suited for this classification outperforming the other two.

Blob analysis along with Support Vector Machine was leveraged by Sadia et al. (2018) with an aim to provide an effective and efficient solution for the classification of weeds and to provide a tool for agri-robots. A combination of manually clicking the photographs and using a camera mounted on a tractor was used to create a relatively small dataset of 80 images, which was further separated in the proportion 90% train set and 10% test set. The model was trained using feature extraction, and the images were resized by 50% in size while ensuring no data loss occurred while processing the images. This was followed by segmentation and binarization to transform the colored images to black and white. These images had a 10% threshold. Blob analysis helped differentiate the background and the object. Results of varying accuracy were obtained when SVM, Blob analysis and binarization were implemented in combination. The type of images dictated the accuracy obtained, that varied between 50% and 95%. Overlapping images had a lower accuracy while images that were distinct and well defined which contained weeds or crops, naturally had a higher accuracy.

Another research work used primarily SVM and Random forest algorithm to classify weeds Kamath, Balachandra and Prabhu (2020). These models were implemented in combination with other techniques to enhance the results. Three types of weeds viz. high weed, low weed and no weed were distinguished on the basis of their density in the images used. Support Vector Machine is implemented with linear and Radial Bias Function (RBF) kernel, while Random Forest is applied as not a lot of hyper parameters require fine tuning. The inference of this research study showed that Random Forest had a better performance as compared to SVM, with 86% accuracy of the output while the SVM model had an accuracy of 73%.

As seen from the above cited research studies, traditional machine learning approaches were quite popular amongst researchers for classification of images during the earlier part of the decade with SVM being a popular choice for image classification, with a few researchers opting for Random forest and Decision trees. These models, however, were primarily implemented on smaller datasets which did not require a lot of processing time.

In the following section, this paper discusses the performance of research studies that made use of Neural Networks.

2.4 Analysis of studies based on neural network

Neural Networks and their advantages have been better recognized and understood due to their capacity to deal with complex and sizeable data and a faster computing power.

A convolutional neural network (CNN) model was implemented by Ferreira et al. (2018) for identifying weeds in soybean crops. The CNN model was judged against the performance of machine learning models such as AdaBoost, AdaBoost, Random Forest and SVM. The dataset used for this research study was obtained with the help of an Unmanned Aerial Vehicle (UAV) followed by image segmentation by leveraging algorithm of superpixels. The segmented images paved the way for a dataset containing classes namely grass, broadleaf and soil and soybean. Feature extraction was performed on this dataset.

ConvNets were trained with the help of the segments obtained and then the images were classified to check if weeds are present or absent in an image. The accuracy across all the images for the CNN model was 99% and the overall accuracy obtained was 98% which made the CNN model outperform the other models.

The problem of weed detection has been present for a while now and this problem has also seen the use of a Single Layer Perceptron model to classify broadleaf weeds and grasses. An accuracy of 93.7 was obtained on a dataset that comprised of 400 images Ishak, Hussain and Mustafa (2009).

A deep residual neural network was leveraged to identify and categorize external defects in tomatoes. Fine tuning and feature extraction were used to train the residual neural network to make it adept at determining the external defects. After tuning and feature extraction, the output average precision obtained was 94.6% Da Costa et al. (2019).

A sizeable dataset containing 17000 images of 8 relevant weed species that are found in 8 different regions in Australia. Olsen et al. (2019) modelled this dataset using deep learning techniques such as Inception-v3 model and a ResNet-50 model for classification purposes. The Inception-v3 model outperformed the ResNet-50 model with a precision of 95.7% as opposed to that of 95.1% of the ResNet-50 model.

Saad, Mohammed and Essaid (2018) leveraged Back-propagation Neural Network (BPNN) model and compared its performance against SVM model. The BPNN model performed slightly better with an obtained accuracy of 96.6% as compared to that of the SVM which was 95.1%. The purpose was to conduct a research study while making use of two classifiers while working on a dataset obtained from a soybean plantation in Brazil.

With time, the use of UAV's gained impetus for capturing images to form a dataset. This application is widely used for detection of weeds in fields. Bahn et al. (2018) made use of a vision-based classification system where in UAV's were used to capture images of the farmland. This step was followed by using CNN's coupled with crop-line information for categorizing weeds in bean fields, spinach, and beets. Due to this being an unsupervised machine learning method, therefore, it was not dependent on the training data. The high-resolution images in the dataset produced wavering results contingent on the type of field. The accuracy for beet fields was 93% whereas for the bean fields it was 69% and finally, the spinach fields had an accuracy of 81%.

The performance of SVM was compared against the of ANN in another research work carried out by Bakhshipour et al. (2017) to detect the presence of weeds in sugar beet fields using shape characteristics which consisted of Fourier descriptors and moment invariant features to detect and differentiate between four commonly found weed species in sugar beet fields. In this case, the SVM model outperformed the ANN model with an accuracy of 95% as opposed to 92.92% if the ANN. Shape extraction features were leveraged to obtain these accuracies, while other methods such as color or texture features were not used.

A low-cost Weed Identification System (WIS) was proposed and implemented by Liang, et al. (2019) in opposition to the otherwise costly models in terms of power consumption and capital used for detection of weeds. The dataset and the farmland observed was based in Taiwan of which images were acquired using a drone mounted with a camera capable of capturing RGB images of the fields. A convolutional neural network (CNN) model containing three convolutional layers, four dropout layers, three pooling layers and one fully connected layer was leveraged. A mobile application was created to enhance user experience and to notify the users when weeds were detected, which contained the location at which the weed was detected on the farmland. An accuracy of 98.8% was obtained using this model.

Kumaraswamy et al. (2019) conducted a research study with the task of detecting weeds that leveraged and compared performances of models like Support Vector Machines (SVM), Artificial Neural Network (ANN) and Convolutional Neural Network (CNN). Weeds such as Para grass and Nutsedge were to be identified in Chrysanthemum. Keras packaged, OpenCV and python were utilized for the classification process on a dataset that consisted of 2560 images. The use of SVM with Radial Bias Function (RBF) and Polynomial function helped it perform slightly better. However, the CNN model that had a dimension of 250*250*3 for the input layer, outperformed the ANN and the SVM model.

Asad and Bais (2019) performed their weed detection study on canola fields in two steps; maximum likelihood classification was used to separate the background from the foreground followed by naming the pixels of weed. Multiple deep learning models were implemented out of which ResNet-50 outperformed the other models with an output accuracy of 99.48%.

Gao et al. (2020) detected the weed Convolvulus sepium (hedge bindweed) in sugar beet fields which were located in Belgium under different lighting conditions with the help of convolutional neural network model. An initial size of 4000*6000 pixels was resized to 800*1200 pixels. The model implemented was YOLOv3 (You Look Only Once) which was then trained by using 2271 synthetic images. Considering speed and accuracy as performance parameters, the implemented model produced steady results and was compatible to be deployed on a mobile platform.

From the literature studies mentioned above, capturing images by using a UAV is gaining an impetus amongst researchers. Using UAV’s can aid in detection objects while using autonomous navigation control. Using Convolutional Neural Network for weed detection is seen as a popular choice amongst researchers. The use of Neural Networks is preferred due to its ability of fast computation and the ability to train the network in a single cycle as compared to other traditional approaches.

2.5 Analysis of studies based on Internet of Things (IoT)

Precision agriculture encourages the use of technology. With the rising applications of Internet of Things (IoT) and the increasing use of UAV’s, can result in enhanced results with respect to efficiency while also minimizing the adverse effects on the environment.

In a study performed by Kiani and Seyyedabbasi (2018) that saw the use of precision agriculture which included the use IoT which divided the farmland into multiple blocks and physical factors such as moisture content in the soil, temperature and humidity in these regions was measured using sensors. This resulted in an efficient cultivation and management of the farmland for the farmers.

Classifying weeds using IoT is a fairly new research area and Dankhara, Patel and Doshi, (2019) proposed a model which involved the use IoT with an aim to minimize human intervention by introducing intelligent robots which would automate the procedure of detecting weeds.

Raspberry PI with a compatible camera of 8MP which comes with an image sensor (Sony IMX219) comprised of the hardware that was under use. For selective and efficient use of herbicides, a sprayer is mounted in order to spray herbicides depending on the type of weed. This approach is gaining popularity in spite of being relatively new as researchers are adopting this method as it automates the process and also reduces the use of herbicides significantly.

The applications of IoT in farming and for weed detection are vast, as intelligent robots can be used to automate the process. These can be modelled using neural networks to classify weeds or detect them or according to the required function which helps in reducing human intervention in farming.

2.6 Summary

The study implements an ANN model, CNN model and a Logistic Regression model with a view to perform a comparison between neural networks and traditional machine learning methods for the purpose of image classification. Most of the earlier studies for image classification implemented CNN models while the popular traditional machine learning model preferred by researchers was SVM. Moreover, IoT and its applications are now being leveraged in agriculture to support and enhance precision agriculture. Thus, taking all these factors into account, this research will make a comparison between ANN model that involves feature extraction, a Logistic Regression model with hyperparameter tuning using SGD classifier, and a CNN model with 12 layers.

3 Methodology

This research work implemented the ANN model, CNN model and a Logistic Regression model to crop and weed images to detect and classify weeds in soybean crops accurately as discussed in 2.6. This paper makes use of Knowledge Discovery in Databases (KDD) methodology.

The objective of this research is to find a machine learning method that will efficiently detect and classify weed images from crops. The research makes a comparison across three models namely Artificial Neural Network with feature extraction and hyper parameter tuning (ANN) model, a Convolutional Neural Network Model (CNN) with 12 layers in total and a Logistic Regression model with feature extraction and hyperparameter tuning. The model performances will be compared based on their accuracies which will determine which among the three is a better model for the purpose of image classification.

The figure below explains the KDD methodology with respect to this research project.

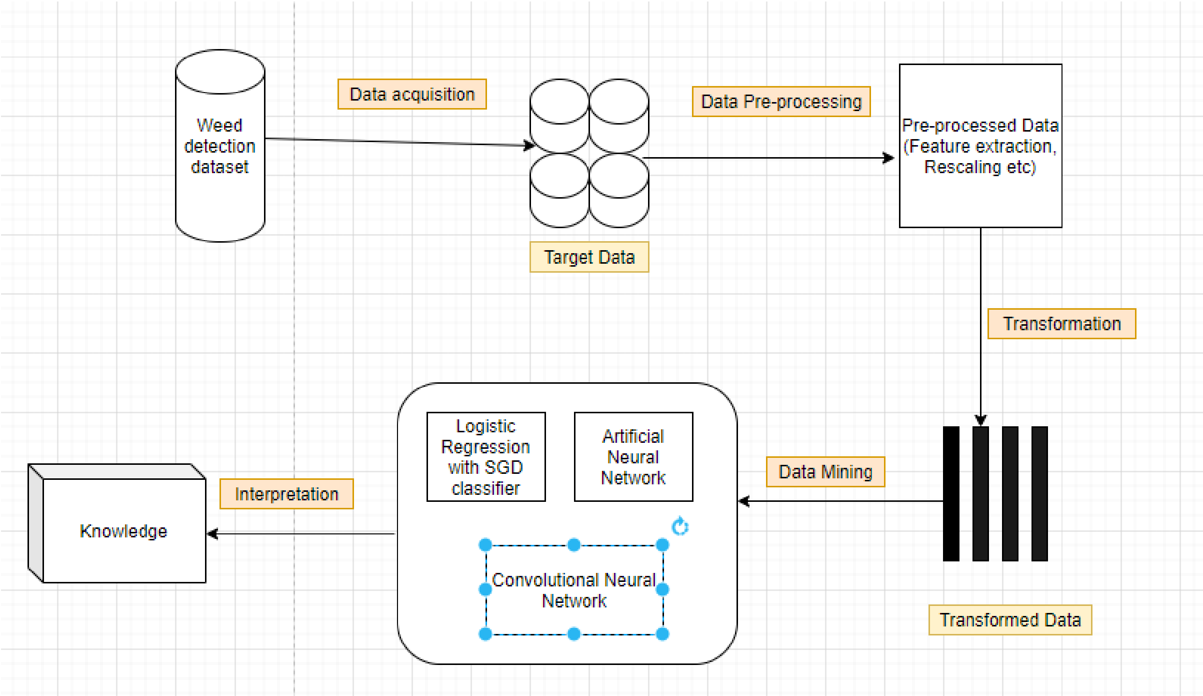

Fig 1. KDD for weed detection in soybean crops.

3.1 Dataset

This section gives a brief idea about the dataset and the operations performed on it to get the data processed and ready before implementing the models for classification

The data was obtained from a public repository Data for: Weed Detection in Soybean Crops Using ConvNets (2020) and the size of this dataset is 1.2GB. As mentioned in Ferreira et al. (2018), the dataset was captured using a UAV. A total of 400 images wherein all the occurrences of weeds were selected and by leveraging Pynovisão software, SLIC algorithm was used to segment these images were then marked out into their respective classes. These segments were used for a creation of a larger dataset that contained 4 classes of images such as broadleaf or weed, grass, soil, and soybean. This image dataset has 15336 image segments, out of which, 3249 belong to soil, 7376 are of soybean, 3520 images of grass and the remaining 1191 of broadleaf weeds.

3.2 Feature Extraction and Hyperparameter Tuning

3.2.1 Feature Extraction

Feature extraction is performed with a view to decrease the number of resources in a dataset to process data without losing relevant information from the data. Features such as shape, colour, texture, image orientation etc are available. However, for the purpose of this research, only the shape feature is extracted in both the models i.e. the ANN model and the logistic regression model. The extracted shape feature is shuffled and then the data is normalised by subtracting the mean image in order to make it indifferent to background changes.

3.2.2 Hyperparameter Tuning

In the logistic regression model, hyperparameter tuning is performed with the help of using the Stochastic Gradient Descent (SGD) classifier. The SGD classifier is used for hyperparameter tuning. Using the mini batch function, a new set of data points is created. The mini batches are nothing but looped over training samples. The arguments passed are epochs/iterations, loss function sigmoid and loglikelihood. The grad function computes the dot product. In case of each mini batch, the data is taken, and the dot product is computed between the data and the weight matrix followed by passing the obtained result through the sigmoid activation function.

3.3 Data pre-processing and Data Transformation

In this section, the important features of the data and the steps involved in pre-processing and transforming the data will be discussed.

3.3.1 Image Augmentation and Rescaling

Image augmentation is the most popular method of data augmentation which can be used to increase the size of the training by generating an altered version of images in dataset in an insecure manner. The method of image augmentation comprises of transforming the images present in the training data which are a member of the same class as the original image. The various transformation operations consist of zoom, rotate, shift, flip and other options. Image augmentation is not the same as data preparation which involves resizing the images or scaling the pixels as data preparation involves applying these in a uniform manner to all the datasets so as they can act as a link between the datasets.

This paper observes the use of image augmentation for the CNN model using the image generator. The image data generator is initiated and to increase the train data. The train set is flipped horizontally and has a rotation range of 40 with a width and height shift of 0.2 and a shear and zoom range of 0.2. After this step, a new training data set is generated which is rescaled to a size of 150*150 pixels with a batch size of 92 which dictates the number of images to be yielded from the generator for each batch. Categorical labels are required as the loss function used is categorical cross-entropy.

The validation and the test dataset are also rescaled according to the training set with images resized to a size of 150*150 with a batch size of 31 for both the directories.

3.3.2 Convolutional Neural Network

Once the data was acquired, then data pre-processing was performed. The dataset was divided into a training directory, test directory and a validation directory. The dataset was split as 70% for training directory while the test directory and validation directory contained 15% each. The images were restructured using the image data generator function and the images were reshaped and resized to a target size of 150*150. The images were also rotated with a slight shift in their height and width.

3.3.3 Artificial Neural Network

The data for the purpose of implementing the ANN model was divided into train data and test data in the ratio of 80% train data and 20% test data. The images were rescaled to a pixel size of 200*200. This enabled normalization of data by using mean image subtraction. This makes the dataset less sensitive to different backgrounds and lighting conditions. Feature extraction was also performed on the images. Using random permutations, the shape feature was extracted to transform the data and make it ready for applying the required model.

3.3.4 Logistic Regression

The data was divided into train data and test data in the ratio of 80% training data and 20% test data respectively for the purpose of implementing the logistic regression model. The images were rescaled to a pixel size of 50*50 for faster processing. Feature extraction was performed on the images with the help of random permutations. The shape feature of the images was extracted and was used to train the dataset. The data was also normalized by subtracting the mean image. This is done in order to reduce the sensitivity of the network to different backgrounds and lighting conditions.

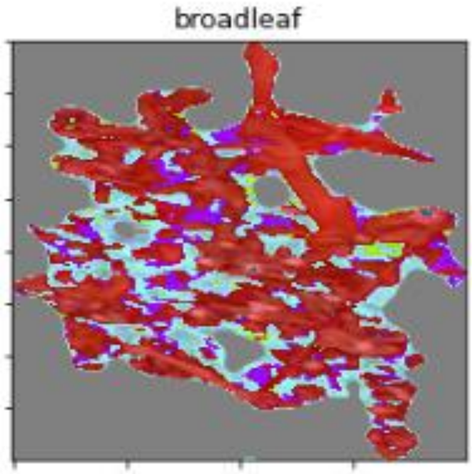

Fig. 2 Image after normalisation

3.4 Data Modelling

This section briefly discusses the implemented algorithms and models in order to classify the images. The use of traditional and deep learning models is observed in this research study. The implemented models are mentioned with the steps involved on

3.4.1 Convolution Neural Network (CNN)

The convolutional neural network model consists of convolutional layers, dense layers, and transition layers. The model consists of 12 layers in total in the CNN architecture. The convolutional layer is used to capture the information from the input. The dot product of the weights and the input image region is computed using the filter. The Rectified Linear Unit (ReLU) enhances the execution speed of the CNN and is therefore used as an activation function in the CNN. Deeper the network, higher is the complexity of the CNN. The MaxPooling layer is leveraged to shrink the dimension size of the image. The output layer includes a softmax for the purpose of classification. The softmax layer improves the interpretability of the model as it normalizes the exponential value to 1.

3.4.2 Artificial Neural Network (ANN)

The artificial neural network (ANN) is based on Artificial Intelligence techniques which enables the computer to perform tasks like classification, recognizing patterns, prediction etc. The ANN comprises of many neurons which are present in the ANN architecture throughout the input, output, and hidden layers. After data normalization is done by subtracting mean image, the ANN model is applied. It is first initiated as a function and the parameters such as input size, output size, hidden layer size etc are specified. The Rectified Linear Unit (ReLU) is used as an activation function.

3.4.3 Logistic Regression Model

The logistic regression is a classification algorithm. To implement this model, the hyperparameters and the batch size were defined. The Stochastic Gradient Descent (SGD) classifier in scikit-learn was used for hyperparameter tuning. The SGD optimizer takes parameters such as learning rates, batch size, epochs etc to train the models and are called hyperparameters. The loss parameter has been set to log as this classifier is used to implement a logistic regression model. SGD is an optimization technique found to be efficient for fitting a linear classifier for logistic regression that has a convex loss function.

4 Design Specification

4.1 Architecture Design

This section gives a description of the architecture and the flow of the research. The majority of these steps have been discussed in the section above i.e. section 3. This section will deal with the design specifications of the models. The further sections will talk about the implementation followed by the evaluation of results and the conclusion.

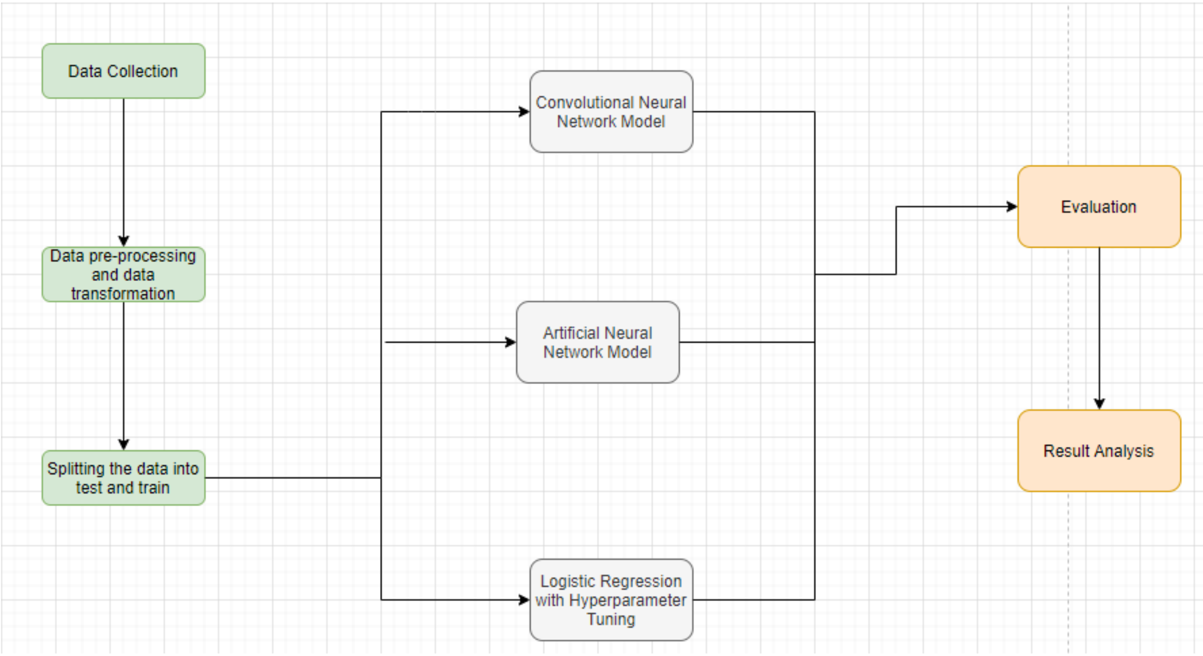

Figure 3: Architecture for classifying images of weeds, crops, soil, and grass

The dataset was extracted and then prepared for processing. The rescaling of images is different for different algorithms. The images are normalized by subtracting mean image for ANN and the Logistic Regression models. Once the data was ready, it was split into training and test data. For the CNN model, the dataset was first split into training, validation and test sets after which image augmentation was done before implementing the model. The models were judged on the basis of the accuracy, which enables to determine the best algorithm for the purpose of image classification.

5 Implementation

This section in brief, describes the working of the implemented models i.e. Convolutional Neural Network, Artificial Neural Network and Logistic Regression to detect weeds and classify them.

The implementation of the models for the purpose of classification was an uphill task. The experiment was performed on Google colaboratory with 15GB storage required on drive. Around 13GB of RAM space is needed and around 20GB of GPU during run-time. The CNN and the ANN models require a higher amount of time for execution due to the number of epochs and the specified layers. The use of GPU during runtime, ensures a faster execution of the models. To implement the models, it is necessary to have Keras and TensorFlow libraries of Python and also to have Python version 3. To get started, the dataset is first uploaded on google drive which is then mounted using Python API for accessing the dataset.

The dataset of size 1.2GB that consists of 15,336 images divided into 4 classes such as broadleaf, grass, soil, and soybean. The study classifies images and aims at successfully identifying weeds and differentiating them from crops.

5.1 Convolutional Neural Network (CNN)

For the convolutional neural network model, the data is first split into 70% training set, and 15% each into test and validation set. using Imagedata generator function, the images are rescaled to a pixel size of 150*150, with the width shift, height shift, zoom and shear range of 0.2.

The model consists of 12 layers in total. The convolution layer uses the ‘rectified linear unit’ (Relu) activation function. The input provided to the convolution layer is the dense block and filter size. The dropout value is set to 0.5. The dense layer is fully connected and placed before the output layer to build up the extracted information from the preceding layers. The model is compiled using model.compile() function with a categorical crossentropy loss function, ‘rmsprop’ optimizer at a default learning rate of 0.001. The plots of training and validation accuracies are plotted and also plots of loss history for training and validation are plotted.

5.2 Artificial Neural Network

The artificial neural network is a two layered model. Feature extraction is used, and the shape feature is leveraged for this purpose. Images are normalised by subtracting the mean image so that the final image is not differently sensitive to varying backgrounds.

The ANN model is defined as the function with all the parameters such as input size, output size, hidden layers, weights of all layers etc. specified in this function. The activation function is Rectified Linear Unit (ReLU) as the model takes lesser time to execute and has a faster convergence due to a cheaper computation cost. The input provided is the product of image height, image width and the channel, at a learning rate of 0.94. The plots of loss history for training and classification accuracies are also plotted so as to better understand the implemented model.

5.3 Logistic Regression

The Logistic Regression model involves the use of hyperparameter tuning using the Stochastic Gradient Descent (SGD) classifier. The data is pre-processed, and the images are scaled to a pixel size of 50*50 with the extracted shape feature and shuffling the training index. To reduce the sensitivity towards different backgrounds, image normalisation is done by subtracting the mean image. The idea behind the image size being scaled to 50*50 is to improve the data pre processing and execution time. Using the SGD classifier, mini batches are created and using these mini batches, the loss and the gradient is calculated. Parameters such as learning rate, input size, output size, batch size etc are fed to the model so as to obtain appropriate results.

6 Evaluation

This section discusses the thorough analysis of the mode of evaluation of the implemented models and also the evaluation metrics used to evaluate the models. The accuracy is calculated with the help of confusion matrix, which is available in the scikit learn library

The implemented models were evaluated on the basis of the accuracy and the loss per epoch for training and validation sets is also calculated along with loss history plots as well as the accuracies of training and validation sets.

The accuracy is calculated as:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

TP (True Positive): The weed images that are classified correctly as weeds by the model.

TN (True Negative): The crop images that are correctly classified as crop by the model.

FP (False Positive): The crop images that were wrongly classified as weed images by the model.

FN (False Negative): The images of weeds that were wrongly classified as crops by the model.

The CNN model underwent a different pre-processing as compared to the ANN and the Logistic Regression model with hyperparameter tuning. Different pre-processing techniques were used with an aim to obtain efficient results. The models were used for weed detection in soybean crops. The CNN model significantly. outperforms the ANN model and the Logistic Regression model with hyperparameter tuning. The accuracy obtained using the CNN model is 97% as opposed to 75% of the ANN model and 55% of the Logistic Regression model with SGD classifier. The below table summarizes the results of all the algorithms.

Table 1 Comparing model performance based on accuracy

Model	Accuracy
Artificial Neural Network	75%
Logistic Regression	55%
Convolution Neural Network	97%

Thus, the table above summarizes the results and we can conclude that CNN has comprehensively performed better than the other two models. Below are the plots for the loss history and the train and validation accuracies for all the implemented models.

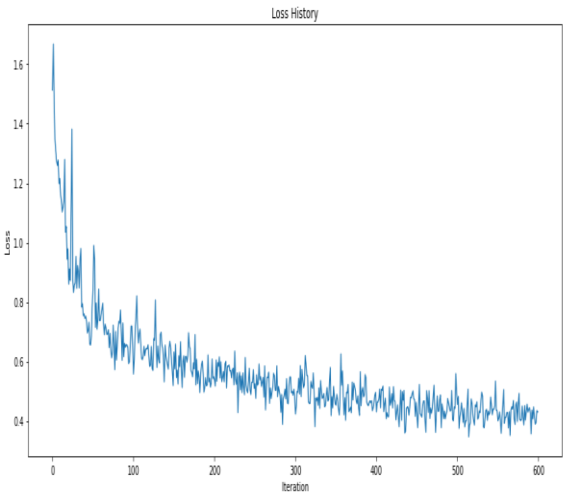

Figure 4 (a) Loss history for ANN model

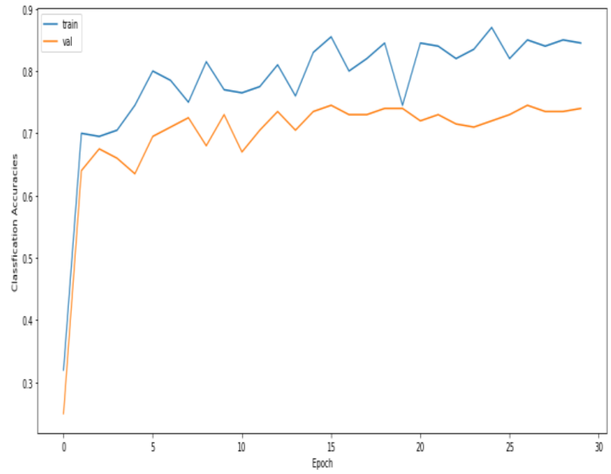

Figure 4 (b) validation accuracy for ANN model

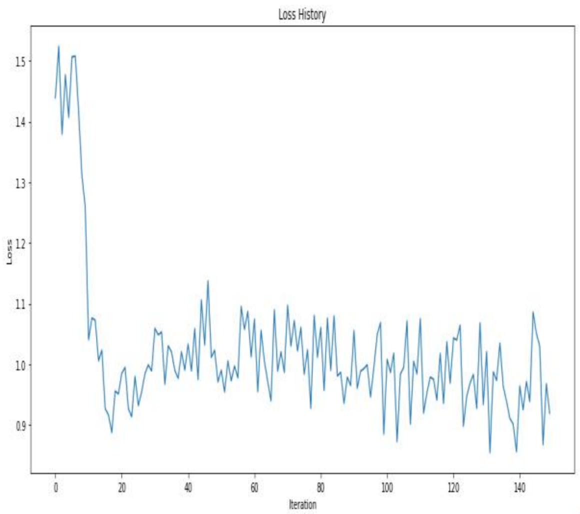

Figure 5 (a) Loss history for LR model

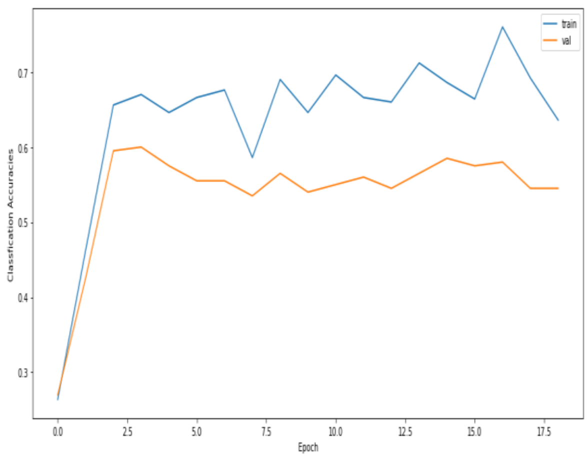

Figure 5 (b) train and validation accuracies for LR model

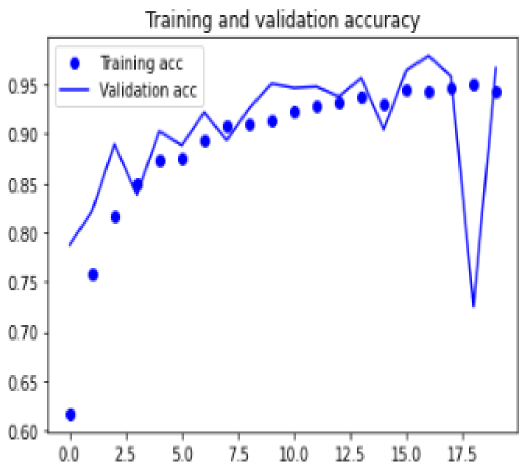

Figure 6 (a) Train and validation accuracies for CNN model

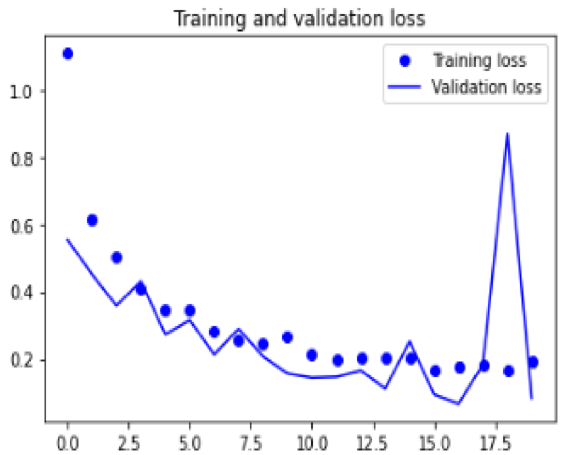

Figure 6 (b) Loss history for CNN model

6.1 Discussion

Thus, from the results discussed above, CNN model is better for the purpose of image classification with an accuracy of 97% while the ANN and Logistic Regression models have an accuracy of 75% and 55% respectively.

The most relevant study related to this research project was conducted by Ferreira et al. (2018). Their research accomplished better results for their CNN model with an overall classification accuracy across all classes higher than 99%. The reason behind obtaining the high classification accuracy could be the use of 8 layers in their implementation along with feature extraction that was performed by the authors which resulted in an extraction of 218 features such as shape, colour, texture etc completing the array of features. This research was aimed at developing a model for weed detection with minimal data pre-processing and feature extraction. The images were rescaled to a pixel size of 256*256 as opposed to 150*150 in this research.

Bakhshipour et al. (2017) implemented ANN and obtained an accuracy of 93.1% for crop-weed discrimination. The reason behind this maybe the use of PCA for dimension reduction and due to the use of wavelet features, removing soil and residues from the images, the extraction of texture coefficients and segmentation for data pre-processing, whereas this research study only extracted the shape feature for pre-processing purposes.

The research study presented, tried to extract feature as well as tune parameters to enhance the performance of models. However, looking at the results, it is safe to say that the Convolutional Neural Network model is the best method for image classification. Decreasing the pixel size of images reduces the execution time. Therefore, the Logistic Regression model with SGD classifier takes the least amount of time while the CNN model takes the maximum tie amongst the three models.

7 Conclusion and Future Work

To ensure a better quality and quantity of yield, it is necessary to identify and tackle the issue of weeds in farmlands while also minimizing the use of herbicides to reduce damage to the crops and the environment. In this research, the aim was to detect weeds in soybean crops using image classification techniques. This research was implemented on a dataset that consisted of more than 15000 images of soil, grass, soybean, and broadleaf weeds. Neural network models such as Convolutional Neural Network, Artificial Neural Network, and a classifier such as Logistic Regression with SGD classifier were implemented. Amongst the implemented models, the CNN model could classify the images with 97% accuracy and fares substantially better than the other two models.

The CNN model outperformed other models by achieving an accuracy of 97%, in comparison with an ANN model that achieved an accuracy of 75% while the Logistic Regression model achieved an accuracy of 55%. The use of ConvNets facilitates the faster processing and training of large datasets due to its processing power. Thus, in conclusion, CNN is best suited for image classification problems.

For the future, researchers can try tuning more parameters for other algorithms like ANN or Logistic Regression or extract more features. However, the integration of CNN in agri-robots or integrating CNN model with IoT for automating the process of weed detection, classification and elimination is a scope worth looking at with the necessary hardware and software resources.

Another application for the future can be the use of transfer learning-based models on this dataset along with fine tuning the parameters to achieve a higher accuracy. Various image enhancement techniques and increasing or decreasing the number of epochs can also be done to check for improvements in the performance of the models.

8 Acknowledgement

The accomplishment of this research project was possible due to the guidance, constant support, and constrictive feedback of Mr. Hicham Rifai. He played an important part in helping me to implement this research and thus complete my thesis. I would also like to express my gratitude to my family members and my loved ones for all the financial and moral support needed to through this master's course.

References

Grand challenges in weed management <https://www.frontiersin.org/articles/10.3389/fagro.2019.00003/full/> (last visited on July 27, 2020)

Llewellyn, R. S., Ronning, D., Ouzman, J., Walker, S., Mayfield, A., and Clarke, M. (2016). *Impact of Weeds on Australian Grain Production: The Cost of Weeds to Australian Grain Growers and the Adoption of Weed Management and Tillage Practices*. Report for GRDC. *CSIRO*, 112.

Gharde, Y., Singh, P., Dubey, R. and Gupta, P., (2018). Assessment of yield and economic losses in agriculture due to weeds in India. *Crop Protection*, 107, pp.12-18.

International: World soybean production, (2019). Soystats, <http://soystats.com/ international-world-soybean-production/> (last visited June 28, 2020).

Rizzardi, M.A. and Fleck, N.G. (2004). Métodos de quantificação da cobertura foliar da infestação de plantas daninhas e da cultura da soja. *Ciência Rural*, 34(1), pp.13–18.
Colin, W.W., Harold C., Koushik S., Jonathan F. (2016). *Encyclopedia of food grains.* Elsevier Academic Press.

Herrera, P., Dorado, J. and Ribeiro, Á. (2014). A Novel Approach for Weed Type Classification Based on Shape Descriptors and a Fuzzy Decision-Making Method. *Sensors*, 14(8), pp.15304–15324.

Dankhara, F., Patel, K. and Doshi, N., (2019). Analysis of robust weed detection techniques based on the Internet of Things (IoT). *Procedia Computer Science*, 160, pp.696-701.

Leonard, E.C., (2016). Precision agriculture. Wrigley, C., Corke, H., Seetharaman, K., Faubion, J. (Eds.), Encyclopedia of Food Grains, vol. 4, Elsevier, Oxford, pp. 162–167.

Bucci, G. *et al.* (2019) 'When accuracy of measurements matter: economic profitability from precision agriculture', *2019 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Metrology for Agriculture and Forestry (MetroAgriFor), 2019 IEEE International Workshop on*, pp. 237–241. doi: 10.1109/MetroAgriFor.2019.8909240

Sciencedirect.com. (2020). *Image Processing - An Overview | Sciencedirect Topics*. [online] Available at: <https://www.sciencedirect.com/topics/engineering/image-processing> [Accessed 7 August 2020].

Anbarjafari, G., 2020. *1. Introduction to Image Processing*. [online] Sisu.ut.ee. Available at: <https://sisu.ut.ee/imageprocessing/book/1> [Accessed 7 August 2020].

Desktop.arcgis.com. (2020). *What Is Image Classification? — Arcgis Help | Documentation*. [online] Available at: <https://desktop.arcgis.com/en/arcmap/latest/extensions/spatial-analyst/image-classification/what-is-image-classification-.htm> [Accessed 9 August 2020].

Homepages.inf.ed.ac.uk. (2020). *Image Analysis - Classification*. [online] Available at: <https://homepages.inf.ed.ac.uk/rbf/HIPR2/classify.htm> [Accessed 8 July 2020].

Hung, C., Xu, Z. and Sukkarieh, S., (2014). Feature Learning Based Approach for Weed Classification Using High Resolution Aerial Images from a Digital Camera Mounted on a UAV. *Remote Sensing*, 6(12), pp.12037-12054.

Ahmed, F., Al-Mamun, H., Bari, A., Hossain, E. and Kwan, P., (2012). Classification of crops and weeds from digital images: A support vector machine approach. Crop Protection, 40, pp.98-104.

Herrera, P., Dorado, J. and Ribeiro, Á. (2014). A Novel Approach for Weed Type Classification Based on Shape Descriptors and a Fuzzy Decision-Making Method. *Sensors*, 14(8), pp.15304–15324.

Saha, D., Hanson, A., Shin, S.Y., (2016). Development of enhanced weed detection system with adaptive thresholding and support vector machine. Proceedings of the International Conference on Research in Adaptive and Convergent Systems, pp. 85–88.

Tellaeche, A., Pajares, G., Burgos-Artizzu, X.P., Ribeiro, A., (2011). A computer vision approach for weeds identification through support vector machines. Appl. Soft Comput. 11 (1), 908–915.

Siddiqi, M.H., Lee, S., Kwan, A.M., (2014). Weed image classification using wavelet transform, stepwise linear discriminant analysis, and support vector machines for an automatic spray control system. J. Inform. Sci. Eng. 30.

Beibei, C., Eric. T. Matson, (2015). A feature-based machine learning agent for automatic rice and weed discrimination. International Conference on Artificial Intelligence and Soft Computing, pp 517–527.

S. Murawwat, A. Qureshi, S. Ahmad & Y. Shahid (2018) ‘Weed Detection Using SVMs’, *Engineering, Technology & Applied Science Research*, (1), p. 2412. Available at: http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib&db=edsdoj&AN=edsdoj.80861e534de495d8f31a1ca8613b&site=eds-live&scope+site (Accessed: 23 March 2020).

Kamath, R., Balachandra, M. and Prabhu, S., (2020). Paddy Crop and Weed Discrimination: Multiple Classifier System Approach. *International Journal of Agronomy*, 2020, pp.1-14.

dos Santos Ferreira, A., Matte Freitas, D., Gonçalves da Silva, G., Pistori, H. and Theophilo Folhes, M., (2017). Weed detection in soybean crops using ConvNets. *Computers and Electronics in Agriculture*, 143, pp.314-324.

Ishak, A., Hussain, A. and Mustafa, M., (2009). Weed image classification using Gabor wavelet and gradient field distribution. *Computers and Electronics in Agriculture*, 66(1), pp.53-61.

Da Costa, A.Z., Figueroa, H.E.H. and Fracarolli, J.A. (2020). Computer vision based detection of external defects on tomatoes using deep learning. *Biosystems Engineering*, 190, pp.131–144.

Olsen, A., Konovalov, D., Philippa, B., Ridd, P., Wood, J., Johns, J., Banks, W., Girgenti, B., Kenny, O., Whinney, J., Calvert, B., Azghadi, M. and White, R., (2019). DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning. *Scientific Reports*, 9(1).

Abouzahir, S., Sadik, M. and Sabir, E. (2018) 'Enhanced Approach for Weeds Species Detection Using Machine Vision', *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), Electronics, Control, Optimization and Computer Science (ICECOCS), 2018 International Conference on*, pp. 1–6. doi: 10.1109/ICECOCS.2018.8610505.

M. Dian Bah *et al.* (2018) 'Deep Learning Based Classification System for Identifying Weeds Using High-Resolution UAV Imagery', *Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2*. (Advances in Intelligent Systems and Computing. 857), p. 176. doi: 10.1007/978-3-030-01177-2_13.

Bakhshipour, A., Jafari, A., Nassiri, S. and Zare, D., (2017). Weed segmentation using texture features extracted from wavelet sub-images. *Biosystems Engineering*, 157, pp.1-12.

Liang, W.-C., Yang, Y.-J. and Chao, C.-M. (2019) 'Low-Cost Weed Identification System Using Drones', *2019 Seventh International Symposium on Computing and Networking Workshops (CANDARW), Computing and Networking Workshops (CANDARW), 2019 Seventh International Symposium on, CANDARW*, pp. 260–263. doi: 10.1109/CANDARW.2019.00052.

T., S., G.S., S.G., S., S., Kumaraswamy, R. (2019) 'Performance Comparison of Weed Detection Algorithms', *2019 International Conference on Communication and Signal Processing (ICCSP), Communication and Signal Processing (ICCSP), 2019 International Conference on*, pp. 0843–0847. doi: 10.1109/ICCSP.2019.8698094.

Asad, M. and Bais, A., (2019). Weed detection in canola fields using maximum likelihood classification and deep convolutional neural network. *Information Processing in Agriculture*.

Gao, J., French, A., Pound, M., He, Y., Pridmore, T. and Pieters, J., (2020). Deep convolutional neural networks for image-based Convolvulus sepium detection in sugar beet fields. *Plant Methods*, 16(1).

Kiani, F. and Seyyedabbasi, A., (2018). Wireless Sensor Network and Internet of Things in Precision Agriculture. *International Journal of Advanced Computer Science and Applications,* 9(6).