National College of Ireland

Detecting Diseases in Rice Leaf Using Deep Learning and Machine Learning Techniques

MSc Research Project
MSc Data Analytics

Shubham Raje
Student ID: x20132158

School of Computing
National College of Ireland

Supervisor: Dr. Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Shubham Raje
Student ID:	x20132158
Programme:	MSc Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Majid Latifi
Submission Due Date:	16/08/2021
Project Title:	Detecting Diseases in Rice Leaf Using Deep Learning and Machine Learning Techniques
Word Count:	
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:	Shubham Suresh Raje
Date:	23rd September 2021

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):	

Detecting Diseases in Rice Leaf Using Deep Learning and Machine Learning Techniques

Shubham Raje
x20132158

Abstract

Plant diseases have a serious affect on the farming industry. Because of these diseases there is a bad impact on the productivity of the crops. This leads to huge losses to farmers. To ensure better quality, quantity and productivity of the yield, it is very crucial for identifying the diseases at early stage for reducing the use of pesticides to reduce damage of the crops and environment. In this research the motive was to detect and classify the diseases in the rice leaf, having four categories of classes as healthy, hispa, brown spot and leaf blast. In this research study convolutionl neural network was used for the feature extraction from the rice images. Whereas, some machine learning classifiers such as Random Forest and K-Nearest Neighbors were used for the classification of the diseases based on the categories. The first model CNN performed well for the feature extraction with the accuracy of 80 percent. Along with this second model was classification of diseases using some machine learning classifiers such as Random Forest and K-Nearest Neighbors, accomplished the accuracy of 96% and 72% respectively.

Keywords - Deep Learning, Machine Learning, CNN, Random Forest, KNN, IOT, Image Processing.

Contents

1 Introduction

Agricultural sector plays a crucial role for economic development of any country. In terms of raw materials, the majority of countries are dependent on agricultural goods. Rice is mostly cultivated crop around the globe. Rice is cultivated in over 100 countries around the world. A total of 158 million hectares are harvested each year, yielding more than 700 million tons of rice. In comparison to other continents, Asia produces the majority of rice [1]. Because of increasing population it is a ecting environment in terms of global warming, rapid climatic shifts Yadav et al. (2021). The agricultural sector is su ering as a result of these changes in the environment.

Crops are becoming infected with a variety of illnesses as a result of environmental changes. This has a signi cant impact on crop quality, quantity, and productivity. Different forms of illnesses are wreaking havoc on rice elds these days, having a negative e ect on crop production around the world. Many illnesses have been seen in recent years, including rice leaf blast, brown spot, Hispa, rice curl disease, and many others (Jiang et al.; 2020). These diseases can be found on any part of the rice plant, including the leaf, neck, and ear. Plant disease has a negative impact on not just agricultural productivity but also on the environment in terms of pollution. Plant diseases are responsible for 10 to 15% of total productivity losses. In the worst-case scenario, farmers could lose up to 50% of their crop, which is a signi cant loss for farmers and the country's economy (Tian et al.; 2021). As a result, it is critical to detect a plant disease early on in order to ensure sustainable and accurate agriculture and to avoid waste of nancial and other resources. As a result, early detection of pests on crops is critical for avoiding large use of fertilizers and pesticides for obtaining higher productivity. For large-scale crops, naked-eye participant observation are not practical nor su cient. Advancement in IT eld, help to increase productivity of the crops with the minimum use of fertilizers. In today's Arti cial Intelligence (AI) environment, convolutional neural networks (CNN) and machine learning can play important role in classifying diseases.

The Image Processing domain may be able to provide a solution to the agricultural sector's challenges. CNN will be used for extracting features from the leaf images. Diseases can be classi ed using machine learning classi ers. CNN has a faster processing speed and is more accurate in categorization. This approach could be useful for disease classi cation and detection.In this study, a convolutional neural network is utilized to extract features from rice photographs. For classifying diseases, machine learning classi ers such as random forest and K-Nearest Neighbors are utilized.

1.1 Research Question

How well can convolutional neural networks, Random Forest and K-Nearest Neighbors help in the detecting and classifying diseases in rice leaf?

1.2 Research Objective

The agriculture industry now supports the majority of countries. The agricultural industry is equally important for a country's economic growth. However, as the population grows, the environment is being impacted more and more. The agriculture industry is

[1] https://en.wikipedia.org/wiki/Rice

su ering as a result of the severe weather. The crops are infected with a variety of illnesses. Early disease identi cation is critical in order to avert output declines and farmer losses. This can be accomplished with the use of data analytics. Main motive of the research is to identify the diseases in rice leaf using CNN and machine learning classi ers.

1.3 Research Report Flow

- Literature Review: In Section 2, literature review for previous researches related to this eld is explained in detail.
- Research Methodology: Knowledge Discovery Databases approach is used in this research work. Steps of this methodology is well explained related to this research work and which is addressed in Section 3.
- Design Speci cation: Overall design of the implementation of research work is presented in the Section 4
- Implementation: A brief summary of how research work is implemented is well explained in the Section 5
- Evaluation: Evaluation metrics which are examined during the implementation of the work are addressed in Section 6
- Conclusion and Future Work: Summary of the research work is summarised with the future work and which is presented in the Section 7

2 Related Works

In the past, various studies in the eld of image processing which were conducted are analysed in this section. Di erent methodologies and approaches for detecting rice leaf disease using Image Processing are reviewed in this literature review. By considering multiple algorithmic models, this paper also considers the impact of Machine Learning and Deep Learning methodologies.

2.1 Neural Network methods

The usage of DCNN (Deep CNN) with a Bayesian learning process for disease categorization in diverse plants is depicted in research undertaken by Sachdeva et al. (2021). The Bayesian Learning approach is used at the top of the board to enhance the pixel dependency. The study used a total of 20,639 photos from the Plant Village database, including 15 di erent classi cations of unhealthy and healthy leaf images. According to the ndings of this study, this model is an e ective tool for detecting and classifying disease in plants. CNN's proposed solution, which employs the Bayesian learning method, achieves a 98.9% overall accuracy with no symptoms of over- tting.

The study Zhang et al. (2021) used R-CNN to investigate the identi cation of soybean fungal diseases from a synthetic image. For the sustainability of soybeans and the economy of agribusiness, accurate diagnosis of soybean plant disease is critical. Many studies have been conducted in this subject, but due to a lack of data and technical challenges, disease identi cation in soybeans is more challenging. This research focuses on

creating a synthetic soybean plant leaf image database to address the issue of a limited database. When model was implemented, it produced an accuracy rate of 83.34 percent. Furthermore, the experimental ndings generated with the synthetic dataset by the MF3 R-CNN model were successful in detecting soybean plant diseases, despite the higher complication. A comparison of the baseline model with the MF3 R-CNN model reveals that the MF3 R-CNN model outperforms the base model.

2.2 Machine Learning methods

Deep feature oriented rice crop diseases detection using Support Vector Machine (SVM) was implemented in the research conducted by Sethy et al. (2020). Because features are vital for picture categorization in the eld of Machine Learning, nding the optimum approach for feature extraction is critical. In total, 5932 on- eld pictures of rice plant disease such as blast, brown spot, tungro, and bacterial blight were included in the study. This study measured the e ectiveness of 11 CNN models in a Transfer Learning and Deep Learning approach, as well as SVM. Small CNN models like shu enet and mobilenetv2 were also put to the test. Speci city, Accuracy, Sensitivity, False Positive Rate, F1 Score were used to evaluate performance. A statistical study was used to determine which classi cation model was the best. With an F1 score of 0.9838 and a training time of 69 seconds, the ResNet50 plus SVM model performed well. The deep feature of mobilenetv2 with SVM produced a result with an F1 score of 0.9796 and a training time of 48 seconds in tiny CNN models. In comparison to the fc7 and fc8 layers, the fc6 layer of vgg16 and vgg19 with AlexNet produced better classi cation results. The CNN classi cation model's F1 score was also compared to classic image classi cation models such as local binary patterns plus SVM and histogram of directed gradients plus SVM.

The study done by Mojjada et al. (2020) used digital image processing to diagnose plant leaf illnesses. Agriculture productivity is strongly vital for economic growth. As a result, accurate identi cation of plant disease is critical, as fungal diseases is a natural occurrence that can occur at any time. If pine trees are not properly cared for, they will have a signi cant impact on plants, as well as the amount, quality, and productivity of their outputs. In the United States, a disease known as tiny leaf disease is an extremely hazardous illness that a ects pine trees. The use of automated systems will be extremely bene cial in detecting illness indications at an early stage on farm locations. For the automatic identi cation and categorization of plant leaf illness in pine trees, a segmentation technique is applied. Surveys of various strategies for classifying sickness are conducted as part of this study. The segmentation of photos is done using genetic algorithms, which is crucial for detecting plant diseases. K-means and Support Vector Machine are used to segment and screening and diagnostic in this study. The classi cation accuracy of this approach is estimated to be 75%. With a larger high-quality image database, this can be enhanced.

The study Kaur and Devendran (2020) focuses on using a Support Vector Machine classi er to identify plant leaf illness using segmentation and law mask feature extraction. Optimization-based segmentation and a law mask framework was used to solve the problem of crop diseases classi cation with many classes is the subject of this paper. In this work, however, SVM is used as a classi er. The classi cation accuracy is used to evaluate performance. Precision and recall are calculated as evaluation measures for categorization success. Various ltering methods, such as the Gaussian lter, median lter, and average lter, were used to pre-process the images. K-means clustering

techniques are also employed in image segmentation to group objects based on multiple attributes. The extraction of characteristics is critical for identifying objects. The images' characteristics are extracted after segmentation. Next, K-nearest neighbor (KNN), support vector machine (SVM), Fuzzy logic-based, neural network (NN) is implemented for disease classi cation and detection, with better results.

Employing computer vision and Machine Learning approaches, the research Chouhan et al. (2021) studied on leaf disease segmentation and classi cation in Jatropha Curcas L. and Pongamia Pinnata L., two biofuel plants. Biofuel made from plant extracts such as Jatropha Curcas L. and Pongamia Pinnata L. is in high demand. However, plant growth is hampered by biotic factors, which limits yield. In this research, computer vision methodology was developed for an automated disease diagnosis system. Following that, a hybrid Neural Network with super pixel clustering is used to segment the illness region area. SIFT and LBP algorithms were used to evaluate shape, color, and features. For the categorization of illnesses in plant leaves, Machine Learning classi cation techniques such as Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbor, Classi cation and Regression Trees, Random Forest, Nave Bayes, and Support Vector Machine were employed. The accuracy of Logistic Regression was 0.9857, and the accuracy rate of Random Forest was 0.9095. In comparison to other models, these two models performed well throughout the research.

2.3 Deep Learning Classification methods

The research Atila et al. (2021) focused on the classi cation of plant leaf diseases using the E cientNet Deep Learning Model. Because certain classic Machine learning algorithms do not provide superior results, the E cientNet deep learning method is used in this work to classify plant leaf diseases. This model's results were compared to that of other deep learning models. The model was trained using the PlantVillage database. Original and augmented database of 55,448 and 61,486 photographs, respectively, were used to train the models. Transfer learning was used to train the E cientNet model and some other deep learning models. In comparison to other models, the B4 and B5 models fared well in the E cientNet architecture. On the enhanced dataset, the B4 model had a precision value of 99.39 percent and an accuracy of 99.97 percent. In the data set, the B5 model has an accuracy of 99.91 percent and a precision value of 98.42 percent.

In this work Sujatha et al. (2021), the e ectiveness of Deep Learning vs. Machine Learning approaches in detecting plant leaf illness is examined. Diseases can strike the plant at any stage during its life cycle. As a result, crop output and selling price are reduced. As a result, in the agricultural sector, identifying crop disease is critical. In this study, multiple Machine Learning and Deep Learning techniques for disease detection are designed and evaluated. For citrus plant disease diagnosis, machine learning techniques such as Support Vector Machine, Random Forest, Stochastic Gradient Descent, and deep learning methods such as Inception-v3, VGG-16, and VGG-19 were trained. Since Deep Learning approaches outperformed Machine Learning methods, the categorization accuracy was fairly excellent. The model's performance is displayed as follows: VGG-19:87.4 percent, Inception-v3:89 percent, VGG-16:89.5 percent, RF:76.8 percent, SGD:86.5 percent, SVM:87 percent, VGG-19:87.4 percent, Inception-v3:89 percent, VGG-16:89.5 percent As a result of the results, I can conclude that RF provides the least accuracy. VGG-16, on the other hand, has the highest classi cation accuracy.

The study Hu et al. (2021) looked into the detection and severity analysis of tea fungal

diseases using deep learning. When machine learning image processing techniques were used to detect tea leaf blight, the ndings were poor and inaccurate. As a result, deep learning methods are used in this research to improve disease classi cation accuracy. The Retinex algorithm is used to improve the original image data as well as reduce light uctuation and shadow. For better detection of blurred, occluded, and small bits of leaves, a Faster Region-based Convolutional Neural Networks model was used. The detected images are fed into the VGG16 trained network, which produces the best results. Overall, the ndings of this deep learning technique fared well when compared to typical machine learning algorithms, according to the report.

Table 1: Related Works Summary

Author	Methods	Objective	Advantage	limitations
Chouhan et al. (2021)	Neural Network, Machine Learning	Detection of disease in biofuel plants	Provided better classi cation accuracy	Lower learning rate in segementation
Bao et al. (2021)	Metric Leraning, Computer Vision	Wheat leaf disease detection	Good classi cation accuracy	Identi cation accuracy is less for metric learning
Azadbakht et al. (2019)	Machine Learning	Wheat leaf rust detection	ML model showed better identi cation	Spectral vegetation showed less accuracy
Sujatha et al. (2021)	SVM, Random Forest	Disease detection	Provided solution	Less accuracy for algorithms
Tian et al. (2021)	Machine learning	Rice leaf blast detection	Accuracy provided is good	Few infections are not detected
Jiang et al. (2020)	CNN, SVM	Disease detection with machine learning and deep learning	SVM provided good accuracy	Deep learning model showed less accuracy

3 Methodology

It is vital to choose the suitable approach for completing the project work properly at the beginning of any research project. After selecting the methodology for the research work the stages should be followed properly in order to get better solution. This research project follows Knowledge Discovery in Database (KDD) methodology for acquiring the solution to the problem. Although this methodology is very vast, it goes well with the image processing domain.

The objective of this research is to build a deep learning and machine model which will detect and classify the diseases in rice crop using the database. In this research Convolutional Neural Network (CNN) is used to extract the features from the rice leaf images. Further some machine learning classi ers such as Random Forest and K-Nearest Neighbors will classify the disease in which class it belongs. This is kind of hybrid model.

For this implementation of whole research work I am using this KDD methodology and the gure 1 shows the process ow of KDD approach:

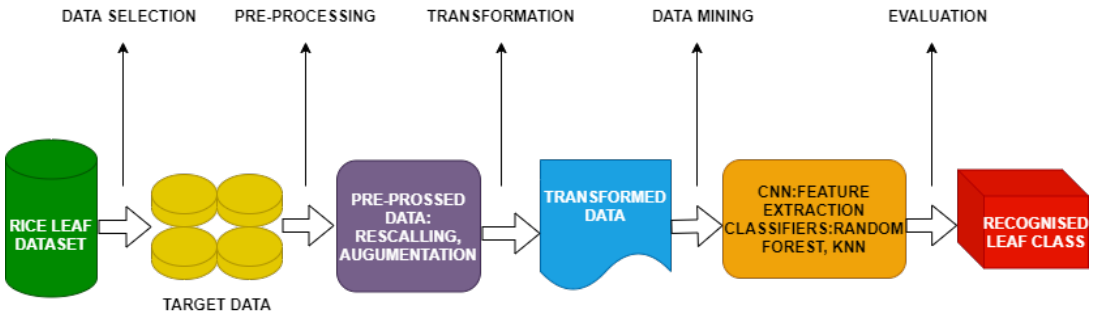

Figure 1: Proposed Methodology

3.1 Data Selection

The database for rice leaf diseases was obtained from the web and is available on kaggle.com[2]. Farmers in Gujrat, India, were consulted to construct the database. The photos were taken in direct sunshine against a white background. Farmers also shared details on the illnesses that a ect rice leaves. In addition, the photos were divided into four groups based on the disorders. Brown spot, Hispa, Leaf Blast, and Healthy are among the four categories. There are 3355 photos in the collection, all of which are in.jpg format. The dataset is 7 GB in size due to a high resolution photos. Further data set was divided into training and validation for each and every class. In total 2684 images are kept for training and 671 images were kept in validation phase. This dataset was generated for the aim of study and can be used to identify and classify rice leaf diseases. This data selection method completes the image acquisition process. The data is kept in the repository after the photographs have been collected. This information can be used in a subsequent step. A data distribution graph was evaluated based on the four categories. The gure 2 shows the data distribution visual based on four labels i.e. Categories available in the database.

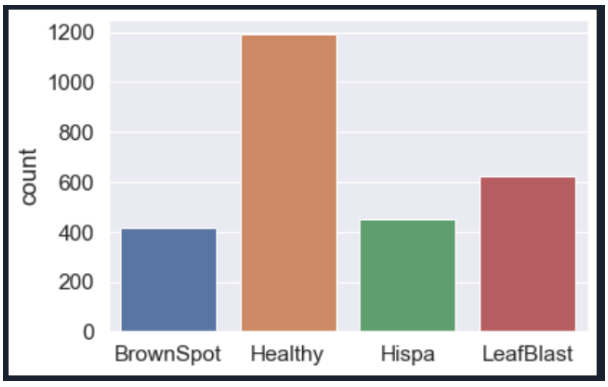

Figure 2: Data Distribution

[2]https://www.kaggle.com/shayanriyaz/riceleafs

3.2 Data Preparation

This section provides steps involved in pre-processing and transformation of data in this research work. This section is divided into following two sub-sections:

3.2.1 Data Pre-processing

This step involves working on database before the implementation phase. For any research work it is crucial to pre-process the data before using for any modelling purpose. In this research project normalization is done on images to use the images in proper format which will help to do modelling faster. Because the original size of images are in di erent size. The main advantage of image resizing is to train the model fast because model trains faster on smaller image size. The original size of images are 1881*1881 pixels and it is resized to 256*256.

3.2.2 Data Transformation

In data transformation process image augmentation is carried out. Image augmentation is one of the most popular method of data augmentation, can be used to increase the size of training by using resized version of images in dataset. Image augmentation process comprises of transforming the pictures which are present in the training data. In this research work image augmentation for CNN is done using image data generator. This process involves various transformation operations such as rotate, shift, zoom, ip and many more. Image augmentation is not similar as data pre-processing which involves re-scaling of images but this process involves certain kind of transformations on images so further it can be used for modeling.

In this research work various steps are done in image augmentation process on training data for CNN model. Initially in this process batch size is set for whole dataset. Batch size is one of the important hyperparameters to tune deep learning models. Batch size is set to 32, which provides 32 images for each batch. Batch size should de ned appropriately for every model. If size is more or less then the it leads to poor generalization. After batch size is set rotation is done on every image. The rotation range given to images is 30 degree clockwise direction. Rotation of image is very important because when the image is rotated there are some chances that pixels are out of range and with the rotation process each and every pixel is fetched. Image ip is another important parameter. Here vertical ip is done on images. Vertical ip is the rotation of image vertically. After all this transformation I will get the transformed data. All this transformations are done on training data and this transformed train data is further passed to the CNN model for model training.

3.3 Data Modeling

This section gives explanation about the algorithms and models that were used to classify the pictures. This research work shows the usage of both classic machine learning and deep learning methods. The models that have been implemented are listed along with the processes involved.

3.3.1 Convolutional Neural Network

The convolutional neural network model works well with the image processing and gives better accuracy and builds a e cient model. So, this reserch work has implemented CNN model for feature extraction from rice leaf images. The CNN architecture consists of dense layer, convolutional layer, max pooling layer, attening layer and dropout layer. The CNN model of this research work have this ve layers. CNN model is used for extracting features from the rice leaf images. In total 16 and 32 lters are added in the convolutional layers for ltering the images. Recti ed Linear Unit (ReLU) is used in the layers to enhance the execution speed of the model, which is used as an activation function in this model. Dense layer is used to connect all the layers with each other. Max pooling layer provides the feature map from the images. Flatten layer will convert the data in 1-dimensional array. The output layer contains softmax function which is used for classi cation of images but in this research work I am not classifying the images in the output layer of the CNN model. Here just I am using Features i.e feature vector from the images which are generated CNN model as proposed. Weights and features generated from the model will be saved and further this feature vector is used in embeddings process, which is further explained in the next subsection.

3.3.2 Embbedings

The features retrieved from the leaf photos are represented by an embedding, which is a vector. The vectors created for other leaf pictures can then be compared to this. Another vector that is close to the rst could be the same leaf class, while another vector that is remote could be a di erent leaf class. The classi cation model will accept an embedding as inputs and predict the leaf's identi cation. I can use the trained model to pre-process a leaf image to generate an embedding that can be saved and used as input to our classi er model, or I can use the trained model to pre-process a leaf image to create an embedding that can be stored and used as input to our classi er model. This embedding create an array and generates feature vector. This feature vector is compared with the feature vector from CNN model and stores the features in the particular label i.e. into four categories. Further this output is given to the classi ers as an input and then it will classify the images according to the four categories present in the database.

3.3.3 Random Forest

Random Forest classi cation algorithm is implemented for classi cation of rice leaf images. This classi cation algorithm is used because it give better accuracy with the CNN model and work better in classi cation problem. Scikit learn metrics used for calculation of evaluation metrics such as accuracy, confusion matrix, precision, recall and F1-score. This classi er is classifying the leaf images based on which class it belongs.

3.3.4 K-Nearest Neighbors

K-Nearest Neighbors classi cation algorithm is implemented for classi cation of rice leaf images. This algorithm is used because it performs better in classi cation task and work well with deep learning model. Scikit learn metrics used for calculation of evaluation metrics such as accuracy, confusion matrix, precision, recall and F1-score. This classi er is classifying the leaf images based on which class it belongs.

3.4 Evaluation

Accuracy, precision, recall and F1-score are some of the evaluation measures which are considered in this research study for measuring the performance of the models. Accuracy measure is checked for both CNN and classification models. Rest others are checked for classification algorithms i.e. Random Forest and KNN. Below are some of the evaluation measures which are evaluated for checking the effectiveness of the models[3].

- Accuracy: Accuracy is the percentage of correct prediction done on test data. It is calculated by dividing correct number of prediction by total number of prediction which as has been given as input.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

- Precision: Precision is total number of positive results which is divided by total number of positive results which are predicted by the classifier.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

- Recall: Recall is defined as total number of positive results which is divided by all of the results which are predicted by classifier.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

- F1-score: F measure helps to calculate precision and recall at the same time. This value is mean of precision and recall.

$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

4 Design Specification

The architecture and flow of the research is described in this section. The majority of these processes have already been described in Section 3. The design specifications of the models will be discussed in this section. The implementation will be discussed in the following parts, followed by the evaluation and the conclusion. The Figure 3 shows the process flow on the implementation.

[3]https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

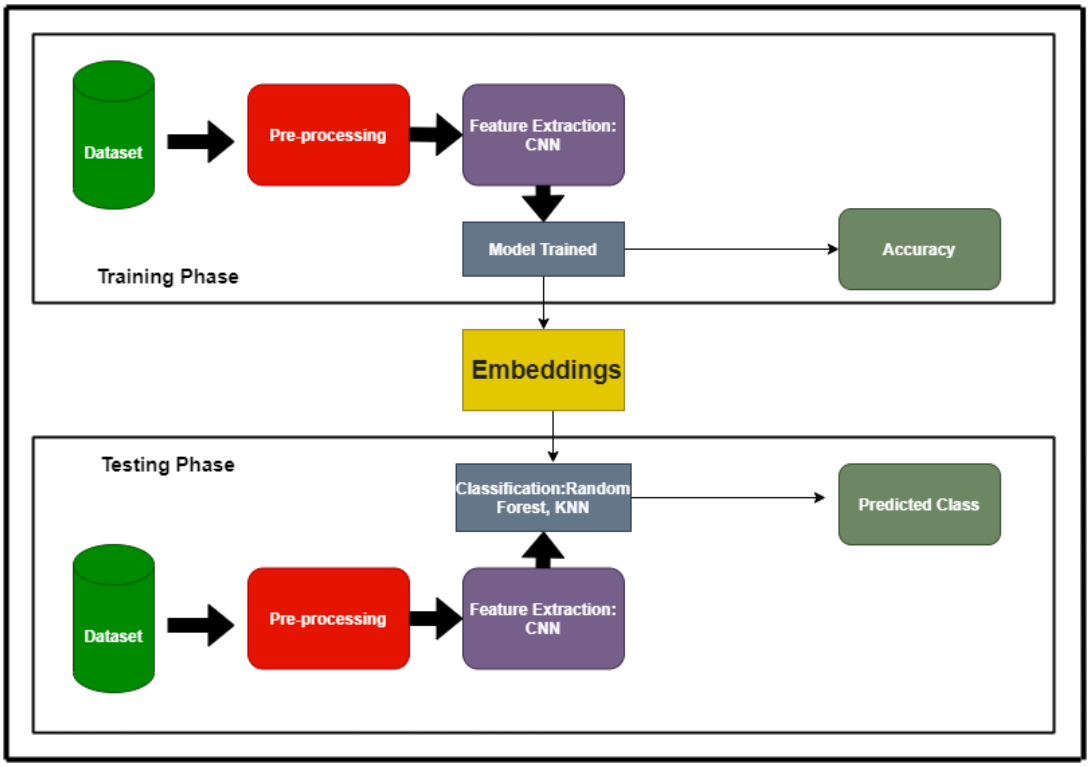

Figure 3: Architecture Design

Initially data was imported and further pre-processing and transformation is done so it can used used for model training. Pre-processing and normalization is very necessary before applying any model on data. Further the results of this is transformed data. Transformed data was split in two for training and testing purpose. Same process has been done on both train and test data. This data then further passed to the CNN model and feature from the images were extracted. Next embbedings process was done to obtained feature vector using array. The feature map from embbedings was compared with the features from the CNN model and further it store in the labels i.e four categories available in the dataset. Further this feature map is given as an input to the classi cation model such as Random Forest and KNN. This two classi ers will classify the images according to the classes present and will show output with the evaluation measures.

5 Implementation

In this section, the models constructed in the project to develop CNN and machine learning models that would detect and categorize rice leaf illnesses based on photographs of healthy and diseased leaves into three categories: brown spot, hispa, and leaf blast are explained.

In this study, Python version 3.8 is used for the implementation of the project work along with the spyder as IDE (Integrated Development Environment). CNN model takes long time to train the model. So, system con guration and GPU should be good and there should be proper installation of all the libraries before training the model and to execute the model smoothly. Various librabries were installed such as TensorFlow, Keras, Scikit-learn, Matplotlib, Pillow and Opencv. TensorFlow is one of the most popular library

which is used in image processing to build an e cient model. Keras is one of the most powerful library for developing deep learning models. It Also acts as an interface for TensorFlow library. Scikit-learn is used for machine learning algorithms for classi cation and regression tasks. Matplotlib is use for plotting in python programming language. Pillow and Opencv is used for working on image data. All these libraries were installed to build an environment for the actual implementation part.

From the the rst model i.e. CNN extraction of features is done from the images. All the necessary libraries for this model building were imported from TensorFlow library using keras. The libraries were installed using pip install command. Second step is embbedings which will create an array and try to extract features from the images and further those features will be compared with the feature obtained from the CNN model and it will store the feature vector according to the particular categories. Second model is a classi cation of images based on the categories. There are two classi cation models which are implemented such as Random Forest and KNN. These classi cation models will take the output i.e. feature vector as an input and then it will classify the images.

5.1 Convolutional Neural Network

For CNN model libraries were installed such as TensorFlow and keras for building an e cient model. After importing libraries pre-processing and transformation was done. In pre-processing resizing of images is done for taking good size as an input. The further batch size is set to 32 images per batch. Image augmentation is carried out using ImageDataGenerator function for normalization of data. This transformed data is split into training and test groups. The split ratio is given as 90:10.

Dense layer, convolutional layer, max pooling layer, attening layer and dropout layer are the layers which are de ned. CNN model is used for extracting features from the rice leaf images. In total 16 and 32 lters are added in the convolutional layers for ltering the images. Recti ed Linear Unit (ReLU) is used in the layers to enhance the execution speed of the model, which is used as an activation function in this model. Dense layer is used to connect all the layers with each other. Max pooling layer provides the feature map from the images. Flatten layer will convert the data in 1-dimensional array. The dropout layer contains softmax function which is used for classi cation of images but in this research work I am not classifying the images in the output layer of the CNN model. All the layers were de ned using keras library. Next the model was trained on 150 epochs and it takes few hours to train the model. After model is trained, visual representation of model accuracy, validation accuracy and training loss have been shown using matplotlib library.

5.2 Embbedings

The 'numpy' library is used to create an array in this process. 'os.path import listdir' library is used to get the list of all the directories and the les in the speci ed directory. The 'PIL' library is used for working on the image les. And further TensorFlow and keras libraries were used for the embbedings process.

5.3 Random Forest

Random Forest is an ensemble method which is used for classi cation. In this research work this classi cation algorithm is used for classifying the rice leaf images. To run this model 'RandomForestClassi er' from `sklearn.ensemble' library is imported. As 'sklearn' library is used for classi cation models. Pickle library is imported for saving the model. 'Sklearn.metric' library is used for checking some evaluation measures such as accuracy score, confusion matrix and classi cation report. 'Matplotlib' library used for plotting the graphs has been imported and 'seaborn' library for visualization purpose. The 'n_estimator' is set to 5 for good performance. It shows the number of trees available in the model. This model then further developed with train and test data with the split ratio of 90:10. And after successful running of model it will show various evaluation measures such as confusion matrix, accuracy, precision, recall and F-1 score for checking the performance of the model.

5.4 K-Nearest Neighbors

K-Nearest Neighbors is an simple supervised machine learning algorithm which is used for classi cation and regression. In this research work KNN classi cation algorithm is used for classifying the rice leaf images. To run this model 'KNeighborsClassi er' from `sklearn.neighbors' library is imported. As 'sklearn' library is used for classi cation models. Pickle library is imported for saving the model. 'Sklearn.metric' library is used for checking some evaluation measures such as accuracy score, confusion matrix and classi cation report. 'Matplotlib' library used for plotting the graphs has been imported and 'seaborn' library for visualization purpose. The 'n_estimator' parameter is set to 5 for better performance. This model then further developed with train and test data with the split ratio of 90:10. And after successful running of model it will show various evaluation measures such as confusion matrix, accuracy, precision, recall and F-1 score for checking the performance of the model.

6 Evaluation

All of the ndings gathered in meeting the project objectives are explained in detail in this chapter. The models and architectures developed in this study attempt to alleviate the di culties experienced by farmers in diagnosing and classifying rice leaf diseases. As a result, CNN is used to demonstrate the performance and robustness of the models developed in comparison to other machine learning algorithms such as Random Forest and K-Nearest Neighbors. The accuracy, recall, precision, and F-1 score were used as evaluation measures in this study to analyze the outcomes. Then, using the negative and positive classi cations of each class, a confusion matrix is generated to determine the right amount of predictions made by model for identifying rice leaf diseases. The performance of the models developed is evaluated using this matrix and the classi cation report. The table bellow summarizes the accuracy ndings for all of the models used in this investigation.

Model	Accuracy
CNN	80%
Random Forest	96%
K-Nearest Neighbors	72%

Table 2: Results Summary

The above table represents the result summary for all the models which are implemented in this research work. In this research project first model CNN performed quite well with the accuracy of 80% for feature extraction from images. Whereas, second classification model contains two classifiers such as Random Forest and K-Nearest Neighbors. Accuracy of random forest is 96 percent, which is very good. And the accuracy of the KNN classifier is 72 percent. Among both the classifications model Random Forest performed well with the CNN model as an feature extractor. Whereas, KNN have not given the best performance with the CNN model. In further subsections results for all implemented models is explained and visualised in detail.

6.1 Evaluation results for CNN

Results obtained from the CNN model is explained and visualized in this section.

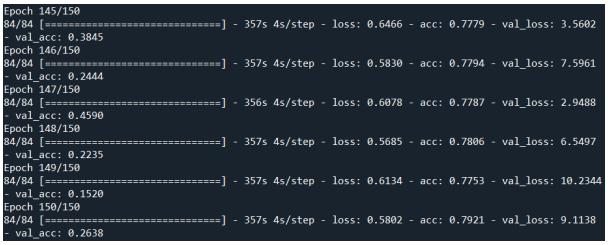

Figure 4: Accuracy of CNN

As CNN model was trained for 150 epochs can be seen in the figure 4. Accuracy of the model can seen in this graphical representation. Along with this loss and validation loss is generated for each and every epoch. CNN model performed quite well in this implementation for feature extraction from the leaf images.

The figure 5 and figure6 shows the graphical representation of accuracy and validation accuracy for the CNN And the graph for loss variation and validation loss of the CNN model.

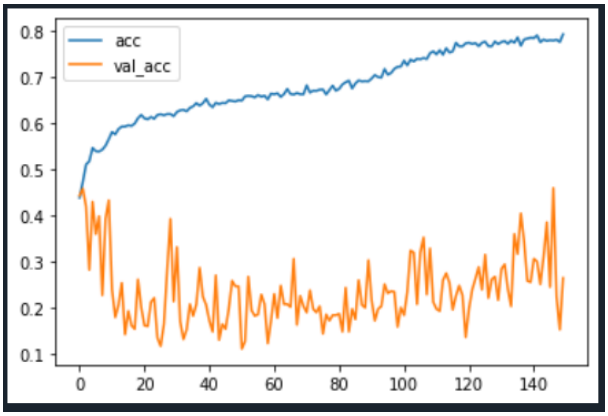

Figure 5: Accuracy Graph

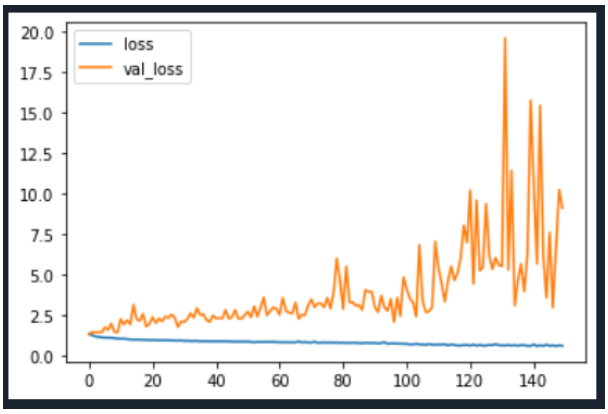

Figure 6: Loss Variation

6.2 Evaluation results for Random Forest

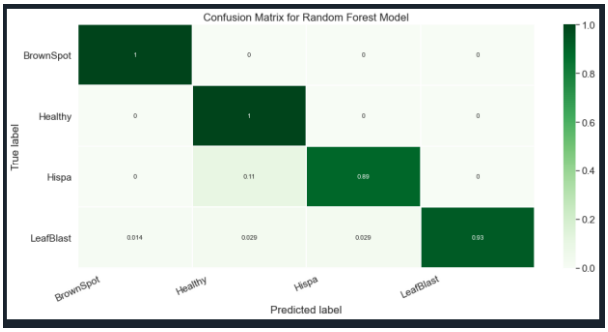

Figure 7: Confusion Matrix for Random Forest

Evaluation measures evaluated for the Random Forest Classi er is explained in this section. In gure 7 visual representation of confusion matrix is shown for Random Forest model. Confusion matrix is used to measure the performance of classi cation model. Which gives the summary of prediction results.

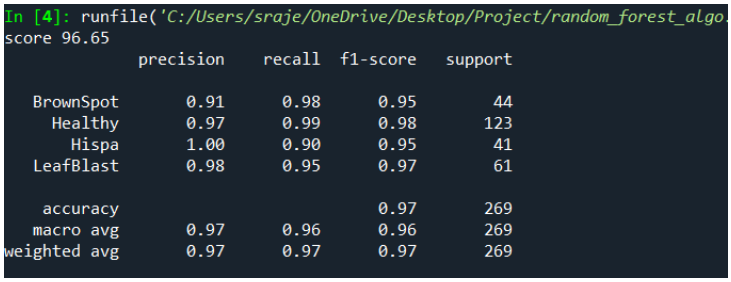

Figure 8: Evaluation Measures for Random Forest

The gure 8 shows the accuracy random forest model. And other evaluation measures such as precision, recall and F1-score for four categories of the images. From the graph precision value can be checked for hispa category, which is 1.00. Whereas, recall and F1-score values are good for healthy class among all others. And the values are 0.99 and 0.98 respectively.

In this research work, Random Forest classi er has been implemented for classi cation of rice leaf images based on the four classes such as healthy, brown spot, hispa and leaf blast. After training of CNN model and embeddings process feature vector was generated. The output i.e., feature vector has been taken as input by the classi cation algorithms. The 'RandomForestClassi er' from the 'Sklearn.ensemble' module was used to run this model. For classi cation models, the 'sklearn' package was employed. Some assessment measures, such as accuracy score, confusion matrix, and classi cation report, were checked using the 'Sklearn.metric' package. The 'Matplotlib' library was used to plot the graphs, and 'seaborn' library was utilized for visualization. For optimal performance, the 'n estimator' parameter was set to 5. It displays the total number of trees in the model. With a split ratio of 90:10, this model was further trained with train and test data. After the model was run successfully, various evaluation metric such as the confusion matrix, accuracy, precision, recall, and F1- score was checked to verify the model's performance.

Precision, recall and F1-score was checked for all four classes. Were precision metric for hispa class was achieved value of 1.0 with the support 41. Precision is calculated by dividing the total number of positive results by the total number of positive results predicted by the classi er with false positive events i.e., TP/TP+FP. Model which produces no false positives gives the precision value has 1.0. In this research work, hispa class shows precision value of 1.0 for random forest model. Were 41 images taken for testing and the value 1.0 says there are no false positives outcomes. The formula for precision is true positive divided by addition of true positive and false positive. In this case support is 41 i.e., 41 images tested for this class and it has been predicted all the images perfectly and there are no false positive events recorded. So, if there are zero false positive values then as per formula the value will be zero and then the true positive values will be divided with true positive events which is 41 and after dividing this the value achieved is 1.0.

6.3 Evaluation results for K-Nearest Neighbors

Evaluation measures evaluated for the KNN Classi er is explained in this section. In gure 9 visual representation of confusion matrix is shown for KNN model. Confusion matrix is used to measure the performance of classi cation model. Which gives the summary of prediction results.

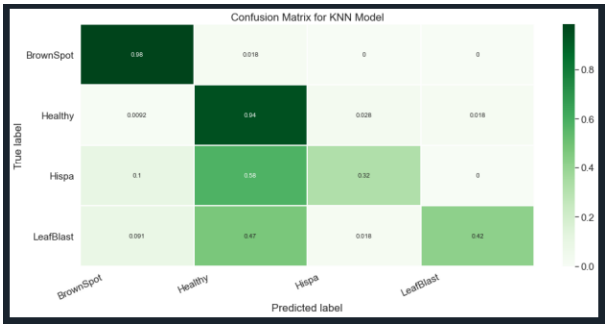

Figure 9: Confusion Matrix for KNN

The gure 10 shows the accuracy for KNN model. And other evaluation measures such as precision, recall and F1-score for four categories of the images. From the graph precision value for leaf blast category has been checked, which is 0.92. Whereas, recall and F1-score values are good for brown spot class among all others. And the values are 0.98 and 0.90 respectively.

```
In [9]: runfile('C:/Users/sraje/OneDrive/Desktop/Project/knn_algo.
score 72.86
              precision    recall  f1-score   support

   BrownSpot       0.83      0.98      0.90        55
     Healthy       0.65      0.94      0.77       109
       Hispa       0.80      0.32      0.46        50
   LeafBlast       0.92      0.42      0.57        55

    accuracy                           0.73       269
   macro avg       0.80      0.67      0.68       269
weighted avg       0.77      0.73      0.70       269
```

Figure 10: Evaluation Measures for KNN

6.4 Discussion

From the results which are discussed above, CNN model have given the accuracy of 80%. Whereas the classi cation models such as random forest and KNN gave an accuracy of 96% and 72% respectively. In this research work Random Forest model have given better performance with the CNN model.

The most relevant study which is related to this research work was conducted by Chouhan et al. (2021). In this study leaf disease detection and classi cation is done using neural network and machine learning classi ers. Their research accomplished the accuracy of 90% for the Random Forest classi er with the neural network. As compared to their research work our research performed well with the CNN model with random forest as a classi er.

Another most relevant study conducted by Sujatha et al. (2021) have implemented deep learning with machine learning techniques for plant disease detection. Their research accomplished the 76% of accuracy for random forest model with the CNN model. In comparison with our research their model showed less accuracy.

7 Conclusion and Future Work

To ensure better quality, quantity and productivity of the yield, it is very crucial for identifying the diseases at early stage for reducing the use of pesticides to reduce damage of the crops and environment. The main aim of this research was to detect and classify the diseases in the rice leaf, having four categories of classes as healthy, hispa, brown spot and leaf blast. In this research study convolutionl neural network was used for the feature extraction from the rice images. Whereas, Some machine learning classi ers such as Random Forest and K-Nearest Neighbors were used for the classi cation of the diseases based on the categories. The implementation was done on total 3,355 images of rice leaf. The rst model CNN performed well for the feature extraction with the accuracy of 80 percent. Along with this second model was classi cation of diseases using some machine learning classi ers such as Random Forest and K-Nearest Neighbors, accomplished the accuracy of 96% and 72% respectively. In this research project Random forest achieved better accuracy than KNN with the CNN model.

In future, accuracy can be improved for CNN model to get more good results and also for KNN model. This can be achieved with more good quality of image database. Various image enhancement techniques and also increase or decrease in number of epochs can be done for checking the improvements in the model performance.

8 Acknowledgment

My heartfelt gratitude to our college, National College of Ireland, and the MSc in Data Analytics department for allowing me to successfully complete this research project over the previous three months. Gratitude to Mr. Majid Lati , my supervisor and mentor, for his help throughout the project. His unwavering support aided me in nishing and presenting this thesis. I appreciate his time and, in particular, his insightful supervision.

References

Atila, Ü., Ucar, M., Akyol, K. and Ucar, E. (2021). Plant leaf disease classi cation using e cientnet deep learning model, *Ecological Informatics* **61**: 101182.

Azadbakht, M., Ashourloo, D., Aghighi, H., Radiom, S. and Alimohammadi, A. (2019). Wheat leaf rust detection at canopy scale under di erent lai levels using machine learning techniques, *Computers and Electronics in Agriculture* **156**: 119{128.

Bao, W., Zhao, J., Hu, G., Zhang, D., Huang, L. and Liang, D. (2021). Identi cation of wheat leaf diseases and their severity based on elliptical-maximum margin criterion metric learning, *Sustainable Computing: Informatics and Systems* **30**: 100526.

Chouhan, S. S., Singh, U. P., Sharma, U. and Jain, S. (2021). Leaf disease segmentation and classi cation of jatropha curcas l. and pongamia pinnata l. biofuel plants using computer vision based approaches, *Measurement* **171**: 108796.

Hu, G., Wang, H., Zhang, Y. and Wan, M. (2021). Detection and severity analysis of tea leaf blight based on deep learning, *Computers & Electrical Engineering* **90**: 107023.

Jiang, F., Lu, Y., Chen, Y., Cai, D. and Li, G. (2020). Image recognition of four rice leaf diseases based on deep learning and support vector machine, *Computers and Electronics in Agriculture* **179**: 105824.

Kaur, N. and Devendran, V. (2020). Novel plant leaf disease detection based on optimize segmentation and law mask feature extraction with svm classi er, *Materials Today: Proceedings* .

Mojjada, R. K., Kumar, K. K., Yadav, A. and Prasad, B. S. V. (2020). Detection of plant leaf disease using digital image processing, *Materials Today: Proceedings* .

Sachdeva, G., Singh, P. and Kaur, P. (2021). Plant leaf disease classi cation using deep convolutional neural network with bayesian learning, *Materials Today: Proceedings* .

Sethy, P. K., Barpanda, N. K., Rath, A. K. and Behera, S. K. (2020). Deep feature based rice leaf disease identi cation using support vector machine, *Computers and Electronics in Agriculture* **175**: 105527.

Sujatha, R., Chatterjee, J. M., Jhanjhi, N. and Brohi, S. N. (2021). Performance of deep learning vs machine learning in plant leaf disease detection, *Microprocessors and Microsystems* **80**: 103615.

Tian, L., Xue, B., Wang, Z., Li, D., Yao, X., Cao, Q., Zhu, Y., Cao, W. and Cheng, T. (2021). Spectroscopic detection of rice leaf blast infection from asymptomatic to mild stages with integrated machine learning and feature selection, *Remote Sensing of Environment* **257**: 112350.

Yadav, S., Sengar, N., Singh, A., Singh, A. and Dutta, M. K. (2021). Identi cation of disease using deep learning and evaluation of bacteriosis in peach leaf, *Ecological Informatics* **61**: 101247.

Zhang, K., Wu, Q. and Chen, Y. (2021). Detecting soybean leaf disease from synthetic image using multi-feature fusion faster r-cnn, *Computers and Electronics in Agriculture* **183**: 106064.