Automated Wheat Stem Rust Detection using Computer Vision

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Rahul Moorthy Mahesh

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Prof.Volkan Isler

May, 2023

Acknowledgements

I would like to thank my advisor Professor Volkan Isler for his continuous support, supervision, and care during the whole process of problem-solving. Without his insights into the problem, I would not have been able to grow as a researcher and gain the satisfaction of solving an impactful problem during my Master's degree. I'm sure to keep growing under his guidance going forward as a Ph.D. student.

I would also like to thank the USDA collaborators Dr. Matthew Rouse, Dr. S. Kianian, and especially Professor Ce Yang for their constant support and insights into the problem from the agriculture perspective. Without their support, I would have had no data and an understanding of the technical jargon to even begin solving the problem.

A special thanks to Professor Nikolaos Papanikolopoulos for his constant dedication towards growing the Masters in Robotics program and for his immense care for all the students. Without his support, this whole journey and growth would not have been possible.

Finally, a big thanks to my fellow lab mates at the Robotic Sensor Networks lab and all my friends for providing me the motivation to strive hard and making this journey enjoyable and memorable.

Dedication

I dedicate this thesis to my mother, father, and sister

Abstract

Wheat is one of the most important cereal crops, contributing significantly to the financial economy and food sources. Currently, the direct consumption of wheat amounts to about 41%. Additionally, in 2019 alone, the global trade value of wheat was about $39.6 billion. Hence, the protection of the yield of such crops from diseases is of immense importance.

Stem rust is a fungal disease that attacks cereal crops. In particular, it is a common disease that occurs in wheat and destroys 50 to 70% of the yield if left unchecked. The loss of yield would in turn affect the economy and food consumption. Thus, there is a need to detect the outbreak early to apply fungicide treatment to the field. The traditional approach for detection involves experts inspecting the fields visually and grading them for stem rust which is a time-consuming process for a large field and can also be affected by human errors. Hence, an automated approach to the grading process would help solve such problems. The availability of an automated grading process will allow mobile robots, popularly being used for activities like irrigation, seed sowing, and precision agriculture to rapidly perform grading and alert the experts in case of detected stem rust. The alert through the automated detection would in turn lead to a timely application of fungicide for preventing the spread of stem rust in an efficient manner.

The thesis focuses on formulating the wheat rust grading as a multi-class classification problem and demonstrating the effectiveness of the visual attention approach for solving it. The thesis also presents the first RGB field dataset with labels from experts for the development of automated stem rust grading approaches. The proposed approach was developed and evaluated on the presented dataset and shows the ability to distinguish between different intensities of stem rust with **86%** accuracy. The reliability of the network is also validated qualitatively through attention maps where the visual

attention approach shows interpretable focus areas compared to traditional detection approaches which fail to identify the general presence area of stem rust.

Contents

List of Tables

List of Figures

Chapter 1

Introduction and Background

1.1 Wheat Stem Rust Grading

The stem rust detection is currently performed through manual grading or phenotyping by experts. In the grading process, the experts visually inspect large fields at regular intervals to detect the outbreak of stem rust. Stem rust causes the occurrence of pustules on the stem. Figure 1.1 shows an example of a healthy and stem rust-infected wheat. The infected wheat can be observed to have dark brown pustules spread across the stem area. The grading process involves determining the Coefficient of Infection (CI) by inspecting the pustules for infection response and severity.

Figure 1.1: Stem rust disease brown pustules observed on the stems

The infection response is determined by visually investigating the color of the pustules. Figure 1.2 shows the categories of infection response where different categories have different colors of the infected area. The four main categories are R, MR, MS, and S, but as humans perform the grading, each plot can also have combinations of categories where the first category represents the most dominant infection response shown by the plant. In the main categories, "M" represents medium, "S" represents susceptible, and "R" represents resistant. Next, the severity estimation examines the percentage area of the pustules. In Figure 1.3, scale "B" represents values given to different area coverage of pustules in which higher area results in higher severity. The range of values is from 0 to 100. The relation between severity and infection response is also of a direct correlation which means higher infection response results in higher severity.

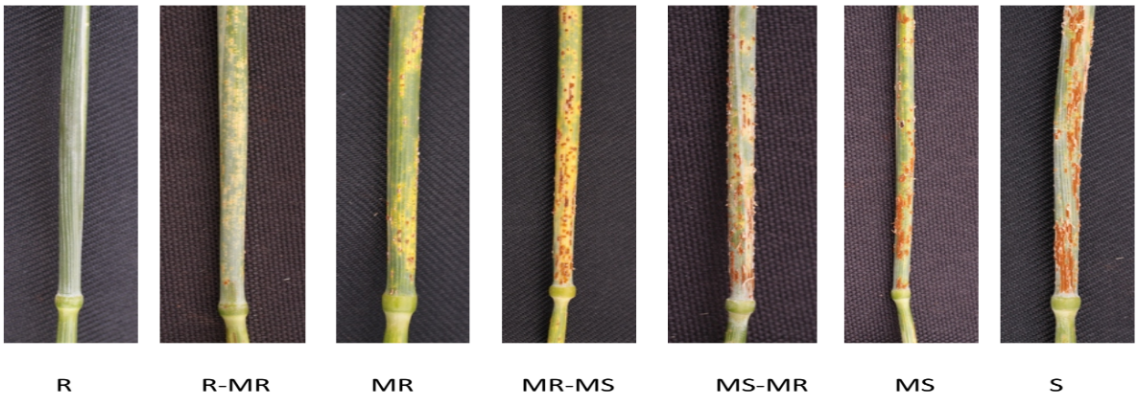

Figure 1.2: Colour-based infection response grading [1]

After evaluating the disease response and severity, the resultant value of the grading is the CI. The CI is the product of disease response and severity. The disease response is converted to a numerical value using a weighted average. The experts assign each main category a number - 0.2 to R, 0.4 to MR, 0.8 to MS, and 1 to S. The resulting computation is the weighted average of the combination of infection response labels assigned to the plot. The first category is given double the weight compared to other

categories. For example, if a plot is assigned R-MR, the resultant infection response is (0.2*2 + 0.4*1)/(2+1) equals 0.26. The range of the disease response is from 0 to 1, while that of CI is from 0 to 100. The final estimation of the CI value is the whole process of grading performed for each plot manually.

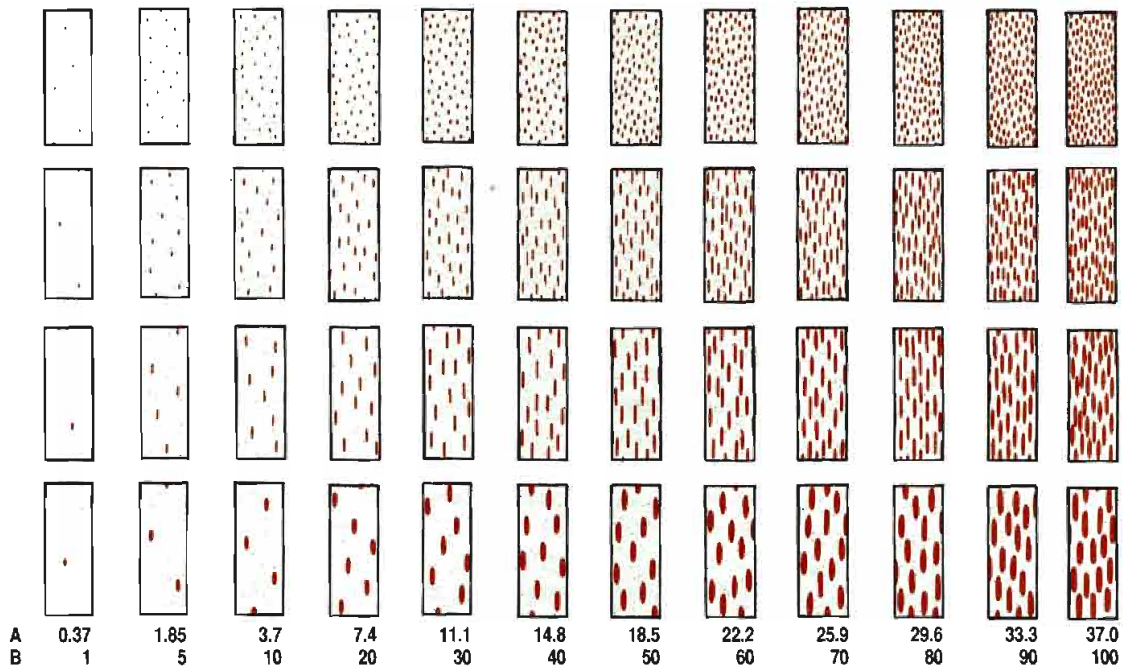

Figure 1.3: Density-based modified Cobb scale severity grading [2]: higher density results in higher severity

1.2 Significance of Automated Grading

The manual grading process is time-consuming and labor-intensive. Furthermore, individual expert bias can also affect grading. Hence, there is a need to automate the manual grading mechanism so that mobile robotic systems can be used to rapidly perform such repetitive processes in an efficient manner. The automation would, in turn, help early detection of stem rust outbreaks without the dependency on human experts.

Furthermore, it can act as a reference grading for experts to reduce the time consumed on a single wheat plot.

1.3 Current Works on Automated Disease Detection

Automated grading or disease detection is an active research area in the agriculture community. The research is motivated to detect diseases early and prevent them from spreading to other plants. The outbreak of smartphones and cameras has shifted the paradigm of automated disease detection using such devices as they can be easily installed in a field and is also accessible to the population. Sannakki et al [3] proposed an image processing methodology to identify diseases and perform automated grading for pomegranates from images. Mohanty et al [4] trained a deep learning-based architecture to detect 26 diseases on 14 crop species. The approach was able to achieve 99.35% accuracy on the test set. Baranwal et al [5] also developed a deep-learning framework for apple leaf disease detection. Wallelign et al [6] used a Lenet architecture to identify soybean plant diseases. Sabrol et al [7] developed a machine-learning approach by extracting features from images for tomato plant disease classification. All the works used images captured by a camera on a smartphone.

Most previously described works are on plant disease identification, with no focus on wheat. The existing literature present for wheat only focuses on detecting or grading leaf rust disease. Johannes et al [8] developed an image-processing algorithm combined with statistical analysis to detect the presence of septoria, rust, and tan spot on leaves. Lu et al [9] developed a wheat disease diagnosis system using weakly supervised deep learning to localize different types of leaf rust. Schirrmann et al [10] used a Resnet architecture to analyze localizing stripe rust during different phases of rust progression. Pan et al [11] developed an ensemble approach to differentiate between healthy wheat, stem, and leaf rust. Tang et al [12] developed a semi-automated image labeling approach to reduce the effort in dataset creation for early stripe rust detection. Maqsood et al [13] used

SRGAN to upsample small images for better feature extraction. The training pipeline of the CNN network uses these upsampled images, achieving 83% accuracy for stripe rust detection. All these works were developed on a dataset collected in a laboratory setting, making it unsuitable for field usage as the environment changes drastically. The only work which uses field data is by Mi et al [14], which developed a visual attention mechanism to grade stripe rust disease. The grading problem was formulated as a classification problem in this work and attained 97% accuracy in the grading process.

These works show sufficient exploration of the stripe rust detection problem, but the research community is yet to study the problem of automating stem rust grading. The potential reason for the lack of study is the non-availability of a labeled dataset. Hence, the thesis focuses on both these aspects of creating a field dataset that the research community can use to study automated stem rust grading and developing an automated approach for the same. The automated approach could act as a stepping stone to begin research for early stem rust detection resulting in reduced loss of yield from wheat due to the disease.

1.4 Thesis Contribution

The thesis focuses on developing an automated detection of wheat stem rust for field usage, which comes down to a grading problem. The contributions through the thesis are as follows:

- Creating the first RGB field dataset for stem rust detection with labels from the experts
- Qualitative and quantitative study of a data-driven approach for determining the Coefficient of Infection (CI)
- Formulating the problem of detection as a multi-class classification problem to provide reliable and interpretable detection focus maps

- Demonstrating the effectiveness of the visual attention-based approach for stem rust detection over a large Convolution Neural Network (CNN) and evaluating it qualitatively and quantitatively
- Analyzing the failure modes of the visual attention approach using embedding visualization techniques and statistical analysis. Performing domain-specific changes for resolving the failures
- Developing a normalization procedure to combine two datasets captured at different scales without loss of detection performance

Chapter 2

Object Detection

2.1 Introduction

The task of object detection focuses on identifying and localizing objects in the scene. These detection approaches have various applications in the domain of robotics like performing highly accurate manipulation, scene understanding for safe navigation, and semantic SLAM. The computer vision research community has heavily progressed in achieving high performance on popular object detection datasets like CoCo [15] and Pascal VOC [16].

2.2 Related Works

Object detection was first performed using handcrafted features. Viola Jones detector [17] used a sliding window to detect the presence of human faces inside the window. The features were represented using the haar wavelet and template feature matching was performed between the window and the template image to identify the faces. P. Felzenszwalb et al [18] proposed a deformable part model which divided the object into different parts and built a HOG features-based part detector to detect various parts of the object. These approaches though worked well, did not generalize across scale, size,

and environmental changes.

The emergence of popular datasets like CoCo [15], and Pascal VOC [16] for object detection allowed the use of data-intensive deep learning architectures. Ross Girshick et al [19] proposed the RCNN architecture which is a two-stage object detector using selective search for object proposals, pre-trained Convolution Neural Network (CNN) for extracting features, and finally a Support Vector Machine model to detect the type of object. The time-consuming nature of the model led to the Faster RCNN [20] model which eliminated selective search and used a separate network to learn object proposals. The model though effective still performed redundant computations for both object proposal and object identification. Hence, YOLO [21] was developed which proposed an end-to-end architecture to perform detection and classification simultaneously.

Currently, with the introduction of attention blocks and transformer-based architectures which provides interpretable features and high performance as compared to traditional CNN architectures. The state-of-the-art results on the CoCo [15] and Pascal VOC [16] datasets have been achieved by attention-based networks. Yanghao Li et al [22] proposed using ViT (Vision Transformer) as a backbone for object detection which achieved state-of-the-art results on both the CoCo [15] and Pascal VOC [16] dataset. Tianhe Ren et al [23] proposed combining the FocalNet-Huge block as a feature backbone with the use of a Stable-DINO detector showing high mean Average Precision on the CoCo [15] test dataset.

2.3 Stem Rust Detection

Stem rust detection is a problem to identify and localize the presence of stem rust disease in wheat. The annotation process for stem rust localization is a very time-consuming and labor-intensive process with the disease spread across the stem in a non-uniform manner. Furthermore, the expert methodology to perform the detection essentially comes down to grading the wheat visually for stem rust intensity. Hence, in the thesis,

the stem rust grading is initially regressed in a data-driven manner. The analysis of the regression approach led to the formulation of the detection as an image classification problem in which various intensities of disease are identified based on the range of the CI values.

2.4 Evaluation Metrics

The standard metrics of classification were chosen for the evaluation of the approach for stem rust detection. The metrics are accuracy, precision, recall, and confusion matrix. The mathematical form of each of the metrics is shown in Figure 2.1. Here, TP represents the true positive samples which means the number of positive samples predicted as positive by the model, TN represents the true negative samples which means the number of negative samples predicted as negative by the model, FP represents the false positive predictions which means the number of negative samples predicted as positive by the model and FN represents the false negative predictions which mean the number of positive samples predicted as negative by the model.

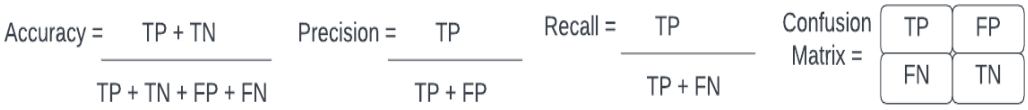

Figure 2.1: Accuracy, precision, recall and confusion matrix mathematical form: TP (True Positive), FP (False Positive), FN (False Negative), and TN (True Negative)

The accuracy measures the prediction performance of the model. It is a metric that estimates the performance of the model in a balanced test set setting while is a bad metric for imbalanced data. Hence, precision, recall, and confusion matrix are estimated which determines per class performance for the model.

Chapter 3

Wheat Stem Rust Dataset Preparation

3.1 Data Collection

The state-of-the-art methods rely on the availability of labeled data for highly accurate performance in detection. One major limitation of wheat stem rust detection is the lack of publicly available datasets as the majority of the literature is focused on studying leaf rust detection. Thus, the thesis provides the first field dataset for studying stem rust detection.

The data collection aimed to collect the data which captures the effect of stem rust on wheat. Additionally, grade the effect of stem rust by plant pathology experts. There were two rounds of the collection performed during consequent summers. For convenience, the naming convention of the data collection cycles was according to the city of the collection - Rosemount and StPaul data collection which are both located in the state of Minnesota, USA.

3.1.1 Rosemount Data Collection

The Rosemount cycle was the first iteration of data collection. The collection cycle focused on capturing close-range videos of stem rust-affected plots using a hand-held camera. The range was chosen manually according to the clear visibility of the wheat stem and the effect of rust in the camera's field of view. Furthermore, plant pathology experts graded the plots. The time of the data collection was matched with the grading to affirm the labels with the data collected.

The cycle captured approximately 1-minute videos of 29 plots. The resolution of the video was 3840×2160 at 30 Frames Per Second (FPS). Furthermore, the camera was moved slowly, capturing all the wheat stems in the plot. A few sample images extracted from the video are shown in Figure 3.1

Figure 3.1: Rosemount sample extracted frames from video

Data Analysis

The data was then analyzed to understand the distribution of the CI values. The performed analysis would be crucial because a diverse dataset is needed to develop a robust approach, as a dataset skewed towards a particular population would prevent the approach from generalizing. The distribution was examined by first computing the CI using the infection response and severity provided by the experts. It is then qualitatively investigated for skewness using a histogram shown in Figure 3.2. The observed range of

CI distribution is between 1 to 60 rather than the full infection effects range of 1 to 100. Furthermore, skewness was present in the distribution towards the lower end of 1 to 10 which was quantified to affirm the qualitative observation using the skewness statistical measure, which calculates the asymmetry in the distribution. The estimated measure was 2.05, which defines a skewed unimodal distribution rather than a diverse normal distribution (0 skewness measure). The high skewness measure could be a problem for directly detecting the infection using CI values.

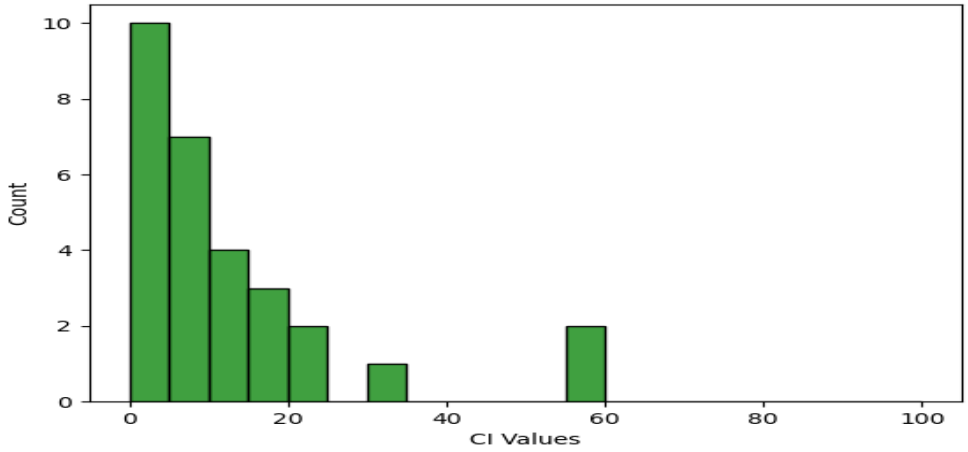

Figure 3.2: Skewed Rosemount CI value distribution: skewed towards the lower end

Additionally, the hypothesis concerning the correlation between severity and infection response was verified to validate the integrity of the grading. Figure 3.3 shows the relationship between severity and infection response across all the plots of Rosemount data. There is the presence of a direct correlation between the severity and infection response through visual observation. A quantitative measure of Pearson correlation, which statistically determines the relationship between two or more variables, was calculated to assert the observation. The estimated correlation was 0.86, which shows a high correlation between the severity and infection response.

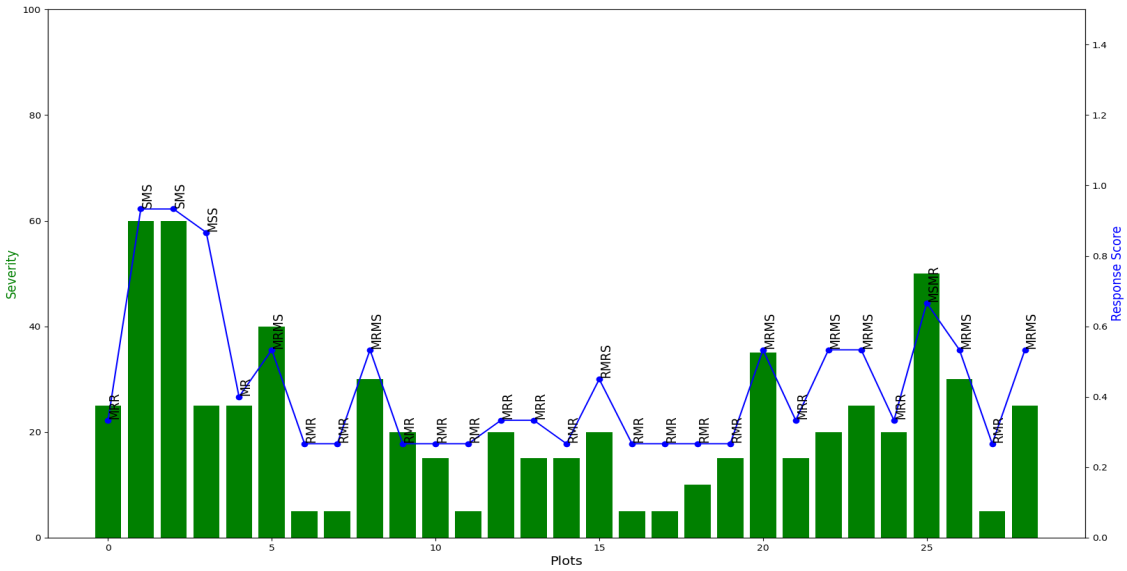

Figure 3.3: Direct Rosemount severity vs infection response correlation validation: observed high severity when high infection response

These two were the only essential aspects considered for analyzing the data collected.

3.1.2 StPaul Data Collection

The StPaul cycle was the second iteration of data collection. The collection cycle focused on a similar setting as Rosemount data capturing close-range videos of slowly covering the whole stem rust-affected plot using a hand-held camera. The range was again chosen manually according to the clear visibility of the wheat stem and the effect of rust in the camera's field of view. Furthermore, the same experts again graded the plots to ensure labeling integrity.

In the StPaul cycle, data collection captured approximately 1-minute videos of 4 rows having 20 varieties of wheat summing up to 80 plots. The resolution of the video was 3840×2160 at 30 FPS. Additionally, 7 - 10 images were also captured for each plot to augment the video data. The resolution of the images was 4032×3024. A few sample images extracted from the video are shown in Figure 3.4.

Figure 3.4: StPaul sample frames extracted from video

Data Analysis

The StPaul data is analyzed similarly to the Rosemount data. The CI value distribution was qualitatively analyzed using a histogram shown in Figure 3.5. The observed range of CI values was again between 1 - 60 for StPaul data. Furthermore, the data had skewness towards the lower values, but relatively less than Rosemount's. The skewness comparison can be quantified using the skewness statistical measure. The estimated measure is 0.792 for StPaul data, which is a skewed unimodal distribution (greater than 0), but it is less skewed than Rosemount's data, which had an estimate of 2.05. Even such tiny skewness could also be a concern for estimating the CI values from data.

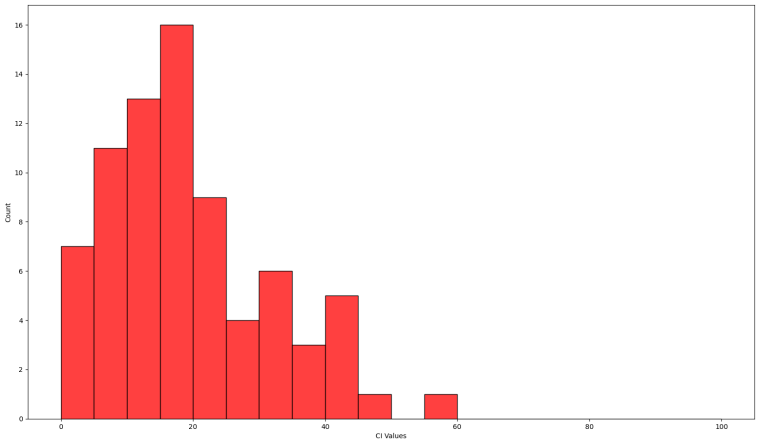

Figure 3.5: Skewed StPaul CI value distribution: skewed towards the lower end

Figure 3.6 verifies the hypothesis concerning the correlation between severity and infection response for all plots in StPaul data. There again exists a direct correlation between the severity and infection response which is quantified using the Pearson measure, estimated to be 0.85, showing similar behavior of high correlation between severity and infection response.

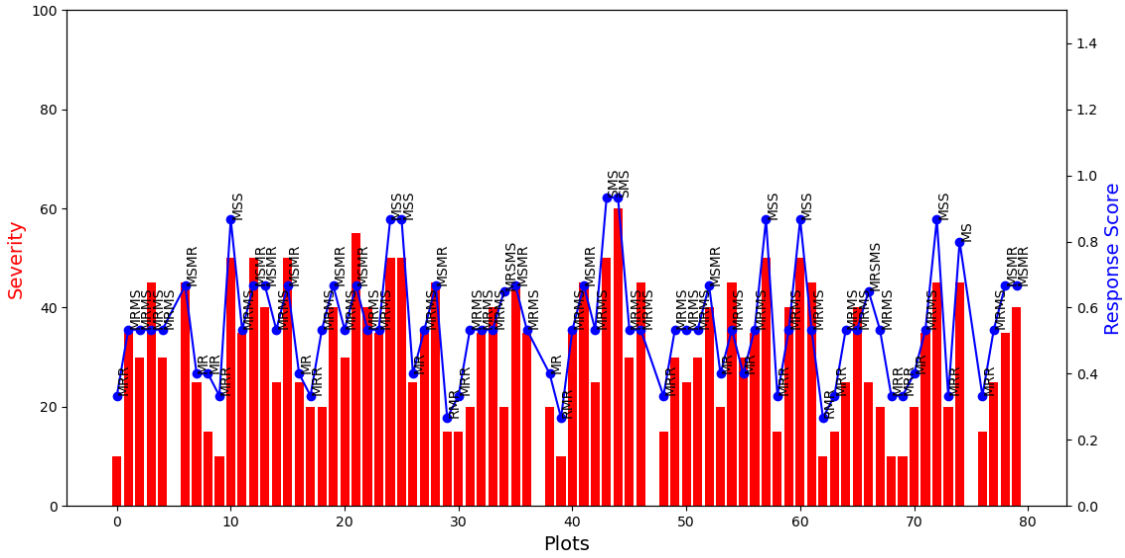

Figure 3.6: Direct StPaul severity vs infection response correlation validation: observed high severity when high infection response

3.2 Data Pre-processing Pipeline

Data collection is always followed by data pre-processing, an essential step for developing an approach to solving any problem. The pre-processing step ensures the quality of the data used for development. The pre-processing steps for the problem of stem rust detection consist of image extraction, FPS reduction, out-of-focus data removal, and field of interest extraction. Both datasets undergo these steps of data pre-processing.

Firstly, the frames were extracted from videos for computational efficiency compared to using videos. During the extraction of frames, the frame rate was reduced to 6 to remove redundant images covering the same field of view. The extracted frames for each video of the plots were then visually observed to discard frames that are highly blurred or with stems not occupying more than half of the field of view. Figure 3.7 shows examples of such frames. The filtered frames then go through the Field of Interest (FOI) extraction which crops the area focused by the camera, as it would be the only region vital for the approach. The decision about the area is determined using visual observation which is kept constant in all images. Figure 3.8 shows the sample cropped FOI area on one image for StPaul and Rosemount dataset.

Figure 3.7: Examples of bad quality data

Figure 3.8: Sample field of interest on Rosemount and StPaul data

The images after the cropping process are of resolution 3520×1440 form the final dataset. The CI is calculated for each plot using the grading by the experts which act as the labels for the images corresponding to that plot. The dataset statistics after the pre-processing are shown in Table 3.1.

Dataset	**Number of Images**
Rosemount	6305
StPaul	16706

Table 3.1: Dataset Statistics

Chapter 4

Wheat Stem Rust Coefficient of Infection Prediction

4.1 Introduction

The grading process for stem rust is to determine the CI directly from manual inspection by experts. If an approach can determine the CI values directly from images. It would help reduce the time consumed to grade a single plot. Thus, the work focuses on developing a data-driven approach to determine the CI from images.

4.2 Problem Formulation

The detection problem was formulated as a regression problem where given an input image of a wheat plot, the work builds an automated stem rust grading approach with the goal to regress the CI by extracting features from the image.

4.3 Approach

Convolutional Neural Networks (CNN) are modern methods used in literature for automatic data-driven feature extraction from images. They have shown promising results on tasks like scene recognition (ImageNet [24]), and object detection (CoCo [15]). Hence, considering their effectiveness, the approach also uses a CNN architecture to determine the CI from data.

4.3.1 Network Architecture

Resnet - 50 [25] is a popular architecture with high performance on the ImageNet dataset popular for scene recognition. Figure 4.1 shows the network architecture of Resnet - 50 [25]. Traditionally, architectures use sequential convolutional layers, which take input from the previous layer. The Resnet introduced a new connection called a residual connection which feeds the previous layer to the next layer but also to the layer, which is two layers apart from the current one. The skip connection is Resnet allows the architecture to learn similarly to the deep networks with less computation complexity. The choice of Resnet - 50 [25] architecture was for computational complexity considerations and understanding if data-driven estimation is possible for CI values in a rapid manner.

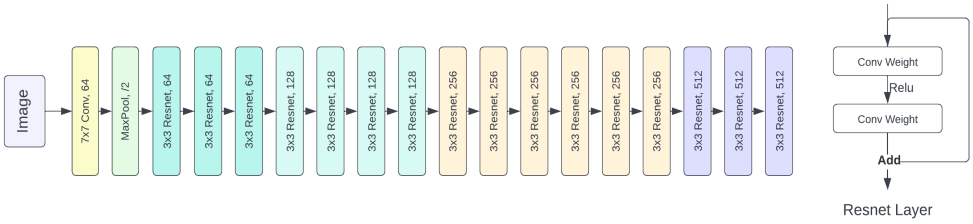

Figure 4.1: Resnet-50 architecture

The architecture is used as a feature extractor, and then an adaptive global average pooling is added to convert to a 1D vector which is then fed to six fully connected layers of nodes - 1024, 512, 256, 128, 64, 32. These fully connected layers would act as a

feature extractor for the 1D feature embedding. The final fully connected layer is then fed to one output node representing a single CI value. The activation function for the fully connected layers is relu, while for the output layer is linear as a number ranging from 0 - 100 is to be determined from the image.

4.3.2 Data Preparation

The network cannot be fed the full 3520×1440 image due to the computational complexity of processing such a large image. Hence, the image was resized by maintaining the aspect ratio. The length and width were divided by four resulting in a size of 880×360. All the images from the dataset were resized to the stated size which was the only preprocessing step that was performed for feeding to the network.

The network had to be trained on a training set. Hence, the 80 plots from StPaul were divided into two distributions. Images from half of it are used for training, while the other half is for testing. The division into two disjoint distributions would ensure that the train and test sets are different. The Rosemount data was not used in testing as the amount of data was really small for being divided into two distributions. Additionally, two sets of experiments were also designed in which, firstly, Rosemount data is also used in the training set, then in the second experiment, only the StPaul dataset alone is used in the training set. These experiments were performed to understand if Rosemount data addition is helping improve the performance of the task of CI value prediction.

4.3.3 Network Training

The training procedure first includes loading ImageNet weights to the Resnet-50 feature extractor which is known as transfer learning and helps faster network loss convergence. The network was then trained for 50 epochs using the Adam optimizer with a learning rate of 0.01, aiming to minimize the mean squared error loss function. Finally, the batch size was selected to be 32.

4.4 Evaluation Metrics

The approach was evaluated based on mean squared error and mean absolute error. The mean squared error is the mean of the squared error between the predicted CI value and the actual CI value. The mean absolute error is the mean of the absolute difference between the actual CI value and the predicted CI value. These metrics are considered a standard evaluation methodology for regression tasks.

4.5 Results

The quantitative evaluation metrics were estimated using the predicted CI values by the trained Resnet-50 model. Table 4.1 shows the metrics estimated from the predictions. The model does not estimate the CI values accurately, given the high mean absolute and squared error for the CI value having a range from 0 to 100. Figure 4.2 explains the observation of improper estimation in a better manner. The figure visualizes the mean predicted and actual CI values for images in half of the test plots, which shows a significant difference between the actual and predicted CI value. The degraded performance might be due to the skewness and lack of data in the training dataset to generalize well for a regression task.

Training Data Experiments	Test Mean Squared Error	Test Mean Absolute Error
StPaul with Rosemount	306.008	13.57
StPaul	**79.65**	**6.53**

Table 4.1: Coefficient of Infection Regression Metrics

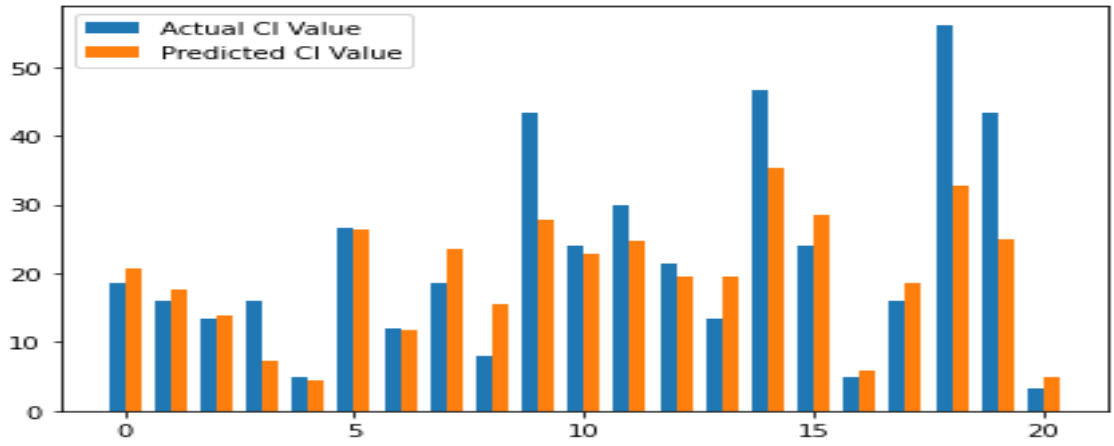

Figure 4.2: Degraded CI value regression qualitative analysis: high difference between the predicted and actual CI value

One another observation from Table 4.1 is that the metrics have less error when StPaul alone is used for training rather than using Rosemount and StPaul data. The observed reduction in performance is opposite to the expected behavior as increased diversity of data improves the prediction of the model. The reasoning for reduced performance is due to the difference in scale between Rosemount and StPaul's data. The Rosemount data collection was from a few feet away compared to StPaul data. The scale difference of a few feet would affect the features extracted when combining two datasets. Hence for further study of the problem, only StPaul data was used for training rather than combining Rosemount and StPaul.

The area of focus for layers of Resnet - 50 [25] was also analyzed for the interpretability of the predictions using attention maps. The Gradcam activation map [26] is an attention map that projects the activations of the layer under consideration onto the original image. These projected activations show the area of focus on the original image. Figure 4.3 shows the visualization of the Gradcam activation maps on one of the images. The ideal focus area for intermediate layers should be on the stem region as stem rust is only present in that area. In Figures 4.3 - c and d, the area of focus is the

ground and bottom of the stem, which puts the reliability in doubt due to the absence of stem rust in those areas. The reason for such focus areas might be due to the CI value labels not providing enough supervision for guiding the network towards the stem area. Hence, there is a need for a problem formulation that supervises the area of the presence of stem rust in an understandable manner.

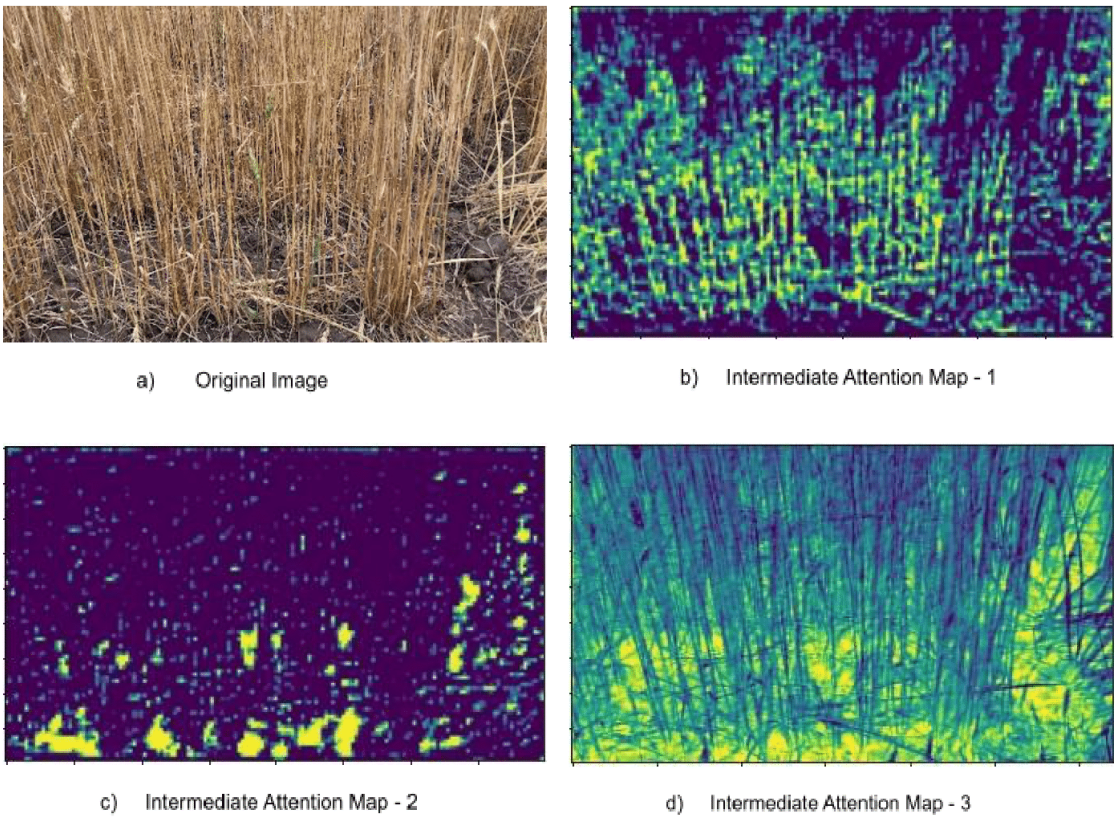

Figure 4.3: Some intermediate attention maps focusing on ground rather than stems

Chapter 5

Wheat Stem Rust Presence Detection

5.1 Introduction

The CI prediction gave some important insights that with the current dataset, the coefficient could not be directly estimated from the image due to the skewed nature and limited diversity of the dataset. The more critical insight was, the model was not able to learn the area of presence of the stem rust from the coefficient of infection labels. Hence, an understanding had to be developed about the ability of the approach to first identify the presence area of stem rust from the image which would be a step toward determining the coefficient directly as it would be a function of the area infected. Furthermore, identifying the presence would also be an easier problem to solve than a regression problem given the nature of the dataset.

5.2 Problem Formulation

The problem was formulated as a binary classification problem in which given an input image of a wheat plot. The work builds an approach with the goal to classify the image

on the basis of the presence of stem rust. It is classified as low if stem rust is absent and as high if there is a presence of stem rust.

5.3 Coefficient of Infection Range Selection

The dataset for identifying the presence of stem rust was built by abstracting ranges of CI values into a particular class. The range of CI values was chosen on the basis of maximum separation between low and high CI value ranges to simplify the identification of the presence of stem rust while keeping the data balance and diversity within the class. Furthermore, there was no presence of plots with no stem rust. Thus, low CI values were used as a proxy for no presence of stem rust. Taking these aspects into consideration the final range for the low class was chosen to be 0 to 6.5 while the range for the high class was chosen to be 35 - 100.

5.4 Approach

The approach uses the same idea of the CNN to identify the presence or absence of stem rust from an image of wheat plot.

5.4.1 Convolution Neural Network-based Detection

The architecture is exactly the same as that of the CI value prediction. The Resnet-50 [25] architecture was again used as a feature extractor similar to the approach for the CI determination. The output of the Resnet - 50 [25] final global pooling layer was fed to six fully connected layers - 1024, 512, 256, 128, 64, 32. These fully connected layers have relu nonlinearity added to extract features from the 1-D embedding representing the features of CNN. The output is then fed to the output layer which is again a single node. The activation function for the single node was a sigmoid compared to the linear activation function used for CI value determination as it is a task of binary classification.

Data Preparation

The image being of high resolution cannot be directly fed into the network due to computational complexity. Hence, the image was resized to 880×360 by maintaining the aspect ratio. All the images were resized to the stated size.

The train test split was similar to that of CI value prediction in which the StPaul plots were divided into two distributions in which the first half was used for training while the other was for testing. The labels were assigned on the basis of the CI value of the plot. A label of low (0) was assigned to the plots with a CI value between 0 - 6.5 while a high (1) for the CI range of 35 - 100. Additionally, the plots not in the range of both classes were ignored. The step was done for both the training and testing set.

After the labeling process, presence of a class imbalance was observed in the training set with majority of the samples being part of low. Hence, Random Under-sampling was also performed to have balanced image distribution. The undersampling process involves the removal of random images from each class. Table 5.1 shows the final train and test distribution.

Dataset	Low Number of Images	High Number of Images
Training	2015	1134
Testing	1116	1079

Table 5.1: Dataset Distribution

Network Training

The training procedure includes loading the Resnet-50 feature extractor ImageNet weights. The fully connected layer was initialized with random weights. The network was trained using binary cross-entropy loss for 50 epochs using the Adam optimizer. The learning rate was set to 0.01 with a batch size of 32.

Results

The trained model was used to determine the presence of stem rust on the test set. Quantitative analysis of the model was performed using the calculation of the evaluation metrics for the test set. Table 5.2 shows the evaluation metrics performance while Figure 5.1 shows the confusion matrix for the test set. It can be observed from the confusion matrix and evaluation metrics that the model is able to separate between low and high disease. Hence, further analysis for understanding the reliability of the approach was performed using qualitative analysis.

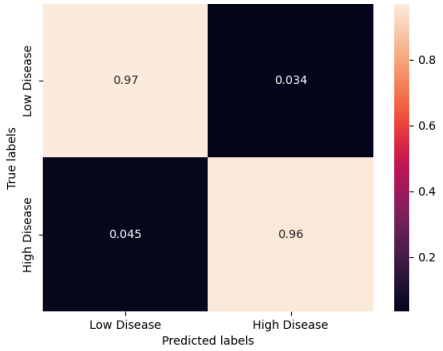

Figure 5.1: Resnet low vs high confusion matrix

Metrics	Score
Accuracy Score	0.96
Precision Score	0.964
Recall Score	0.955

Table 5.2: Resnet Low vs High Metrics

The qualitative analysis was performed similarly to the CI value predictions. The Gradcam attention maps [26] were drawn for the intermediate layer to understand the focus area of the network on the original image for the predictions. Figure 5.2 and Figure 5.3 show the attention maps for high and low stem rust disease. It can be observed that the new stem rust presence labels used for training the model supervised it to focus more on the stem area for the high disease image as increased activation can be observed in the stem region. The low disease still shows the prediction is based on different parts of the image as the focus area is spread all across the image.

The observed spread of attention is a similar behavior to that of the CI value estimation and might be because of the small receptive (network area of focus) field of the

CNN which allows the network to only learn the features important for the task based on the multiple small independent neighborhood fields. Hence, different focus areas are independently considered important causing the spread of attention map. The dependency between focus areas could be added if the receptive field of the whole image is considered during feature learning resulting in attention being centered on one area.

Visual attention mechanisms have been known to have a full receptive field allowing it to learn the proper area of focus in a data-driven manner and solve classification tasks where the focus needs to be on one particular part of the object. Therefore, a visual attention approach is needed which will result in predictions only based on the area of the stems which would assure the reliability of the predictions.

Figure 5.2: Attention focused on stems for high disease

Figure 5.3: Attention spread out rather than on stems for low disease

5.4.2 Attention-based Convolution Neural Network Detection

Fine-grained visual classification is the problem of differentiating between subcategories within the same category which is a particularly difficult problem to solve as the subcategories differ subtly. Hence, subtle differences need to be focused upon for distinction between subcategories. Visual Attention architectures [27, 28, 29, 30, 31, 32] have attained state-of-the-art performance for the fine-grained visual classification problem, particularly on the Stanford Cars [33], Dogs [34], and CUB-200-2011 [35] dataset. The mechanism divides the image into patches and learns to focus on the right patch in a data-driven manner. Furthermore, the receptive field of attention approach is also of a full image which means the network observes the full image and learns to give importance to parts of it which in CNN, depends on the small kernel size neighborhood of the network. The full receptive field is essentially the reason for the success of visual attention mechanisms on fine-grained visual classification tasks. Thus, the visual attention mechanism provides a promising direction for our problem as well because the idea is to learn the important area of the presence of stem rust using the data which would ensure the predictions are on the basis of only stems rather than other areas in turn increasing the reliability of the predictions. The Attention-based Convolution Neural Network approach uses the Focalnet [36] architecture for understanding the focus area for identifying the presence of stem rust.

The architecture presents a structure known as the Focal Modulation [36] block which is an alternative to the self-attention mechanism. The block also has shown the ability to provide interpretable attention maps and have better computational efficiency compared to the self-attention architecture. Figure 5.4 shows the comparison between the Focal modulation block and the self-attention block. It can be observed that Focalnet [36] aggregates the input first to extract the features at different levels from global to local parts of the image, unlike late aggregation which is performed in self-attention. The idea is, early aggregation would reduce the computational complexity compared to

self-attention by interacting with a static number of features than being directly proportionate to the number of visual tokens. This also results in learning more interpretable areas than the self-attention mechanism. Furthermore, the task of stem rust detection needs high throughput. Hence, such an architecture would be much more useful for solving the problem.

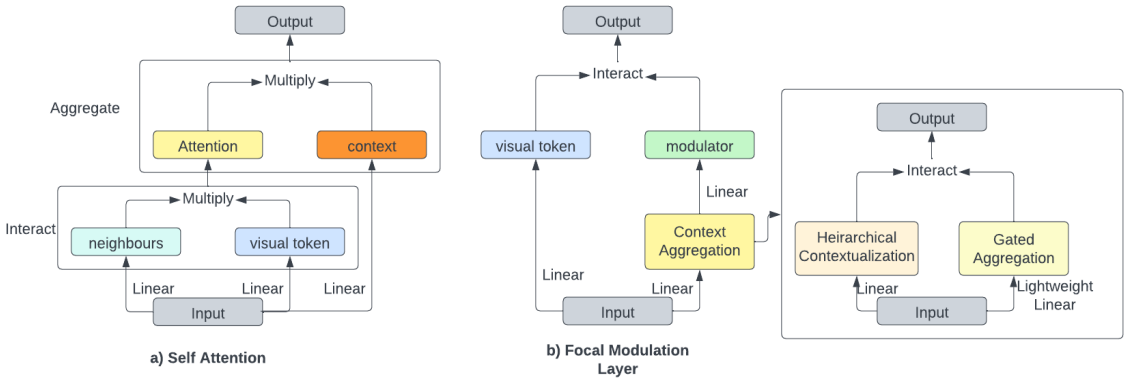

Figure 5.4: Efficient computation complexity through early aggregation for Focalnet compared to late aggregation in self-attention

The FocalNet-T feature extractor model with one output node added to the final average pooling layer was used for stem rust presence detection. The output was again applied with sigmoid activation due to the binary classification nature of the problem.

Network Training

The data preparation was performed in the same manner as CNN-based presence detection. The Focalnet [36] network was loaded with ImageNet weights and the network was then trained for 50 epochs for optimizing the binary cross-entropy error. The optimizer was chosen to be AdamW optimizer with a learning rate of 0.000005, Beta1 and Beta2 values to be 0.9 and 0.999, epsilon to be 1e-8, and a weight decay of 0.05. The batch size was chosen to be 32.

5.4.3 Results

The trained model is then evaluated on the test set. The quantitative analysis was performed using the evaluation metrics defined for the CNN-based presence detection. Table 5.3 shows the performance of the evaluation metrics on the test set and Figure 5.5 shows the confusion matrix of the attention approach. It can be observed that the attention model is also able to highly discriminate between the low and the high class. Though the quantitative measures are lesser than Resnet, the idea of using the attention module is to have the inference area only on the basis of stems. Hence, qualitative analysis was performed to understand if the predictions are only based on the stem area.

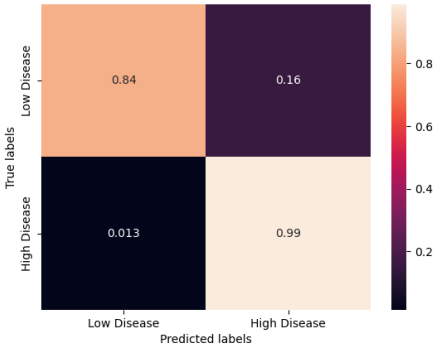

Figure 5.5: Focalnet low vs high confusion matrix

Metrics	Score
Accuracy Score	0.9123
Precision Score	0.856
Recall Score	0.987

Table 5.3: Focalnet Low vs High Metrics

The qualitative analysis was performed using the Gradcam attention maps. The maps were drawn for the modulator of the Focalnet [36] which determines the focus area of the network. Figure 5.6 and Figure 5.7 show the attention map for low and high disease. It can be observed that, unlike the CNN-based detection, the area of focus for the Focalnet [36] is based on the stem area for both low and high disease which is the behavior expected from the approach and is more human-interpretable. Now that, the focus area has been tuned to the requirement of the problem, the approach needs to be developed to move closer to automated CI value estimation.

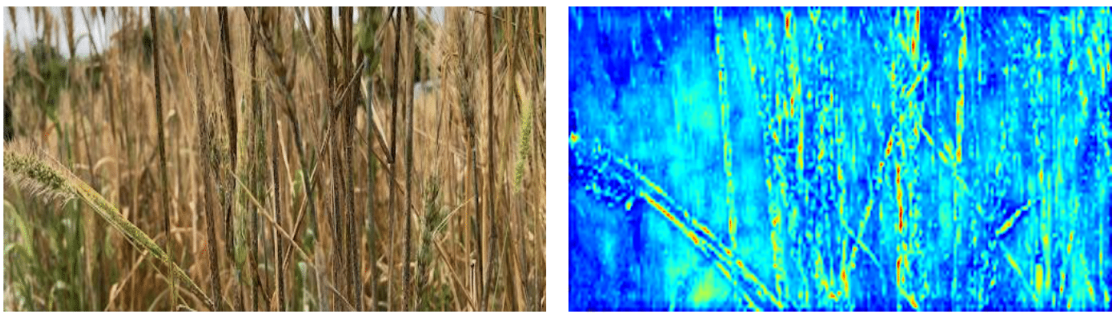

Figure 5.6: Attention focuses on stem area for high disease

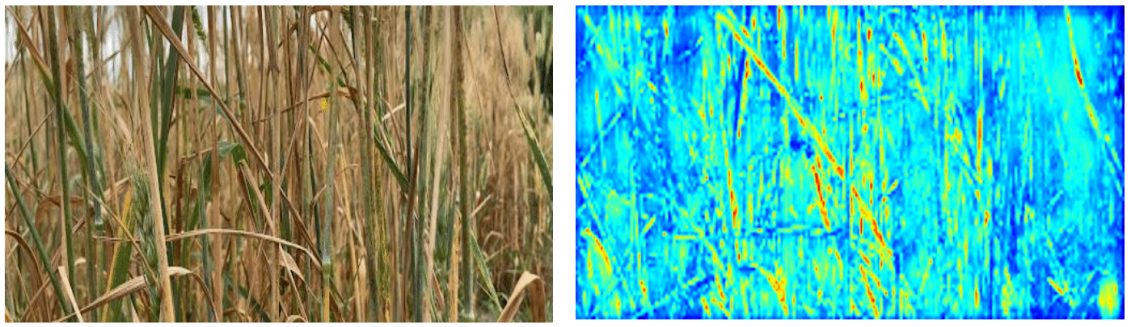

Figure 5.7: Attention focuses on stem area for low disease

Chapter 6

Wheat Stem Rust Intensity Detection

6.1 Introduction

Automated grading aims to determine the coefficient of infection directly from the image. The initial approach for CI value suggested the nature of the dataset prevents building an accurate model for grading from images directly. The stem rust presence detection indicated that the developed approach is able to reliably determine the presence of stem rust. Considering the insights from both, the study formulates the problem of automated grading as a multi-class classification problem in which the approach determines the abstract range of CI values from the image from the plot. The abstract range of CI values would be informative for the experts to grade the plot as it provides a narrow view of the CI values for the plot. Furthermore, it will be also useful for being deployed on the fields through mobile robots to warn the experts during different stages of stem rust.

6.2 Problem Formulation

The problem formulation for the study is that of a multi-class classification problem. Given the input image of a wheat plot. The work builds an approach with the goal to classify the image on the basis of the intensity of stem rust. It is classified as low, medium, and high which has a different range of CI values representing different intensities of stem rust.

6.3 Coefficient of Infection Range Selection

The CI range for the low, medium, and high was determined on the basis of the high separation between classes for easy distinction and an equal number of plots for balanced data points within the classes. Thus, the range selected for the low class was 0 to 6.5, the medium class was 12.88 to 18.66, and the high class was 31.5 to 100.

6.4 Attention-based Convolution Neural Network Detection

The Focalnet-T [36] attention model was used for the distinction of different intensities of stem rust. The only change performed from the presence detection is in the output layer. The output layer in presence detection had the size of one node with sigmoid activation as the task was binary classification. In the intensity detection problem, the output node has the size of three nodes with softmax activation as the task is multi-class classification. The softmax activation is used as it provides the probability distribution over the classes.

6.4.1 Data Prepration

The data preparation is the same as the presence detection study. The images were resized to 880×360 and the labels to the plots were allocated on the basis of the CI

ranges for the classes. The plots not belonging to any of the CI ranges were ignored. The number of images per class after the preparation was again imbalanced in nature. Hence, Random Undersampling was performed to attain equal distribution between classes. Table 6.1 shows the statistics per class after the processing.

Dataset	Low Number of Images	Medium Number of Images	High Number of Images
Training	2015	1453	1602
Testing	1116	1736	1079

Table 6.1: Dataset Distribution

6.4.2 Network Training

The Focalnet [36] model was loaded with ImageNet weights to ensure faster convergence. It was trained for 50 epochs for optimizing the Cross-Entropy Loss. The setting of the optimizer was similar to presence detection. The AdamW optimizer was used with a learning rate of 0.000005, Beta1 and Beta2 values to be 0.9 and 0.999, epsilon to be 1e-8, and a weight decay of 0.05. The batch size was chosen to be 32

6.4.3 Results

The trained model was then used to infer the test set. The quantitative analysis was performed using the evaluation metrics which were the same as the presence detection. Table 6.2 shows the performance of the evaluation metrics while Figure 6.1 shows the confusion matrix of the intensity detection. It can be observed that the low and high are able to be distinguished properly but the medium is getting ambiguous predictions between low and high. Hence, a better understanding of the ambiguity is done using embedding visualization analysis.

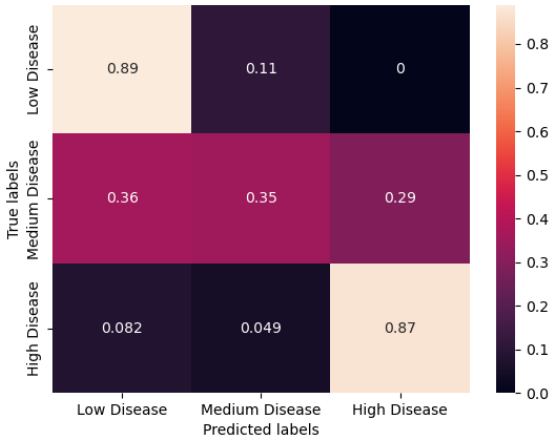

Figure 6.1: Focalnet low vs medium vs high confusion matrix: ambiguous medium class

Metrics	Score
Accuracy Score	0.703
Precision Score	0.65
Recall Score	0.45

Table 6.2: Focalnet Low vs Medium vs High Metrics

The embedding visualization analysis was performed by graphically visualizing the global average pooling embedding for data points of each class. The embedding inference was performed for each image in the test set. The embedding is dimensionally reduced to two dimensions for visualization using UMAP [37] dimensional reduction due to the nonlinear form of data. Figure 6.2 shows the clustering analysis of the embeddings. It can be observed from the embedding visualization that the medium class is equally distributed among low and high resulting in the ambiguity between classes.

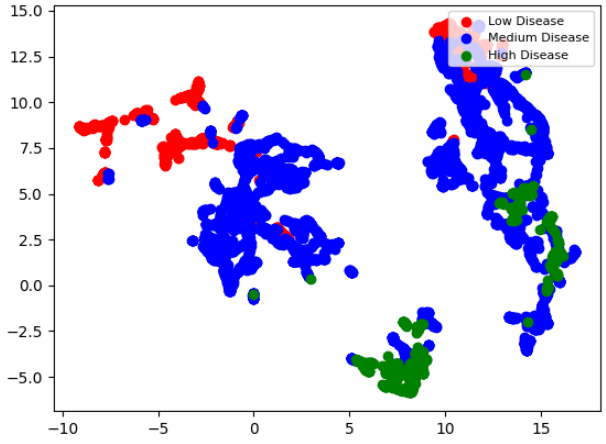

Figure 6.2: Medium cluster ambiguously distributed among low and high cluster

6.5 Improving Stem Rust Intensity Detection

6.5.1 Data Preparation Modification Experiment

Visual attention was not able to differentiate between the different intensity classes. The data was then visually inspected for manual distinction in which it was observed that the current resizing of images causes high information loss which does not even allow manual distinction. Hence, the resizing factor was reduced to have an equal trade-off between proper manual distinction and computational complexity. The factor was then chosen to be two instead of four which means the images were now resized to 1760×720 instead of 880×360.

The CI values of the plots in the three classes were also inspected for diversity. It was observed that the medium class had plots belonging to only two CI values which shows that the medium class did not have any diversity. Hence, to increase diversity the CI value range of the medium class was changed. The final range was 14 to 24 instead of 12.88 to 18.66.

The model was trained using the stated setting and then the trained model was used to infer on the test set. The quantitative analysis was performed using the evaluation metrics in the same way as the original intensity detection approach. Table 6.3 shows the evaluation metrics performance on the test set while Figure 6.3 shows the confusion matrix for the three class classifications after modifying the data preparation. It can be observed that the medium class is correctly distinguished as compared to being ambiguous between low and high. The only issue observed is the model marginally getting overfitted to the medium class as the detection of low and high have reduced. The overfitting problem can be solved using data augmentation.

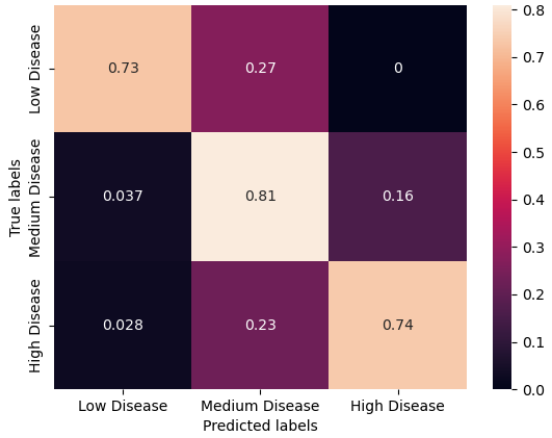

Figure 6.3: Modified data preparation confusion matrix: medium class distinguished from low and high

Metrics	Score
Accuracy Score	0.7679
Precision Score	0.777
Recall Score	0.767

Table 6.3: Modified Data Preparation Metrics

6.5.2 Increasing Detection Robustness

Image augmentation is a process of increasing the diversity of the dataset by introducing variations in color, brightness, and position in the images which would in turn increase the robustness of the model to react to such changes in the test scenario resulting in increased generalization and reduced overfitting. Image horizontal flipping, random blurring, and random brightness change were the augmentations introduced in the training dataset. These operations would ensure positional, blur and brightness invariance in the model which is essential when working in a field setting.

The Focalnet-T [36] model is then trained on the augmented dataset and evaluated on the test set using the quantitative metrics defined in the original intensity detection approach. Table 6.4 shows the evaluation metrics performance on the test set. Figure 6.4 shows the confusion matrix of the stem rust varied intensity detection. It can be observed that image augmentation has increased the generalization performance of the model by having equal detection accuracy of the three classes.

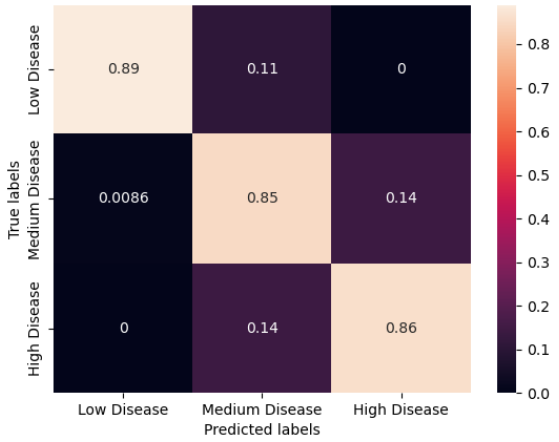

Figure 6.4: Image augmentation confusion matrix: overfitting to medium class reduced

Metrics	Score
Accuracy Score	0.8659
Precision Score	0.870
Recall Score	0.865

Table 6.4: Image Augmentation Metrics

The detection performance is also compared with the Resnet-50 model trained on the same setting as the Focalnet model. Figure 6.5 shows the confusion matrix of the low, medium, and high classes detection with the Resnet architecture. It is observed that with the varied intensity detection, the Focalnet visual attention approach performs better than the CNN approach both qualitatively and quantitatively. Hence, centering the attention does provide reliable and better predictions than the large CNN-based approaches.

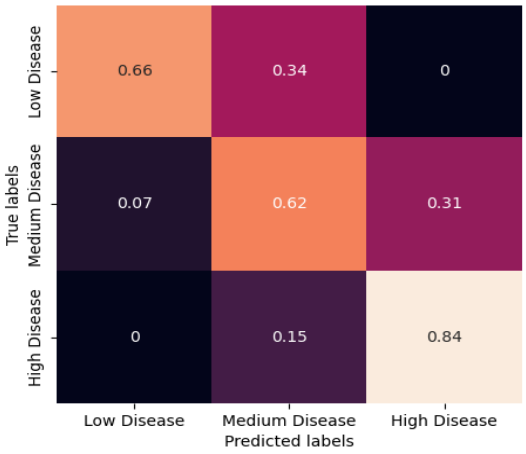

Figure 6.5: Resnet performance on varied intensity detection

6.6 Scale Normalization Combining Procedure

The low, medium, and high classes are being distinguished in a generalized manner using the data augmentation procedure. Further improvement of the detection performance requires more diverse data points. Hence, a procedure had to be developed to introduce Rosemount data which would add diversity to the StPaul dataset. It was observed previously during the CI value regression experiment that the introduction of Rosemount data reduces the performance due to the scale being different in the two datasets. Hence, a scale normalization procedure had to be developed to combine both datasets.

The scale normalization process was experimented both ways in which the StPaul Scale was normalized for Rosemount, and Rosemount Scale was normalized for StPaul. The procedure included calculating the stem pixel width for both datasets and using it as a resizing factor to normalize the scale. The Rosemount to StPaul scale normalization involves resizing the images of StPaul to 1760×720. The pixel stem width is then measured of random ten stems from different images for both datasets. The mean ratio of the stem width between Rosemount to StPaul is then used to resize the Rosemount dataset. The resultant Rosemount dataset is again resized to 1760×720 for creating the final combined dataset. The StPaul to Rosemount scale normalization does not resize the StPaul data. The mean ratio of the stem width between StPaul to Rosemount is then used to resize the StPaul dataset. After both the StPaul and Rosemount are brought to the same scale, both datasets are resized to 1760×720 to create the final combined dataset.

The combined datasets are then used for training the Focalnet-T [36] model and evaluation is performed on the test set using the same steps as original intensity detection. Table 6.5 and 6.6 show the performance of the evaluation metrics on the test set. 6.6 shows the confusion matrix of per-class detection after combining the Rosemount and StPaul dataset. It can be observed that the behavior similar to CI value regression is not observed as the performance is maintained even after combining the datasets.

The model is overfitted to the low disease due to the Rosemount data being skewed to the lower end. Hence, the addition of diversity in other classes would improve the detection performance of all classes.

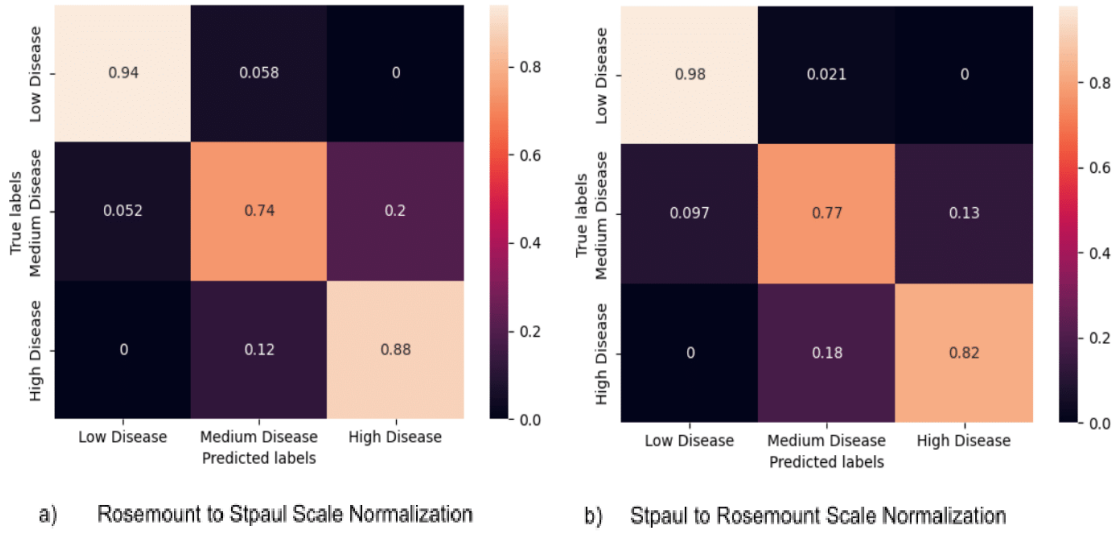

Figure 6.6: Scale normalization results: observed no degradation in performance after combining scale normalized datasets. Overfitting to low due to skewed Rosemount dataset.

Metrics	Score
Accuracy Score	0.838
Precision Score	0.846
Recall Score	0.838

Table 6.5: Rosemount to StPaul Metrics

Metrics	Score
Accuracy Score	0.843
Precision Score	0.843
Recall Score	0.843

Table 6.6: StPaul to Rosemount Metrics

Chapter 7

Conclusion and Discussion

This research studies the critical problem of automated stem rust grading. The automated grading methodology should determine the CI value based on a single image of a wheat plot. It was observed that the approach of learning to estimate the CI values directly from image-CI value pairs did not provide strong supervision for the network. Hence, this research shows an intensity-based multi-class classification formulation of the grading problem provides a better estimate compared to a regression task as it guides the network to determine the intensity based on only the stem region where the disease is generally present. Additionally, the effectiveness of the Focalnet visual attention approach for stem rust grading was also shown as it reliably detected different intensities of the disease both qualitatively and quantitatively compared to a large CNN approach. There still is a lot of scope for improvement in the detection performance which could be refined with the addition of more diverse data points to each class. This improvement was also illustrated by using the developed normalization procedure for combining two datasets captured at different scales which resulted in better detection performance of the low disease class after the addition of new diverse data points from the Rosemount dataset.

The future work includes on-field deployment of the visual attention approach on

mobile robots for high throughput automated intensity detection but the main focus would be to develop multiple finer classes covering the full range of CI values rather than three coarse range classes used in this research resulting in an orderly translation to an automated grading approach. Furthermore, the robots can also be equipped with fungicide treatment mechanisms which can be sprayed based on different intensities of stem rust for high quality yield extraction of wheat. View planning aspects can also be explored for capturing the most optimal and clear view of the plot for proper and efficient stem rust detection which would help reduce the time spent by mobile robots on each plot for the grading process. Lastly, this research can also be extended to multiple wheat diseases like Leaf Rust and Tan Spot which would allow having one general reusable framework for wheat disease identification.

References

[1] Alan P Roelfs. *Rust diseases of wheat: concepts and methods of disease management*. Cimmyt, 1992.

[2] H. Jesse Dubin, Maarten van Ginkel, and Shanmugasundaram Nagarajan. Rust diseases of wheat. 2009.

[3] SANNAKKI SS, RAJPUROHIT VS, NARGUND VB, ARUN KUMAR, and YALLUR PS. A hybrid intelligent system for automated pomegranate disease detection and grading. *Int. J. Mach. Intell*, 3(2):36–44, 2011.

[4] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016.

[5] Saraansh Baranwal, Siddhant Khandelwal, and Anuja Arora. Deep learning convolutional neural network for apple leaves disease detection. In *Proceedings of international conference on sustainable computing in science, technology and management (SUSCOM), Amity University Rajasthan, Jaipur-India*, 2019.

[6] Serawork Wallelign, Mihai Polceanu, and Cédric Buche. Soybean plant disease identification using convolutional neural network. In *FLAIRS conference*, pages 146–151, 2018.

[7] H Sabrol and K Satish. Tomato plant disease classification in digital images using classification tree. In *2016 international conference on communication and signal processing (ICCSP)*, pages 1242–1246. IEEE, 2016.

[8] Alexander Johannes, Artzai Picon, Aitor Alvarez-Gila, Jone Echazarra, Sergio Rodriguez-Vaamonde, Ana Díez Navajas, and Amaia Ortiz-Barredo. Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. *Computers and electronics in agriculture*, 138:200–209, 2017.

[9] Jiang Lu, Jie Hu, Guannan Zhao, Fenghua Mei, and Changshui Zhang. An in-field automatic wheat disease diagnosis system. *Computers and electronics in agriculture*, 142:369–379, 2017.

[10] Michael Schirrmann, Niels Landwehr, Antje Giebel, Andreas Garz, and Karl-Heinz Dammer. Early detection of stripe rust in winter wheat using deep residual neural networks. *Frontiers in Plant Science*, 12:469689, 2021.

[11] Qian Pan, Maofang Gao, Pingbo Wu, Jingwen Yan, and Mohamed AE Abdel-Rahman. Image classification of wheat rust based on ensemble learning. *Sensors*, 22(16):6047, 2022.

[12] Zhou Tang, Meinan Wang, Michael Schirrmann, Karl-Heinz Dammer, Xianran Li, Robert Brueggeman, Sindhuja Sankaran, Arron H Carter, Michael O Pumphrey, Yang Hu, et al. Affordable high throughput field detection of wheat stripe rust using deep learning with semi-automated image labeling. *Computers and Electronics in Agriculture*, 207:107709, 2023.

[13] Muhammad Hassan Maqsood, Rafia Mumtaz, Ihsan Ul Haq, Uferah Shafi, Syed Mohammad Hassan Zaidi, and Maryam Hafeez. Super resolution generative adversarial network (srgans) for wheat stripe rust classification. *Sensors*, 21(23):7903, 2021.

[14] Zhiwen Mi, Xudong Zhang, Jinya Su, Dejun Han, and Baofeng Su. Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices. *Frontiers in plant science*, 11:558126, 2020.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[17] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pages I–I. Ieee, 2001.

[18] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[20] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[21] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[22] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer, 2022.

[23] Tianhe Ren, Jianwei Yang, Shilong Liu, Ailing Zeng, Feng Li, Hao Zhang, Hongyang Li, Zhaoyang Zeng, and Lei Zhang. A strong and reproducible object detector with only public datasets. *arXiv preprint arXiv:2304.13027*, 2023.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[27] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[28] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *Computer Vision–ECCV 2022: 17th*

European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV, pages 74–92. Springer, 2022.

[29] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2956–2964, 2015.

[30] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4438–4446, 2017.

[31] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[32] Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, et al. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems*, 34:25346–25358, 2021.

[33] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

[34] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[35] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[36] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 35:4203–4217, 2022.

[37] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.