Norges miljø- og
biovitenskapelige
universitet

Master's Thesis 2024 30 ECTS

Faculty of Science and Technology

Precision Agriculture: Leveraging Deep Learning for Classification and Segmentation of Paddy Diseases

Mahrin Tasfe

Master of Science in Data Science

Abstract

The early detection of paddy disease is essential for reducing the usage of chemical substances and pesticides, and preventing local and global transmission of diseases. An automated paddy disease diagnosis system makes this possible, and helps to improve crop production and the overall health of rice plants. For this study, we have explored the most prevalent paddy diseases and their visible symptoms. Additionally, the four segmentation models (UNet, VGG16 UNet, TransUNet and Deep Residual UNet) and four classification models (DenseNet, MobileNet, Vision transformer (ViT) and a custom ensemble model of DenseNet121 and Xception), have been reviewed with a comparative analysis highlighting their structural differences, advantages, and limitations. The significant research gaps in this domain have been identified and to address the lack of open-access paddy disease segmentation datasets, we have created a novel paddy disease segmentation dataset using image processing techniques. The applicability of the above-mentioned segmentation models has been evaluated for paddy diseases using this newly created dataset and we have identified that Deep Residual UNet is the most suitable model to be used in resource constraint applications—considering its quantitative and qualitative performance, model size and structural advantages and limitations. Furthermore, we have investigated the impact of training these models with a significantly higher number of augmented images—more than double the original dataset—and observed that while the quantitative performance increased with increased data, the qualitative performance degraded in a few cases. Moreover, due to the huge computational requirements and the data-hungry nature of ViTs, we assessed whether its performance could be achieved with traditional models or their ensembles and found it to be feasible. Additionally, we have explored the effect of augmentation intensity on the above-mentioned classification models.

Declaration of Originality

I hereby declare that the work presented in this thesis is my own unless otherwise stated. To the best of my knowledge the work is original and ideas developed in collaboration with others have been appropriately referenced.

Acknowledgments

I am grateful to everyone who has supported me during this thesis with academic feedback, guidance, and both moral and emotional support.

First, I would like to thank my supervisors, Associate Professor Dr. Habib Ullah and Associate Professor Dr. Habil. Fadi Al Machot for their excellent guidance, support and feedback. Their support, prompt email responses—even during late hours—and patience with my numerous questions have been immensely helpful, especially in resolving my confusion related to coding and structuring the contents. Their encouragement to remain curious and explorative has also been very impactful.

Additionally, I express my gratitude to Associate Professor Dr. Martin Thomas Horsch for his significant guidance in writing a comprehensive scientific research document like this one. His directions during the DAT390 course have resulted in the immense improvement of my research writing capabilities, which I hope is appropriately reflected in this thesis.

Lastly, I am thankful to all my friends, especially AKM Nivrito, Mats Hoem Olsen, and Synne Wu Kofoed for their unwavering support. During hectic moments, their support was crucial in keeping me going. Good friends are invaluable, and I am forever grateful for their friendship.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Introduction

Globally, rice has been one of the most impactful and widely consumed foods as over half of the global population consumes it [1]–[5]. Rice plants before harvesting are usually called 'paddy' and when the crop has been harvested, it is called 'rice'. Nations such as China, India, Bangladesh etc. are majorly dependent on rice and rice production has an immense impact on their economic growth and food security [1], [2], [6]–[8]. The importance of rice production is also increasing due to the expected global population growing over 9 billion people by 2050—increasing the demand for the crops massively [2], [3], [5]. However, various factors such as climate change, global warming, and diseases provide major challenges in paddy production—where disease can cause massive production losses up to 40% [4], [8], [9].

Recent developments in the field of precision agriculture involving computer vision and machine learning techniques can immensely help the paddy disease diagnosis by mitigating the associated challenges [7], [8]. In addition to providing correct diagnosis, these automated disease diagnosis systems can also offer disease management advice to the farmers which helps to limit the overuse of chemical substances, optimise resource usage, reduce environmental degradation, and improve the overall ecosystem health. Additionally, it also aids in minimising the economic and yield losses, the need for erroneous manual disease detection, and the limitations related to human bias and the availability of the plant experts [1], [4], [7], [9]–[12].

To conclude, the high importance of rice production in the global context demands researchers in this field to provide solutions using a combination of precision agriculture, and advanced image processing and machine learning techniques to alleviate challenges related to paddy disease diagnosis. Such technologies have demonstrated promising results on paddy disease detection in various studies [6]–[8], [10], [13]–[26] and can benefit farmers globally by aiding early disease detection, and therefore, ensuring a positive impact on the global food security.

1.2 Thesis motivation

Even with the recent research efforts, the process of paddy disease detection faces several challenges. The motivating factors for our research have been presented in Figure 1.1, with further details provided below.

- **Limitations of manual detection:** Manual disease detection is often plagued with many limitations. Manual observations are subjective and fallible as paddy diseases can show overlapping and similar symptoms. The lack of insights, the slow turnaround of laboratory diagnosis, and the lack of access to plant experts can render farmers' observations faulty [1], [10], [13]. Also, it can be demanding with time and effort—especially in vast fields—making such inspection an arduous and expensive task [4], [7], [9], [11], [12]. Hence, there is a crucial need for fast, accessible, and automated solutions to aid this [1], [10], [11].
- **Need for efficient disease diagnosis:** Efficient automated disease detection systems can decrease the need for human consultations, the time needed for diagnosis, total resource expenditure, and the economic damage [4], [6], [9], [27].

Figure 1.1: Illustration of the motivating factors behind our research.

- **Limited access to specialists:** Lack of availability of expert plant consultants, particularly in remote areas, can lead to inaccurate disease diagnosis and expensive disease management. This often results in loss of yield and quality of the crop [3], [6], [8], [9], [27].
- **Global spread of diseases:** Based on recent research, there is an urgent need for a correct, prompt, and globally applicable image-based disease identification system. Because, disease diagnosis and disease management can become complicated due to the variable growth rates and stages of inception for such diseases, and their potential to spread globally and form new diseases [11].
- **Advances in image processing:** Advancements in image processing technologies can contribute to disease detection to be more accurate, rapid, and less resource demanding—while easing large-scale production monitoring and aiding in increased yield [4], [10].

In summary, applications of emerging technologies to craft effective automated plant disease detection can benefit farmers, consumers, and crop-producing nations as they alleviate farming challenges with prompt and precise interventions that can lead to improved production, sustainability, efficient management, and healthy ecosystem [9], [10].

1.3 Problem statement

Due to the problems associated with manual paddy disease diagnosis, we need an automated disease diagnosis system capable of handling complex samples with high accuracy for proper disease management initiatives. Paddy disease data has several complications related to background shadows, noises, similar and overlapping disease patterns, and lighting variations. To resolve these challenges, segmentation is needed as a preprocessing step or incorporated within the classification process to extract the diseases. Currently, there are no open-access paddy disease segmentation datasets available. We will address this gap by providing a segmentation dataset and analysing the applicability of existing segmentation models to paddy disease. Furthermore, for farmers with constrained computational resources, state-of-the-art vision transformers (ViT) demanding large datasets and high computational resources might not always be suitable. Hence, for this thesis, we will analyse whether the performance of ViTs can be achieved with traditional models or their ensembles. Moreover, researchers often have to rely on data augmentation strategies to overcome the limitations of training small datasets and collecting large and diverse datasets. Therefore, we will analyse the effect of various augmentation intensities on the performance of both segmentation and classification models. To advance precision agriculture, this thesis aims to provide all required analysis to create an effective automated paddy diagnosis system.

1.4 Research objectives and research questions

The research objectives and goals that we plan to achieve in this thesis can be grouped into three categories such as (a) literature review-based research objectives, (b) classification task-based research objectives, and (c) segmentation task-based research objectives. The details of goals and objectives undertaken in each category have been given below.

(a) Literature review-based research objectives

We will provide an extensive review of the prevalent paddy diseases which includes their causative agents, visual symptoms, and potentially affected areas. In addition to this, the selected segmentation and classification models to be used for paddy disease diagnosis will be analysed—highlighting their limitations and advantages. Lastly, the prominent research gaps within this domain will be identified.

(b) Segmentation task-based research objectives and research questions

From the literature review presented in Chapter 2, we have identified that there are currently no open-access paddy disease segmentation datasets available and apart from the study by Daniya et al. [28], there have been no studies on paddy disease segmentation using masks. To address this research gap, we will introduce a paddy disease segmentation dataset and perform a comparative analysis on four segmentation models (UNet, Vgg16 UNet, TransUNet and Deep Residual UNet) previously applied in remote sensing [29], biomedical imaging [30]–[32], urban scene analysis [33], [34], and precision agriculture for crops other than paddy [35]–[38]. This study will evaluate their applicability for paddy disease segmentation. We will also provide a comparative analysis of the selected models in terms of their structure, advantages and disadvantages. Lastly, we aim to address the following research questions for the segmentation task.

- RQ1: Which mask-based segmentation model achieves the highest and lowest performance scores in intersection over union (IoU)?
- RQ2: Considering all parameters including model size, and qualitative and quantitative test performance, which mask-based segmentation model is the most suitable to be used as a preprocessing step or incorporated within the classification process?
- RQ3: If the number of augmented images is higher (more than double) than the original number of samples, does it have any negative or positive effect on the test performance using unmodified test samples (both qualitative and quantitative)?

(c) Classification task-based research objectives and research questions

The main objective of the classification task is to assist farmers in identifying or diagnosing paddy diseases quickly, minimising the dependency on plant specialists, and error-prone and time-intensive manual detection. Due to the inherent characteristics of state-of-the-art ViTs, these models are typically larger and require substantial computational resources, time, and data. Thus, ViTs are often not suitable for resource-constrained applications such as mobile phones, which are often the only available technology for many farmers—particularly in developing countries. Given the limited computational capabilities available to the farmers, there is a need for a memory-efficient model that offers performance comparable to ViTs. This study will assess whether traditional convolutional neural network (CNN) models and their ensembles can perform as well as the ViT models. Furthermore, since disease patterns are smaller in size and have high similarities, this research will also explore how varying augmentation intensities impact disease detection in a complex agricultural environment. Lastly, we aim to address the following research questions for the classification task.

- RQ4: For paddy disease classification with complex field data, can the performance of ViT be achieved with traditional CNN models (MobileNet, and DensNet121) or an ensemble of the traditional models?
- RQ5: How does the level of data augmentation (none, basic, and extensive) impact the performance of models classifying paddy diseases using datasets with complex field environments?

1.5 Thesis contribution

Through this thesis, we are making the following contributions to paddy disease diagnosis and, on a broader scale, to precision agriculture:

- We have identified the significant research gaps in this domain, which will guide future researchers to address these gaps.
- Research on paddy disease segmentation is limited, and currently there are no open-access datasets available for this purpose. To address this gap, we have created a paddy disease segmentation dataset using image processing techniques and this dataset can serve as a starting point for future researchers.
- We have compared the recent segmentation models in terms of their architecture, advantages, limitations, qualitative and quantitative results, and overall, their applicability to this domain which were previously used in other domains.
- Analysed the effects of using a significantly higher number of augmented images—more than double the original dataset—on both qualitative and quantitative test performance of the selected segmentation models.
- Analysed whether traditional models or their ensembles are capable of demonstrating comparable performance to the ViTs in classifying paddy diseases.
- Analysed the impact of various augmentation intensities on the test performance of the selected classification models.

1.6 Thesis outline

- Chapter 1 gives a concise overview of the problem statement, motivating factors behind this thesis, research objectives, goals and questions, and our contributions through this thesis.
- Chapter 2 provides a review of the selected segmentation and classification models for this study and a short description of the main research gaps identified in this domain. An extensive review of the relevant paddy diseases has also been provided in this chapter.
- Chapter 3 includes our detailed research methodology for the segmentation and classification tasks intending to address identified research gaps in paddy disease diagnosis.
- Chapter 4 holds a brief comparative analysis and illustration of the quantitative and qualitative results that we have obtained and analysis of the model learning curves.
- Chapter 5 provides a discussion focused on answering the research questions in addition to a brief discussion of our additional observations from the results. This chapter also includes limitations faced by this research work.
- Chapter 6 has an overview of our findings with a detailed outline for our future work.

2 Literature Review

This chapter provides a detailed overview of the common paddy diseases and their symptoms. Additionally, it gives an overview of the segmentation and classification models selected for this study and their comparative analysis—highlighting their structural differences, advantages and disadvantages.

2.1 Paddy diseases and their symptoms

Diseases caused by unfavourable conditions (imbalances in temperature and soil quality) and by microorganisms (fungi, viroids, bacteria, nematodes, and viruses) severely threaten global paddy cultivation and lead to various symptoms [1], [8]. The symptoms of such diseases in different plant parts (e.g., leaves, stems, and grains) can be unique as well as overlapping due to varied environmental factors—resulting in demands of precise identification [6], [11]. For an overview of the diseases in this context, they have been classified into four categories based on causative factors: bacteria, fungi, viruses, and pests. A concise description of prevailing diseases has been provided in this section—along with Figure 2.1 and Figure 2.2 illustrating such diseases and their subclasses. Table 2.1 provides a tabular representation of the most common paddy diseases highlighting their causative factors, disease symptoms, and possible affected areas.

2.1.1 Fungi-caused diseases

Various fungi attacking leaves, sheaths, grains, and other parts of the plant result in fungal infection. These infected plants generally have noticeable lesions, discolouration, and rotting—producing serious adverse impacts on the health and yield of rice crops [1], [3], [5], [10], [13], [39].

- **Blast disease (BD):** Throughout paddy's complete life cycle, this disease adversely affects leaves, collars, nodes, necks, panicles, and seeds [1], [3]. Mainly impacting nodes and neck tissues, it reduces the grain quality and quantity, and forms dark brownish-black spots on grains [39]. Also, it is visible on leaves as green-grey diamond-shaped lesions with dark green borders. The lesions later expand and form grey centres and dark brownish borders, and the leaf dies in the end [2], [3], [5].

- **Brown spot disease (BSD):** This disease initially materialises as small brown spots, which later develop into cylindrical, oval, and circular forms. It mainly infects leaves, coleoptiles, panicle branches, leaf sheaths, spikelets, and glumes [1], [6], [13], [40]. Contaminated seeds yield unfilled or discoloured grains and act as the primary source of infection—impacting future yields [10].

  - **Thin brown spot disease:** This disease mainly infects leaves at the later stages of paddy. It is predominant in potassium-deficient soils at temperatures between 25 to 28 degrees Celsius and appears as dark brown lesions along leaf veins [41].

- **Sheath blight:** Initially, this disease appears as greenish-grey spots on leaf sheaths near water levels, then progresses with irregular purple-brown and blackish-brown borders [5].

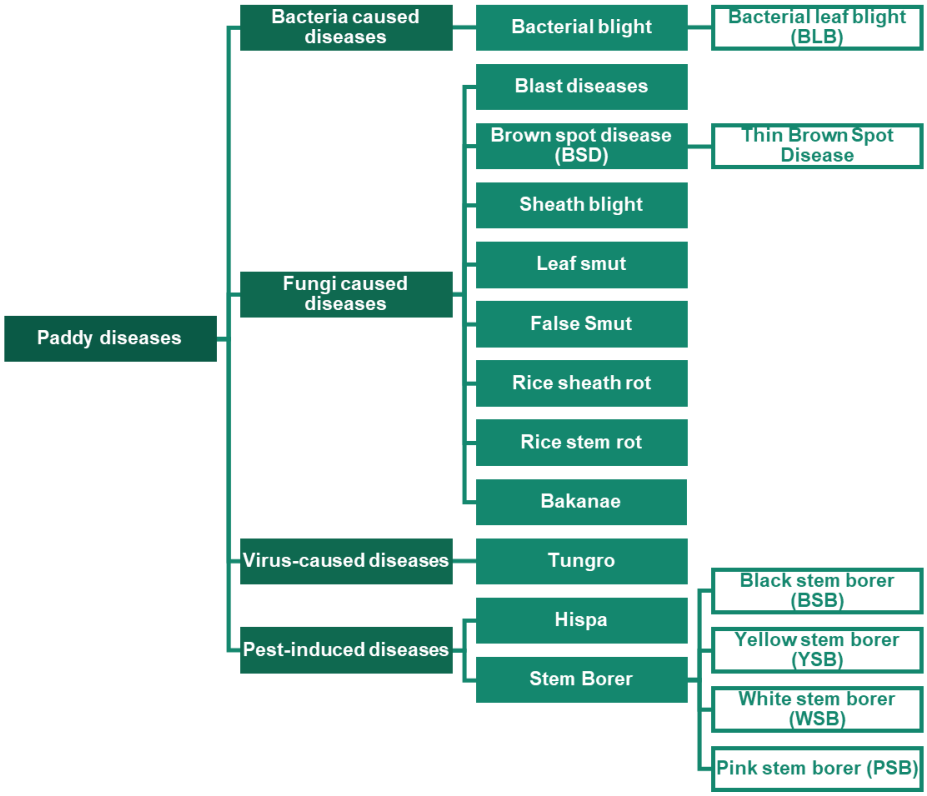


Figure 2.1: Illustration of common paddy diseases and their categories.

- **False smut:** This disease causes the individual grains to form clusters of velvety spore masses or yellow fruiting bodies. In the early stages, these clusters exhibit yellow or orange hues. When the spores mature, they transform into a greenish-black colour—indicating mature black fungal mycelium [3], [10]. Also, it often results in grain chalkiness and reduces both their quality and overall crop yield [3].

- **Leaf smut:** This disease appears in small black lesions with brown or golden circles on paddy leaf blades [3], [18]. When severe, the leaves turn yellow and the tips appear grey due to their demise [40].

- **Rice sheath rot:** This disease mainly transmits through seeds and crop residues [1], [10]. The infection starts as elongated or inconsistent dark markings on the top leaf sheath surrounding the young panicles. It often shows reddish-brown or brownish-tan patches with a grey or brownish-grey centre. When the disease advances, the darker dots within the patches increase [3], [10]. It disturbs panicle emergence, forming sterile panicles and unfilled discoloured seeds [3], [10]. Thus, it can significantly lower the yield while spreading to uninfected crops [1].

- **Rice stem rot:** This disease appears as small black lesions on the outer leaf sheath extending to the inner sheath—particularly at the water line in water-sown rice [3], [42]. The sheaths die and detach as the disease develops. In severe cases, it penetrates the culm and leads to visible white and black sclerotia and mycelium within. It causes lodging and affects grain with unfilled panicles and chalky grains, and is influenced by panicle moisture and nitrogen fertiliser [43]. This disease grows on wounded plants.

- **Bakanae:** This disease spreads primarily through contaminated seeds. However, it can also spread from infected plant material or soil—moving between plants via wind or water and during agricultural activities like harvesting and seed soaking. The pathogens attack the paddy's roots or crowns, resulting in symptoms as unusually tall plants with pale, slender, yellow-shaded, and dry leaves. It causes diminished tillering

Table 2.1: Analysis of the common paddy diseases.

Disease Name	Causative agent	Disease symptoms	Affected areas	References
Rice blast	Fungus	Dark brownish-black spots on grains; diamond-shaped lesions on leaves	Leaves, collars, necks, panicles, seeds, and grains	[1]–[3], [5], [13], [39]
Brown spot	Fungus	Circular dark brown lesions	Leaves, leaf sheath, panicle branches, and glumes	[1], [4]–[6], [10], [12], [13], [40]
Thin brown spots	Fungus	Dark brown lesions along leaf veins	Leaves	[41]
Sheath blight	Fungus	Greenish-grey spots with irregular purple-brown and blackish-brown borders	Leaf sheaths	[2], [5], [6], [39]
False smut	Fungus	Clusters of yellow fruiting bodies, orange to greenish-black spores	Rice grains	[3], [5], [6], [10], [39]
Leaf Smut	Fungus	Black linear lesions with dark gold or light brown halos	Leaves	[3], [18], [40]
Rice sheath rot	Fungus	Irregular spots with dark reddish-brown margins, grey centre, and later-phase darker dots	Leaf-sheaths	[1], [3], [10]
Rice stem rot	Fungus	Black lesions on leaf sheaths, expanding to the culm, with white and black sclerotia and mycelium inside infected stems, unfilled panicles and chalky grains	leaf sheaths and culms	[3], [42], [43]
Bakanae	Fungus	Pale, slender, yellow-shaded leaves, less tillering, white powdery growth on the lower portion of the plants, and lesions on roots	Seeds, seedbed, lower portion of the plants (roots, and crowns)	[39], [44]–[46]
Bacterial blight	Bacterium	Wilting seedlings, yellowing, drying leaves, small grey to olive spots	Leaves, spikes, and seedlings	[8], [10], [13]
Bacterial leaf blight	Bacterium	Grey-white lesions on leaf veins, yellow lesions on leaves	Leaves	[1], [5], [10], [18], [40]
Tungro	Viruses	Yellowing, mottled or striped appearance, rust-coloured spots, and interveinal necrosis	Leaves	[13], [47]
Hispa	Insect (pest)	White membranous leaves leading to wilting	Leaves	[4], [6], [12]
Yellow stem borer	Insect (pest)	“Dead heart” and “white ear” like shapes on leaves, dried, yellow shoots and chaffy ear heads	Shoots and leaves	[48], [49]
Pink stem borer	Insect (pest)	Insect waste and eggs on leaf sheath and stems	Leaf sheath and stems	[48], [50]

and grains that are either empty or partially filled [39], [45], [46]. It also results in white powdery growth on the lower portion of the plants [45].

2.1.2 Bacteria-caused diseases

Various types of bacteria cause paddy bacterial diseases, leading to wilting, yellowing, and lesions in different parts of the plant [2], [5], [18], [40].

- **Bacterial blight:** This disease starts as expanding wet brown streaks on leaves, drying into yellow droplets. Later, these leaves develop grey-white lesions indicating the infection’s end. The infected seedlings typically die within 2 to 3 weeks after the initial infection [8], [10].
  - **Bacterial leaf blight (BLB):** It shows elongated yellow lesions and often grey-to-white lesions on the leaves and leaf veins respectively [5], [18]. Through plant debris and insect feeding, this disease can be transmitted [1].

2.1.3 Virus-caused diseases

Various strains of viruses transmitted through insects can cause this disease and majorly impact crop yield and quality [13].

- **Tungro:** Frequently occurring during the vegetative phase, this disease is transmitted by the infected seedlings and stubble from previous crops and leafhoppers feeding on the infected plants. The infected plants develop yellow leaves which start at the tip and spread downward [13]. Also, mottled or striped patterns, rust-coloured spots, and necrosis between veins can appear on infected leaves [47].

Figure 2.2: Common paddy disease examples.

2.1.4 Pest-induced diseases

Hispa and stem borers are the most common pest-induced diseases.

- **Hispa:** *Dicladispa armigera* scrapes the paddy leaf surface leaving distinct white lines for female insects to lay eggs. After hatching, these larvae accelerate the leaf scraping process leading to the demise of the leaf [4], [12]. Hispa disease can spread across a field in three weeks without pesticide intervention [12].
- **Stem borers:** Six types of stem borer insects attack the rice stems and grains [6]. The symptoms include a "dead heart" in the vegetative stage, and "white heads" during the reproductive stage resulting in unfilled panicles and stem holes. Effective treatment requires distinguishing these symptoms from similar damage caused by other diseases and insects [48]. Depending on the insects causing it, stem borer diseases have their sub-classes such as Black stem borer (BSB), Yellow stem borer (YSB), Pink stem borer (PSB), White stem borer (PSB) etc.

2.2 Related works

Equipped with the recent developments in image processing and machine learning, researchers have developed various techniques for automated paddy disease diagnosis—enhancing performance while reducing manual effort and resource usage. A short overview of the selected segmentation and classification models, and the rationale behind their selection for this study has been presented below. We have also outlined the structural differences, advantages and limitations of the chosen models.

2.2.1 Selected segmentation models

Using the segmentation process, diseased regions are isolated from healthy parts of a leaf, removing the complex backgrounds and healthy parts of the plants which simplifies the disease classification task [51]. Sev-

eral studies have used this process employing masks [21], [28], image processing techniques [41], [51]–[54], and neural networks [53] for accurate disease classification. However, this process is challenging due to the variations in illumination, distance, and complex backgrounds [51]. The procedure for segmentation-based disease classification is shown in Figure 2.3.

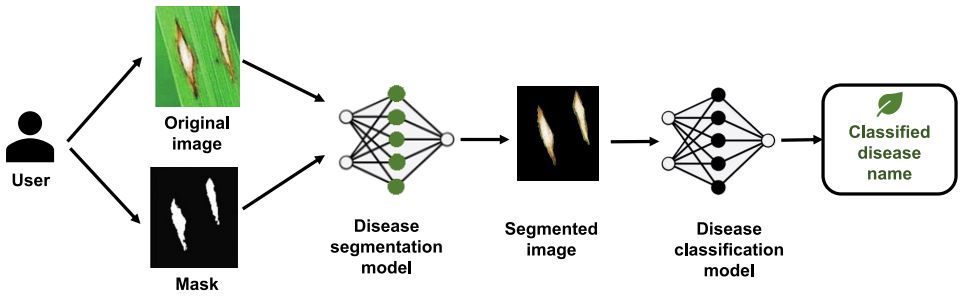


Figure 2.3: Demonstrating the segmentation-based paddy disease classification process.

For the segmentation task, we have selected four models which include UNet [31] and three of UNet's variants: (1) transfer learning-based: VGG16 UNet; (2) attention mechanism-based: TransUNet [55]; (3) residual connection-based: Deep Residual UNet [32]. A comparative analysis of the chosen segmentation models has been provided in Table 2.2 and a concise overview of these models and the rationale behind their selection have been provided below.

Table 2.2: Comparative analysis of the chosen segmentation models (UNet, Vgg16 UNet, TransUNet & Deep Residual UNet).

Properties	UNet	Vgg16 UNet	TransUNet	Deep residual UNet
Model category	Original UNet	UNet variant: transfer learning-based	UNet variant: attention mechanism-based	UNet variant: residual connection based
Model size (MB)	**8.23**	53.05	**384.85**	31.41
Activation function	Sigmoid	Sigmoid	Sigmoid	Sigmoid
Number of Conv layers	23	23	**41**	**22**
Trainable parameters	**2158705**	13905953	**93427025**	8223809
Non-trainable parameters	**0**	1792	**7459136**	9216
Total parameters	**2158705**	13907745	**100886161**	8233025
Advantages	Leverages skip connection and encoder-decoder structure advantages; compact size; faster training; fewer parameters (total & trainable); ideal as a benchmark.	Leverages skip connection and encoder-decoder structure advantages; benefits from transfer learning for increased performance; shows nearly equal performance as the original UNet.	Leverages skip connection and encoder-decoder structure advantages; capable of capturing detailed spatial information and long-range dependencies using hybrid CNN-ViT structure.	Leverages residual connection, skip connection and encoder-decoder structure advantages.
Limitations	Lower or equal performance compared to other models; struggles to capture long-range dependencies and global context.	Model size is almost 6.5 times larger than UNet; larger model size makes the training process time-intensive.	Model size almost 46.5 times larger than UNet; requires more computational power and time; high parameter count (total and trainable) due to ViT integration; risk of overfitting on smaller datasets.	Model size almost 4 times larger than UNet.

UNet

UNet addresses the challenge of training deep neural networks for semantic segmentation with limited training data by leveraging its distinctive U-shaped encoder-decoder structure and skip connections. Their model structure effectively combines the low-level details captured by the encoder with high-level semantic information extracted by the decoder which enables precise segmentation outcomes even with smaller datasets [29], [31]. The usage of skip connections enhances the model by facilitating the integration of detailed context

into the higher-level semantic features, supporting efficient backward propagation and feature compensation during training [29]. However, UNet struggles to capture long-range dependencies and global context [55]. Figure 2.4 demonstrates the UNet architecture.

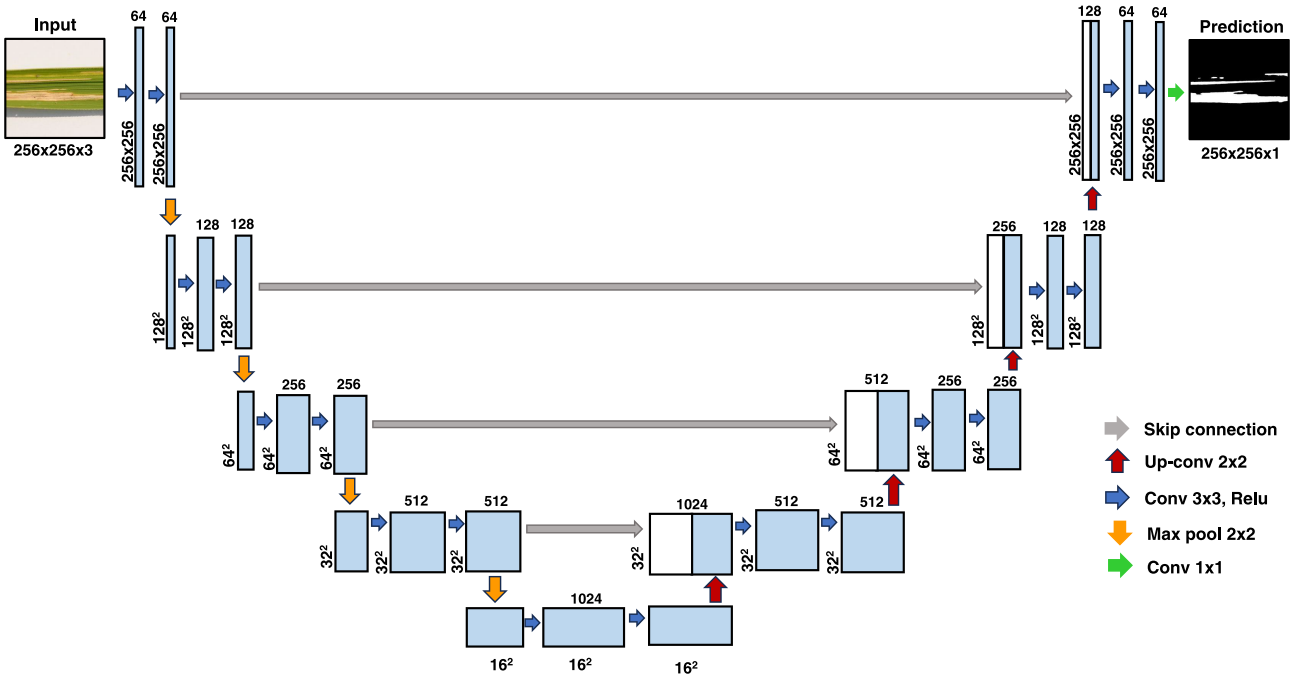


Figure 2.4: Visualisation of the UNet architecture.

Since UNet is the pioneering segmentation model capable of solving the problems associated with model training with small datasets, its applicability has been noticed across various domains [29]–[32], [36]–[38]. Because of its added structural benefits with skip connectors and encoder-decoder structure, and the high adaptability to different domains, it has inspired multiple authors to provide various enhancements to this model [29]–[32], [36]–[38]. Hence, UNet has been chosen as a benchmark for this study. The overview of the chosen UNet variants is given below.

1. **Transfer learning-based:** The UNet [31] model's encoder has been substituted with VGG16 [56] pre-trained on the ImageNet dataset for increased performance with three skip connections. Due to VGG16's large model size, training this model is time-intensive. Even though VGG16 suffers from vanishing gradient problems due to its simplistic architecture, we have chosen this model for our study to evaluate the impact of using a VGG16-based encoder in the modified UNet. The model architecture of VGG16-based UNet has been shown in Figure 2.5.

2. **Attention mechanism-based:** The attention mechanism enhances the model performance by focusing on significant features [30], [55]. Chen et al. [55] introduced a hybrid CNN-transformer with an encoder-decoder framework named TransUNet, optimising the strengths and mitigating the limitations of both CNNs and Transformers. CNNs face challenges in capturing long-range dependencies and struggle with the variations in texture, shape, and size of regions of interest (ROI) across samples. But it effectively captures detailed spatial information. On the other hand, transformers can effectively capture long-range dependencies and global context through the global self-attention mechanisms, but have limitations in obtaining localised detailed low-level features. In addition to this, the authors have also identified that, compared to using ViT directly as a feature extractor, using a hybrid of CNN and ViT—where the high-resolution spatial information is extracted by the CNN and global context is extracted by ViT—is proven to be a better solution. For being used as an input for the transformer, these extracted features are tokenized (vectorised) into 2D embeddings by linear projection. Then, the global context is encoded by

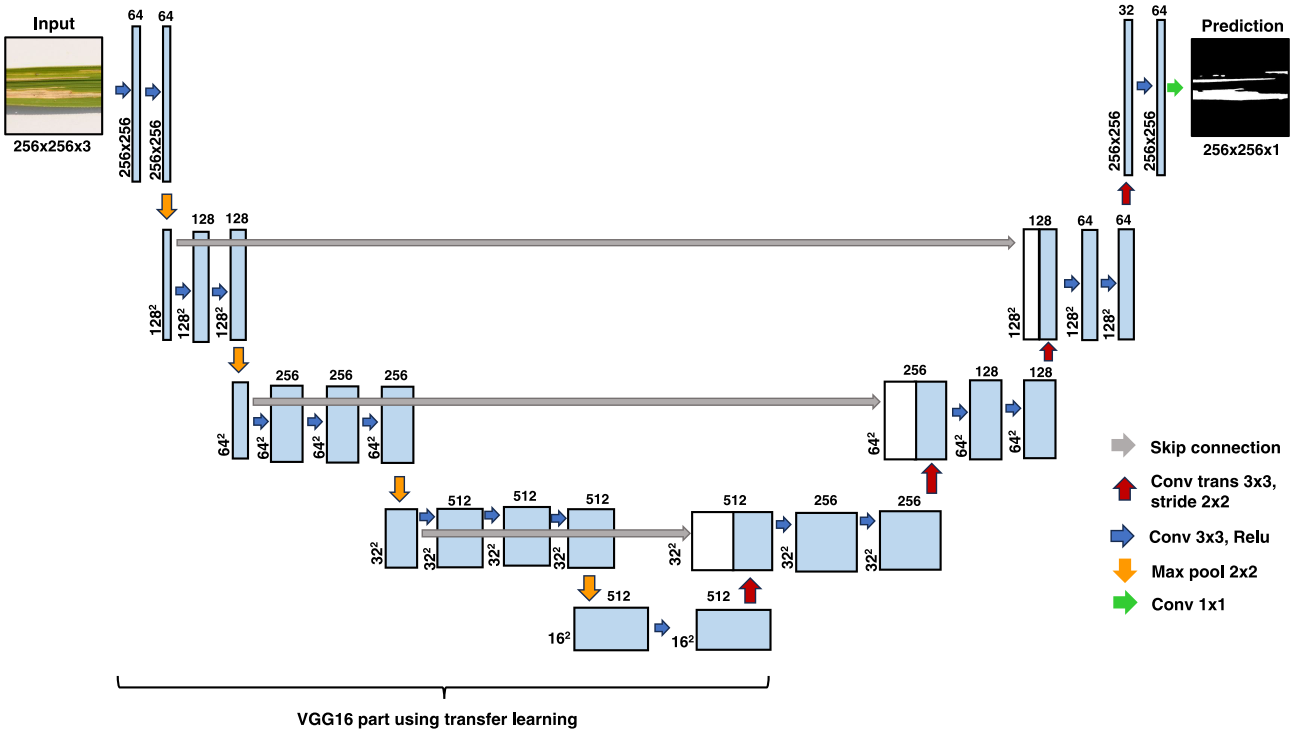


Figure 2.5: Visualisation of the VGG16-based UNet architecture.

the ViT. The feature maps are extracted in three different levels of CNN and are concatenated with the decoder path of the same level using skip connections. The decoder part consists of several upsampling steps for obtaining the final mask. The transformer part utilised for this study had 12 layers and 12 attention heads, and incorporated 3 skip connections for passing information to the decoder. Figure 2.6 shows the TransUNet architecture. Due to the incorporation of ViT in the model structure, the TransUNet requires a significant amount of computational power, time and data because of the inherited nature of ViTs [57]. To assess the impact of the unique hybrid structure of the hybrid CNN-transformer on the model performance, we have chosen this model.

3. **Residual connection based:** Residual connections solved the vanishing gradient problems associated with the training of deep networks [58]. Zhang et al. [29] introduced the Deep Residual UNet by augmenting the UNet with deep residual blocks instead of standard convolutional blocks. Their model leverages the residual connections to simplify the training of deeper networks and uses skip connections to preserve the detailed spatial information in the decoder. The innovative integration of both skip connections and residual connections in their model enhanced the information flow and minimised the total parameter count. Their model comprises an encoder and a decoder with an equal number of deep residual blocks, and a bridge with a single deep residual block without the residual connections. Figure 2.7 illustrates the Deep Residual UNet architecture. To assess the performance impact of integrating deep residual blocks into the UNet, this model has been chosen for this study.

2.2.2 Selected classification models

Various deep-learning architectures have been used in recent years for the paddy and various plant disease diagnosis, which cover basic CNN-based architectures [13]–[17], ViT-based architectures [8], [18]–[24], and memory efficient models [6], [25], [26]. For the classification task, we have chosen four categories of models, including (1) traditional: DenseNet121 [59], (2) vision transformer-based: ViT [57], (3) memory-efficient: MobileNet [60], and (4) ensemble-based: a custom ensemble model of Densenet121 and Xception. A comparative analysis of the chosen classification models has been provided in Table 2.3 and a concise overview of these models and the rationale behind their selection have been provided in this section.

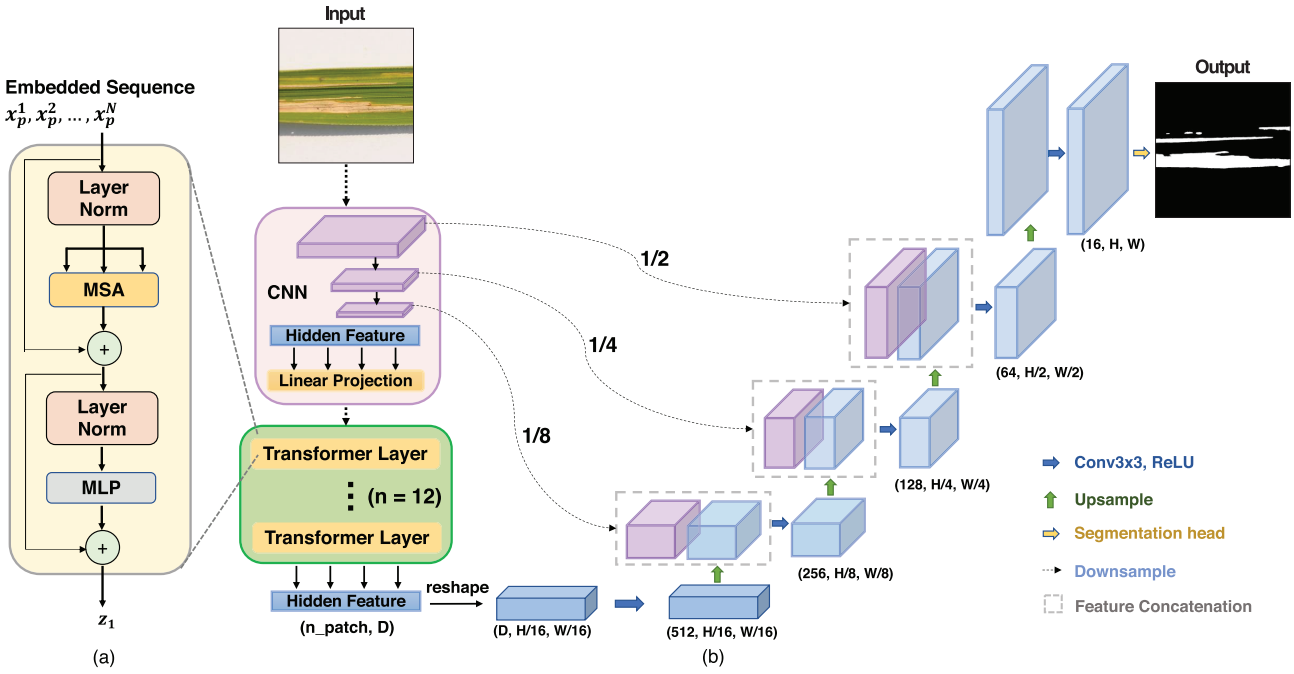


Figure 2.6: Visualisation of the TransUNet architecture.

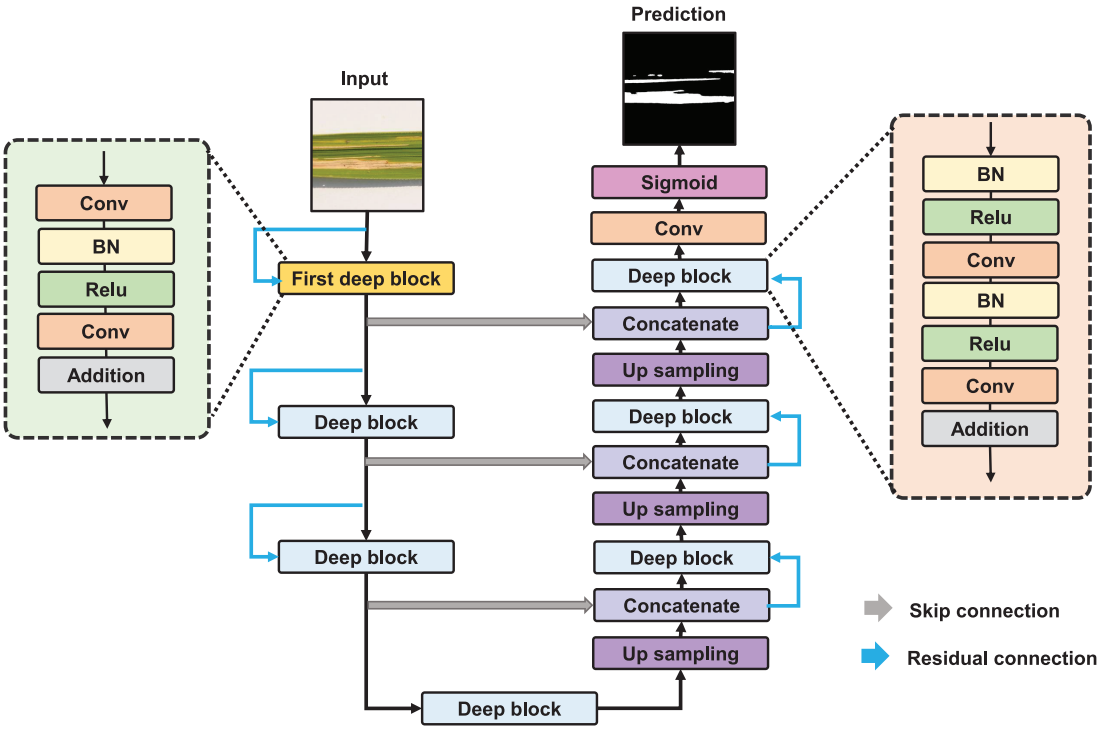


Figure 2.7: Visualisation of the Deep Residual UNet architecture.

Table 2.3: Comparative analysis of the chosen classification models (DenseNet121, MobileNet, Ensemble2 & ViT).

Properties	DenseNet121	MobileNet	Ensemble2	ViT
Model category	Traditional	Memory-efficient	Ensemble-based	ViT-based
Model size (MB)	26.9	**12.37**	**122.52**	73.73
Activation function	softmax	softmax	softmax	softmax
Trainable parameters	6967181	**3220301**	4213773	**19328514**
Non-trainable parameters	83648	21888	**27903080**	**0**
Total parameters	7050829	**3242189**	**32116853**	19328514
Advantages	Due to the usage of shorter connections between layers, has additional deep supervision via the loss function; due to the reuse of feature maps, fewer parameters and small model size; improved information and gradient flow in the network.	Suitable to be used in memory-constraint applications; usage of depthwise separable convolution reduces the computational requirements; high adaptability to a wide range of classification and object detection tasks.	Shows comparable performance to ViT in most cases and improved performance compared to the traditional models.	Due to the usage of transformers, can capture long-range dependencies.
Limitations	Increased GPU memory requirement due to concatenation of features; longer training time due to the usage of small convolutions rather than compact large convolutions.	Comparatively lower performance than the other models in this study.	Comparatively large model size than the traditional models and smaller model size than the ViTs; comparatively high parameter count than the traditional models and low parameter count than the ViTs.	Requires a high amount of data to achieve a good performance; due to the incorporation of the self-attention mechanism needs high computation; smaller patch size increases performance with the cost of increased computational requirements, model complexities and bias in texture.

(1) Traditional models

For the traditional model category, we have selected DenseNet [59] because it has effectively addressed the common issues associated with training deep CNN-based networks such as vanishing gradient and low feature propagation by concatenating features from all preceding layers. Due to the feature reuse, this model is comparatively compact, has a smaller size and fewer parameters. The collective feature memory achieved through the feature concatenation in DenseNet has enhanced the model training, improved gradient and information flow, and effectively addressed the vanishing gradient problem [59], [61]. It also has a regularising effect on the training data, which assists in mitigating the risk of overfitting [59]. Even though this model has several advantages, it demands more GPU memory due to feature concatenation and necessitates longer training time due to the usage of small convolutions rather than large conventions [61]. Figure 2.8 illustrates the process of DenseNet concatenating features from all previous layers. Due to its enhanced accuracy in classification tasks [59] and the aforementioned advantages, this model has been chosen for our study.

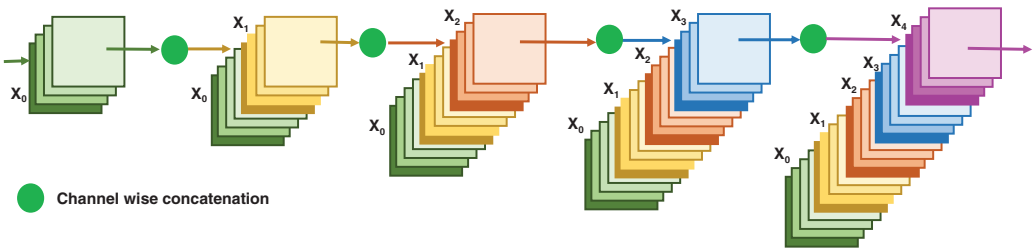


Figure 2.8: Illustration of DenseNet concatenating features from all previous layers.

(2) Vision transformer-based model

Inspired by the transformers used in natural language processing (NLP), Dosovitsky et al. [62] introduced ViT in computer vision which has proficiently mitigated the long-range dependency issues of CNNs. On

benchmark datasets (such as ImageNet, CIFAR-100, Oxford-IIIT Pets, etc.), ViT has produced outstanding classification performance with substantially lower memory consumption [19]. ViT divides the images into several fixed-size patches and utilises multi-head self-attention (MHSA) to recognise intricate patterns. Then, these patches are flattened into a sequence and processed by a transformer encoder to learn the inter-patch relationships, generating a feature vector representing the image. For the final label prediction, this vector is processed using a multi-layer perception (MLP) [8].

Several crops have been classified using ViT for various plant diseases [8], [18]–[24]. ViTs have also successfully minimised the long-range dependency issues of CNNs. The study by Bhojanapalli et al. [63] indicates that decreasing patch sizes enhances both performance and robustness in models. But it also increases model complexity and texture bias. Due to the nature of ViTs, they are prone to overfitting with small datasets and are usually computationally expensive. On the other hand, most paddy disease datasets are relatively small. Therefore, we selected this model to assess its behaviour with various levels of data augmentation in paddy disease classification and to compare its performance against traditional models. Using Keras Tuner, we have optimised the model to have 16 attention heads, 6 layers in depth, an MLP with 1024 units, and a dropout rate of 0.1. Since the input images were resized to 128x128 pixels, we have selected a patch size of 16. Figure 2.9 illustrates the ViT model architecture.

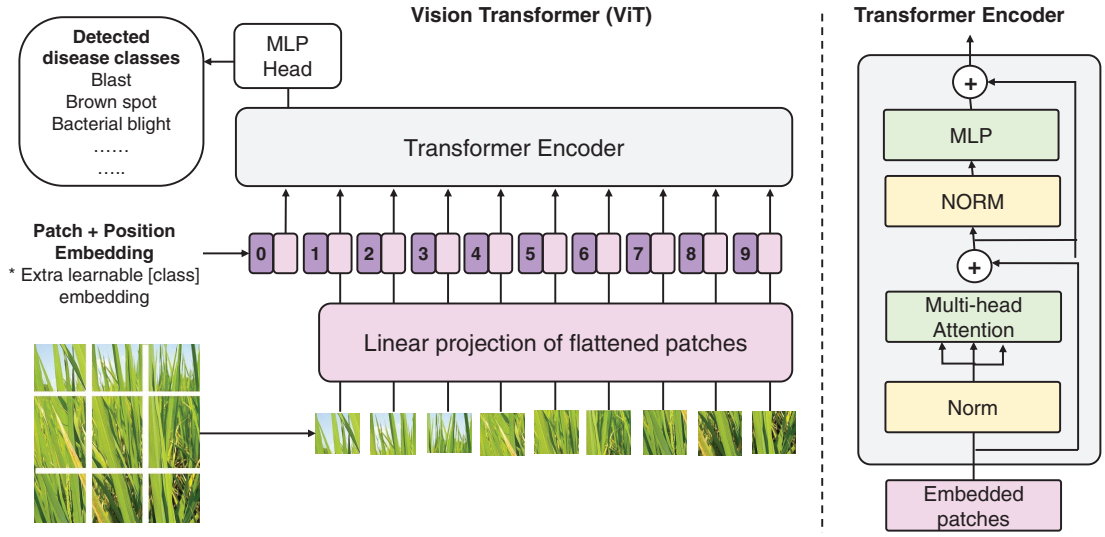


Figure 2.9: Illustration of the custom ViT architecture.

(3) Memory-efficient model

Howard et al. [60] have created MobileNet specifically to be used in memory-constraint applications such as mobile phones. Instead of standard convolutions, depthwise separable convolution was used in this model to make it memory efficient and reduce the computational requirements. Since this model is smaller in size, it is less susceptible to overfitting and requires minimal usage of regularisation and data augmentation compared to the other heavy models. This model also has high adaptability to a wide range of classification and object detection tasks. Since the primary users of paddy disease diagnosis applications are farmers with limited computational resources, a memory-efficient model is essential. Hence, we have selected this model for this study to evaluate its performance for paddy disease diagnosis.

(4) Ensemble-based model

We have created an ensemble model that uses the traditional models, DenseNet121 [59] and Xception [64] for feature extraction and additional layers to predict the final classification label. This model combines features from two models and trades increased model size for enhanced performance. On the other hand, it leverages DenseNet121's dense blocks and Xception's depthwise separable convolutions incorporating residual connections for improved performance. The architecture of this model has been illustrated in Figure 2.10. This model was

developed for this study to achieve performance comparable to ViTs while maintaining a lower parameter count and smaller model size. Using the Keras Tuner, we have identified the best dropout value to be 0.1 and the number of dense layer units to be 1024 to optimise the model performance.

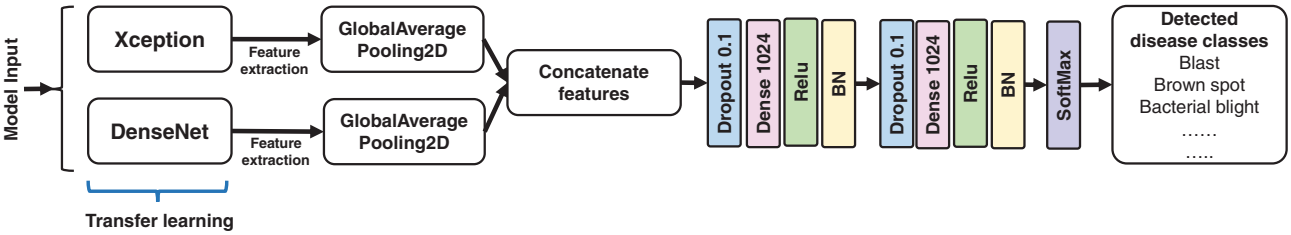


Figure 2.10: Illustration of the architecture of a custom classification model named Ensemble2.

2.3 Identified research gaps

For training robust models using deep learning (DL), we need large datasets covering diverse classes for practical real-life implementations. Currently except for the imbalanced "Paddy Doctor" [16] dataset, we do not have any large datasets [65]–[72]. Additionally, high-sensory (multi-spectral, hyperspectral, fluorescence and thermal) and UAV image-based datasets are also lacking [73]. In addition to this, there are currently no open-access paddy disease segmentation datasets which creates a significant challenge for new researchers. Lastly, estimating and formulating an effective disease treatment strategy requires an accurate assessment of the disease severity. Except for studies by Lamba et al. [74] and Pal et al. [75], there is no work on paddy disease severity analysis.

3 Research Methodology

This chapter provides a detailed overview of the steps incorporated into our research methodology. Since our research belonged to two different tasks related to paddy diseases: (1) segmentation and (2) classification, the research methodology also has two sections covering the individual tasks. Detailed research methodologies for identifying paddy diseases through two computer vision-based tasks (segmentation and classification) have been given below.

3.1 Research methodology for segmentation

The methodology includes the creation of a new paddy disease segmentation dataset, the pre-processing steps incorporating different augmentation strategies, post-processing, implementation setup with the model training specifications, and chosen model performance evaluation metrics. Outlined below are the steps incorporated into our research methodology. Figure 3.1 provides a comprehensive visual representation of the entire process.

3.1.1 Paddy disease segmentation dataset creation process

This section contains the description of the dataset used and the step-by-step process of creating masks to identify paddy diseases. This novel paddy disease segmentation dataset has been further used for the performance analysis of four segmentation models.

Dataset

Due to the time constraint and the complexity associated with the field images, the Rice Leaf Diseases Dataset [66] from Kaggle has been chosen to develop a new paddy segmentation dataset. It contains 120 images of diseased paddy leaves against simple white or black backgrounds. The dataset covered Leaf smut, Brown spots, and Bacterial leaf blight with 40 samples for each disease. Out of total 120 samples, 2 outliers were taken out belonging to Bacterial leaf blight disease. These 2 samples were completely different from the rest of the 38 samples belonging to this class/category and had comparatively low image quality. The samples of this dataset have been illustrated in Figure 3.2.

Pre-processing

The dimensions of the samples in this dataset [66] varies with the samples. Hence, the images were resized to 256x256 for uniformity in dimension. An 80-20 train-test split was implemented and the training set was further expanded through augmentation techniques from 96 to 616. These techniques include colour-space augmentations (green colour shift, random changes in brightness, contrast, saturation, and hue) and geometric augmentations (horizontal flip, vertical flip, rotation, width shift, height shift, and elastic transformation). The intensity of these augmentations has been chosen in such a way that they do not drastically diverge from the original samples and overall help in enhancing the dataset which has been proven by the results in Section 4.1. An example of image augmentation techniques used for the original image is illustrated in Figure 3.3. The 23 test data have been used for mask creation without any altercations.

Figure 3.1: Illustration of the research methodology for the segmentation task.

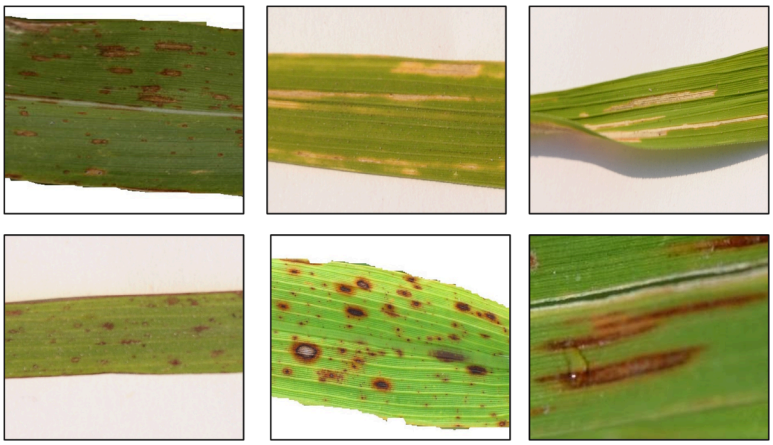


Figure 3.2: Samples of Rice Leaf Diseases Dataset [66] from Kaggle.

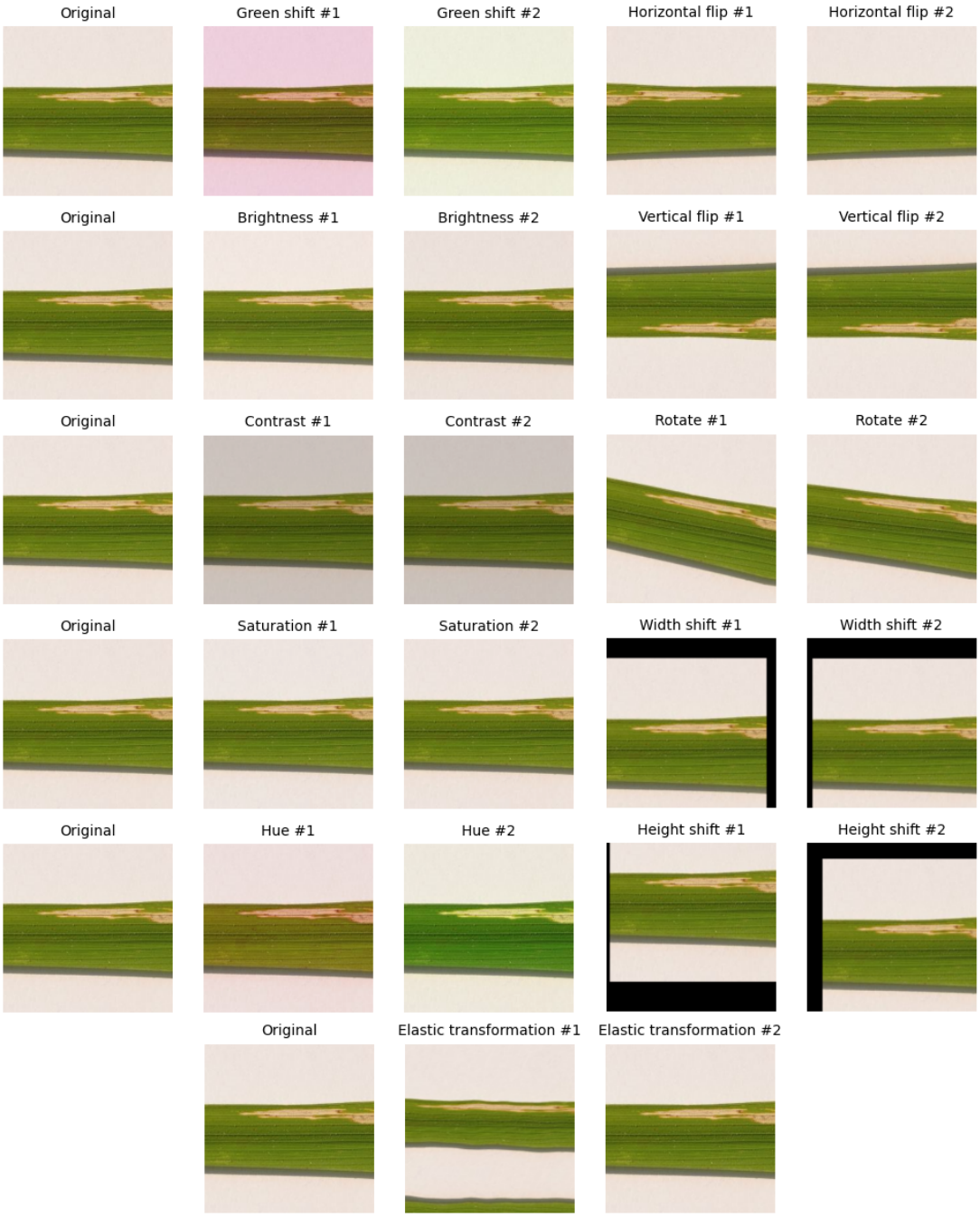


Figure 3.3: Visualisation of the utilised augmentation techniques for mask creation (green shift, brightness, contrast, saturation, hue, vertical flip, horizontal flip, rotate, width shift, height shift, and elastic transformation).

Mask creation process

For creating masks to identify the diseased areas from the RGB images, the following steps have been followed and the whole process has been illustrated in Figure 3.4. The mask-creation process of a sample is shown in Figure 3.5. The dataset is available in this link. Due to the complexity associated with the mask creation using image processing techniques and to ensure the quality of the masks, all the generated masks have been manually rechecked with side-by-side comparisons to rectify minor inaccuracies.

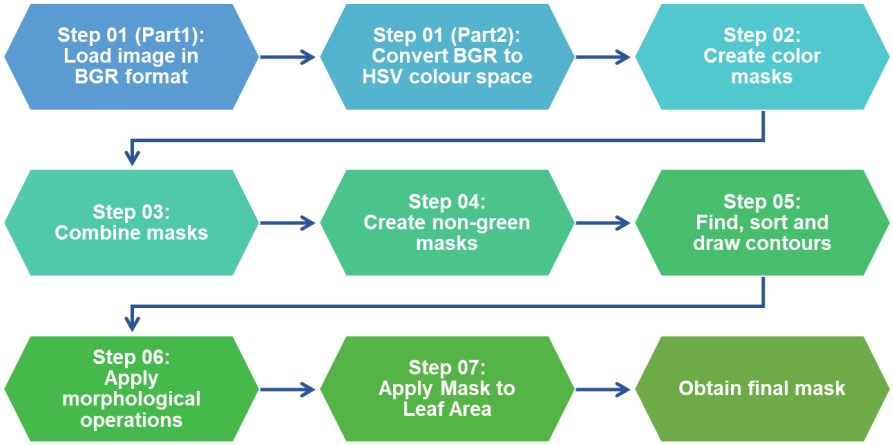


Figure 3.4: Layout of paddy disease mask creation process using image processing techniques.

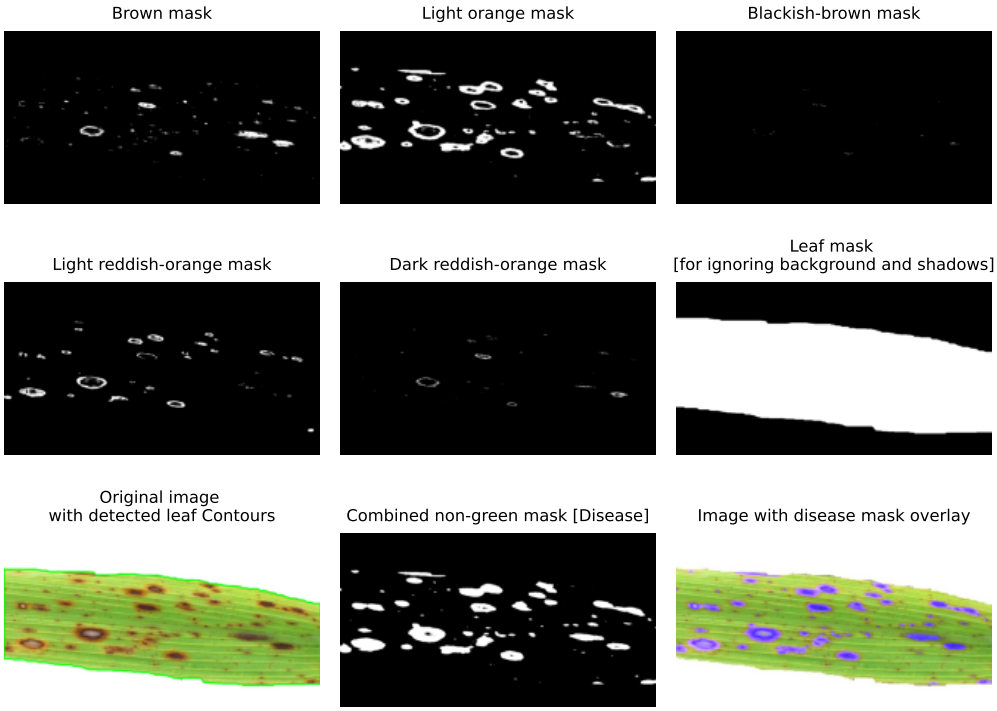


Figure 3.5: Sample of mask-creation process.

- **Step 1: Load the image and convert it to HSV colour space:** Using CV2, the input images are loaded in BGR (Blue, Green, Red) format. Then, they are converted to the HSV (Hue, Saturation, Value) format for colour segmentation.

- **Step 2: Create colour masks to capture leaves:** Different masks are created to capture various colours present in the leaf image: green, brown, orange, blackish brown, light reddish-orange, and dark reddish-orange. Each mask targets a specific colour range in the HSV space to capture different aspects of the leaf and possible diseases.
- **Step 3: Combine masks:** All the colour masks are combined using a bitwise OR operation to form a single mask that includes all the coloured areas of the leaves including the disease portions. This ensures that the entire leaf is included, regardless of its colour due to diseases or image-capturing variations.
- **Step 4: Apply morphological operations:** Due to disease or image-capturing conditions, a few portions of the leaf might be completely white or other colours which might not get captured in the combined masks and might look like small holes/noises. Morphological operations such as closing and opening are applied to the combined mask to fill in these holes within the leaf area.
- **Step 5: Find, sort, and draw contours:** Contours are detected in the mask created from Step 4, which are then sorted based on their area size. This step is crucial for identifying the main leaf or leaves in the image by focusing on the largest contiguous areas. It also helps in discarding background noises and shadows present in the image. The largest two contours are drawn onto the original image which helps in visualisation and further mask creation process.
- **Step 6: Create non-green masks:** Individual masks for colours of brown, orange, blackish brown, light reddish-orange, and dark reddish-orange are created and combined to capture all non-green colours which give the indication of diseases or stress on paddy leaves.
- **Step 7: Apply mask to leaf area:** The non-green mask is applied to the area identified as the leaf in Step 5 with contours. This step ensures that only non-green areas on the actual leaf are highlighted, excluding any background or noise.

3.1.2 Data pre-processing and post-processing

The original 616 train images and their corresponding masks have been augmented to make the train set have 1500 and 2500 samples which results in 884 and 1884 additional images respectively. Horizontal flip, vertical flip, rotation, width shift, height shift, elastic transformation, gaussian blur, zoom-in, and zoom-out have been used for this augmentation. Figure 3.6 illustrates the utilised augmentation techniques. Figure 3.7 illustrates the distribution of disease pixel percentages in the segmentation masks across training datasets of varying sizes for a detailed pixel-level analysis. All three histograms are skewed towards lower percentages which indicates that most masks in each training dataset contain a small area of disease pixels/ROI. As the training dataset size increased from 616 to 2500 samples, the histograms showed smoother distributions, a decrease in the frequency of low disease pixel occurrences, and greater variability in disease pixel percentages. These data changes with augmentation might improve model generalisation.

The training dataset has been further split into a ratio of 80:20 for training and validation. The original images and masks were in size 256x256. We normalised the images by diving with 256 to have the values within the range of 0 to 1. A threshold value of 0.5 has been used to obtain the final mask.

3.1.3 Implementation setup/model training parameters

Using image processing techniques and OpenCV, a novel paddy disease segmentation dataset has been created as described in Section 3.1.1. This dataset has been used for the performance analysis of four segmentation models and all of these models have been implemented on Keras with the TensorFlow framework. The Google Colaboratory with 51 GB RAM has been used for training the models and generating the results. The processing units were NVIDIA A100 GPU, V100 GPU, T4 GPU, and Google TPU. Additionally, an Intel Core i9 laptop with NVIDIA Geforce RTX 4060 Portable GPU, 32 GB RAM, and 2.2 GHz CPU base clock frequency has been used for training the models. The learning rate scheduler, ReduceLROnPlateau [76] has been used to

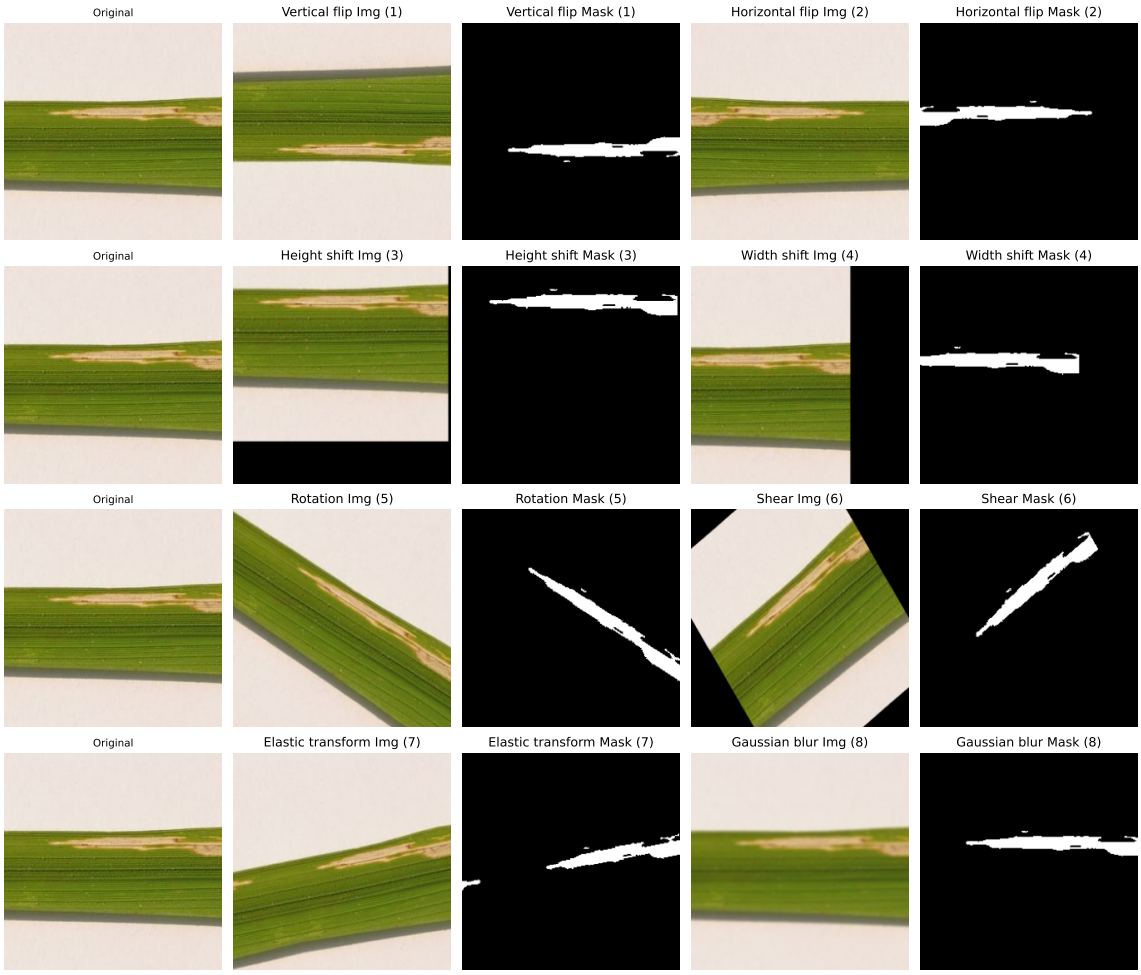
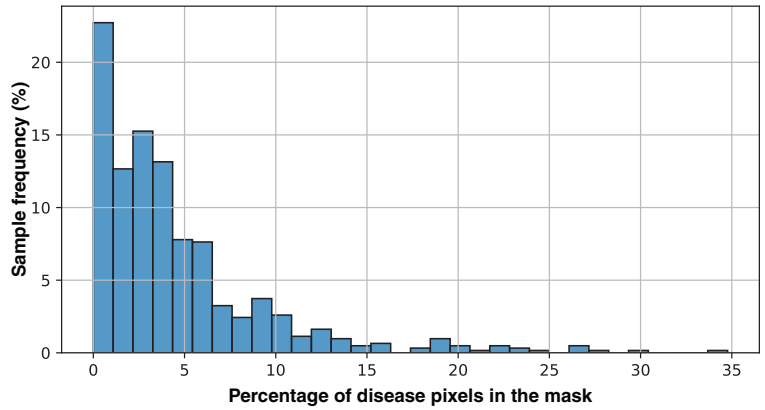
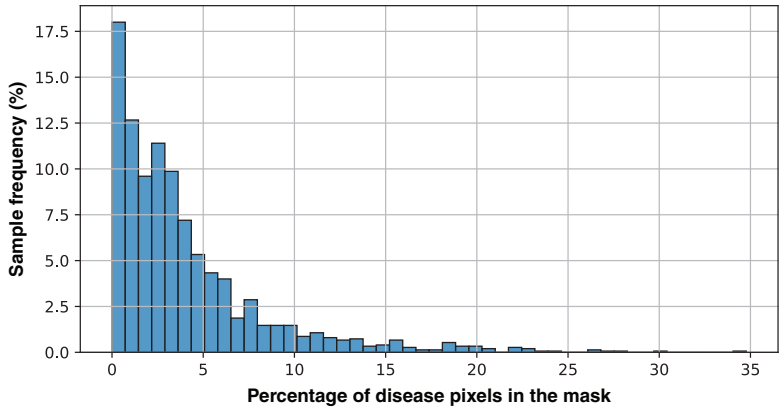


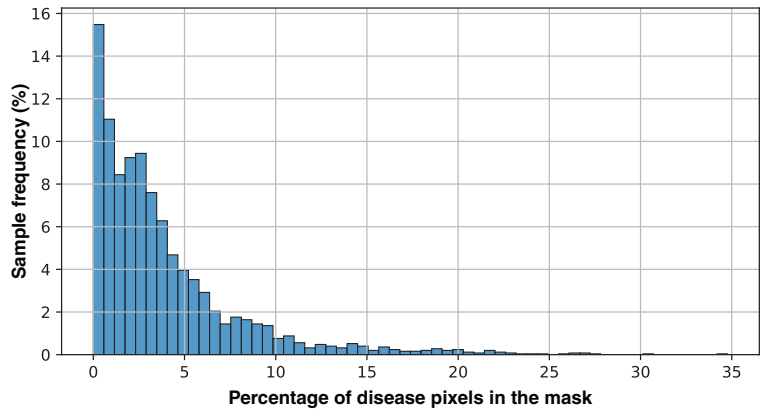
Figure 3.6: Visualisation of the augmentation techniques used for the original image and the mask image.



(a) Disease pixel distribution for 616 train data



(b) Disease pixel distribution for 1500 train data



(c) Disease pixel distribution for 2500 train data

Figure 3.7: Illustration of the disease pixel distributions across three training data sizes.

Table 3.1: Model training specifications for the segmentation task.

Training parameters	Parameter values
Optimizer	Adam
Loss function	Binary cross-entropy
Learning rate scheduler: Initial learning rate	0.001
Learning rate scheduler: Minimum learning rate	1e-8 = 0.00000001
Learning rate scheduler: learning rate reduction factor	0.5
Learning rate scheduler: learning rate reduction patience	3
Batch size	10, except TransUNet with 616 data. For a small number of samples, the model was extremely unstable. For this reason, after trial and error, we chose a batch size of 15, which provides the most stability for the model
Maximum number of epochs	70 ~80 (chose the model with the best validation IoU)

improve convergence, overcome plateaus, and avoid overfitting of the models. Additionally, different batch sizes have been trialled to identify the batch size that makes most of the model stable. The training parameters used for all the segmentation models are given in Table 3.1.

3.1.4 Performance evaluation metrics for segmentation

Accuracy (ACC), Precision (Pr), Recall (Rec), F1-Score, Intersection over union (IoU) and confusion matrix (CM) have been used for the performance analysis of the models. Their equations have been presented in Equations 3.1, 3.2, 3.3, 3.4, and 3.5 respectively. We have also utilised confusion matrices for a better understanding of the misclassifications and the confusion matrix used for the segmentation task has been illustrated in Figure 3.8. A concise description of the performance evaluation metrics to be used for the segmentation task has been given below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.4}$$

$$IoU = \frac{1}{n} \sum \frac{\text{Area of overlap}}{\text{Area of union}} \tag{3.5}$$

Where, TP represents True Positives; pixels correctly classified in both segmented and original mask images, TN represents True Negatives; pixels correctly identified as not part of the segmented image in both images, FP represents False Positives; pixels incorrectly marked in the segmented image, not in the original mask, FN represents False Negatives; pixels missed in the segmented image but present in the original mask.

- **Accuracy:** The accuracy measures the proportion of correct predictions made by the model out of the total predictions [4], [19], [21]–[24], [27], [40], [74], [77], [78]. For segmentation, it gives pixel-wise accuracy. In most cases, the background comprises the majority of pixels compared to the ROI which often increases the accuracy value. Hence, pixel accuracy provides limited insight into the effectiveness of disease segmentation and will not be considered for performance comparison. From Table 4.1, Table

4.2, and Table 4.3, we can see that with different train data sizes and for all the models, the accuracy is almost equal which further validates our reason for putting low significance in model accuracy.

- **Precision/Positive predicate value (PPV):** The precision measures the model's accuracy in predicting positive instances. It is calculated as the ratio of true positives to the total predicted positives [19], [22], [24], [40].
- **Recall/Sensitivity (Sen)/True positive rate (TPR):** The recall is calculated as the ratio of true positives to the sum of true positives and false negatives [4], [12], [21], [24], [40], [74], [78]. It measures the model's ability to correctly identify all positive instances.
- **F1-Score (F1):** The F1-score is a weighted harmonic mean of precision and recall which combines the insights of both precision and recall to assess prediction performance. It is particularly valued for its ability to balance precision and recall trade-offs [4], [12], [19], [21], [22], [24], [40], [74], [77], [79]. For segmentation tasks, It is also called the Dice similarity coefficient (DSC).
- **IoU index:** IoU evaluates the segmentation performance by measuring the similarity between the ground-truth labelled mask and the predicted mask on a pixel level [21]. It is also called the Jaccard Index (JAC). In the case of segmentation tasks, ROI typically consists of a small portion of the whole image. Due to the accurate detection of backgrounds present in the sample, performance evaluation metrics such as accuracy, recall, and precision provide an inaccurate perception of superiority which is mostly wrong in reality. Therefore, researchers consider IoU to be the most informative metric for understanding the performance of segmentation models, and it is commonly used [32].
- **Confusion matrix (CM) for segmentation:** The confusion matrix is a tabular or graphical performance measurement tool which is an N x N array/matrix that evaluates model performance by comparing predicted outputs (pixel classes) against actual outputs (ground truth mask) [12], [74], [77]. In Figure 3.8, the diagonal values represent the correctly classified pixels (true positives), and off-diagonal values represent the misclassification. The rows represent the actual pixel values (ground truth) and the columns represent the predicted pixel category. It provides insight into the analysis of model performance, model credibility and error types [24], [74], [77].

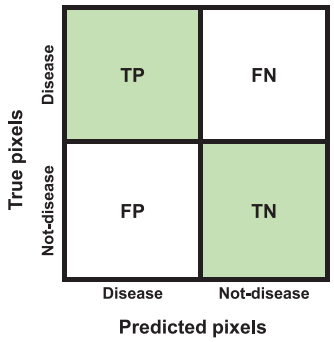


Figure 3.8: Illustration of the confusion matrix used for disease segmentation.

3.2 Research methodology for classification

The classification methodology includes a description of the used datasets, including their class distributions, the pre-processing step incorporating various augmentation strategies, and the implementation setup details, including model training specifications and performance evaluation metrics. Figure 3.9 provides a comprehensive visual representation of the entire process.

Figure 3.9: Illustration of the research methodology for the classification task.

3.2.1 Datasets

For the classification tasks, we have used two datasets named Paddy Doctor [16] and Rice Leaf Disease [65]. The Paddy Doctor [16] dataset was collected from the Tirunelveli district of Tamil Nadu, India. It has 16225 samples and covered the diseases Bacterial leaf blight (BLB), Bacterial leaf streak (BLS), Bacterial panicle blight (BPB), Black stem borer (BSB), Blast, Brown spot, Downy mildew, Hispa, Leaf roller, Tungro, White stem borer(WSB) and Yellow stem borer (YSB) having samples 648, 505, 450, 506, 2351, 1257, 868, 2151, 1095, 1951, 1273, and 765 respectively and 2405 normal samples. The class distribution of Paddy Doctor has been presented in Figure 3.10. On the other hand, the Rice Leaf Disease was collected from the Western tract of Odisha, India. It has 5932 samples and covers the diseases Bacterial blight, Blast, Brown spot, and Tungro having samples 1584, 1440, 1600, and 1308 respectively. The class distribution of this dataset has been presented in Figure 3.11. Both datasets were collected in field environments.

3.2.2 Data pre-processing

Since dataset [16] is comparatively larger than dataset [65], we split the training data into ratios of 70:30 and 80:20 respectively for training and validation. For augmentation, we employed two levels of intensity: one

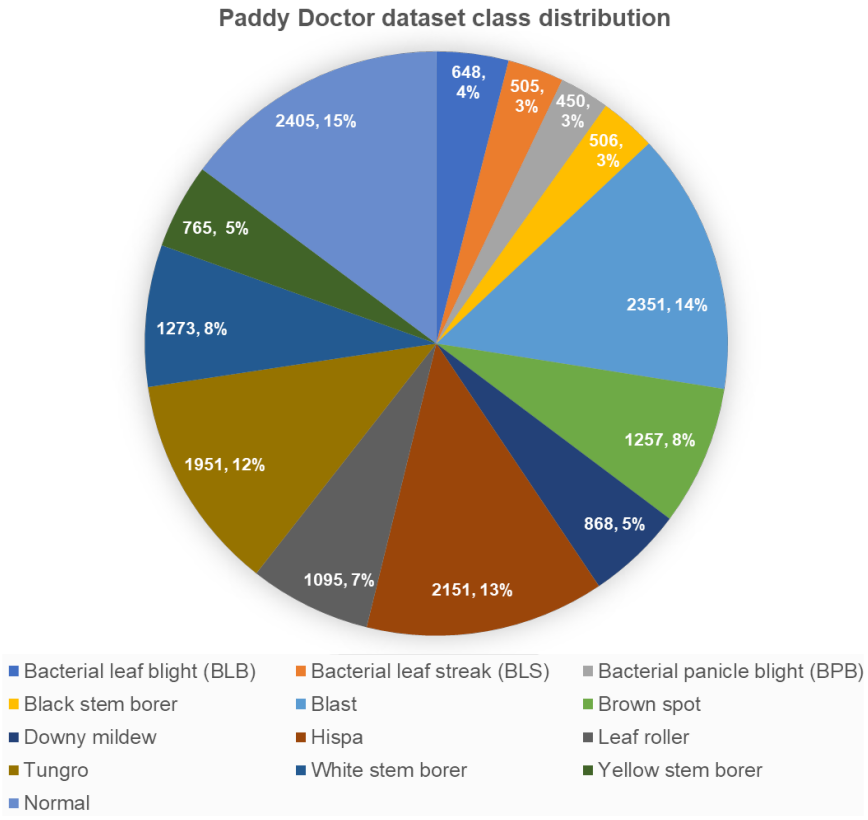


Figure 3.10: Class distribution of the Paddy Doctor dataset.

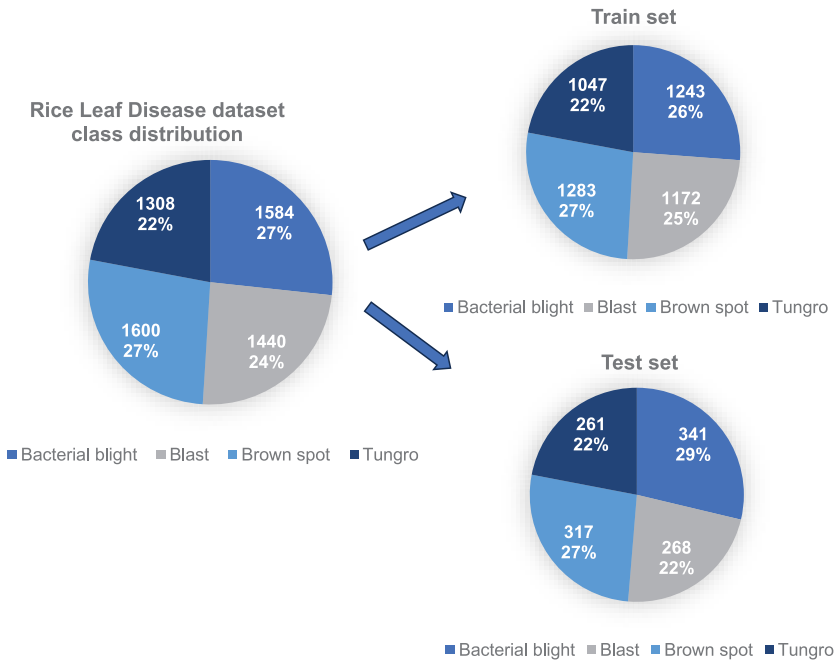


Figure 3.11: Class distribution of the Rice Leaf Disease dataset.

involving only basic augmentations, and another incorporating all relevant augmentations. For basic augmentations, we applied zoom and horizontal flip. For extensive augmentations, we included rotation, width and height shifts, zoom, horizontal and vertical flips, brightness shifts, and elastic deformations. Figure 3.12 demonstrates the utilised augmentation techniques. The original images from datasets [16] and [65] were in sizes 256x256 and 300x300 respectively. We resized these original images to 128x128 and normalised the images by dividing with 128 to have values within the range of 0 to 1.

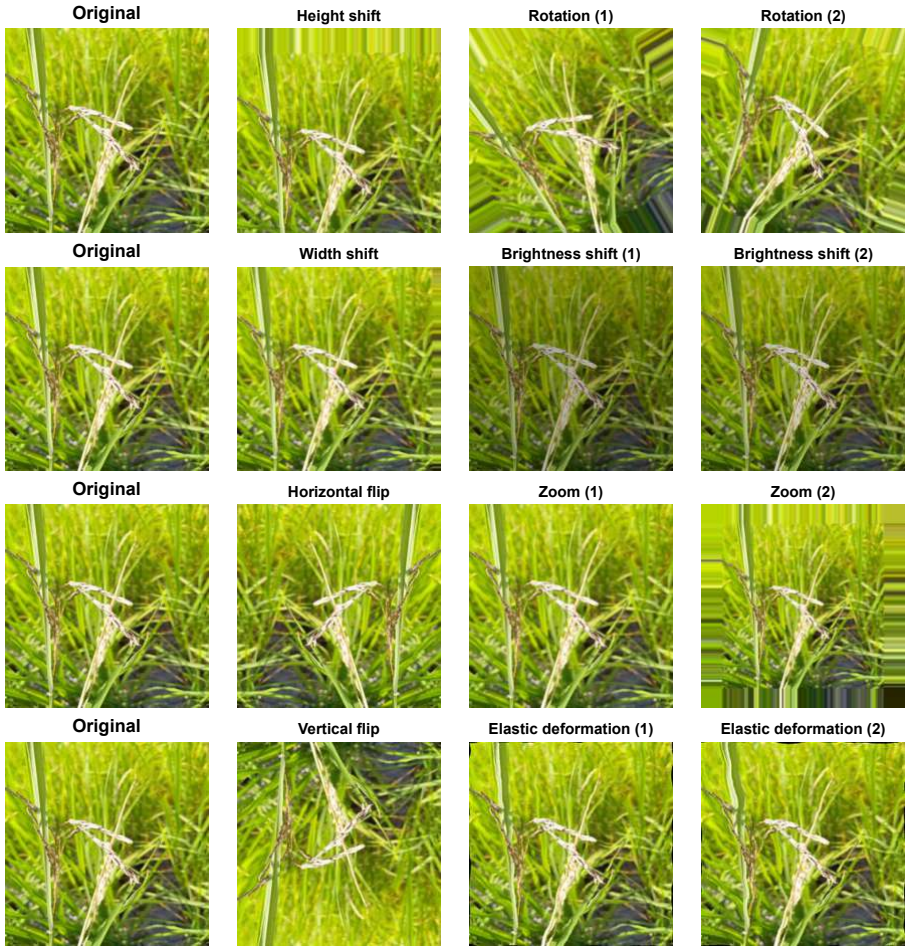


Figure 3.12: Visualisation of the augmentation techniques utilised for classification task (height shift, channel shift, shear, zoom, elastic deformation, horizontal flip, vertical flip, rotation, width shift, and brightness shift).

3.2.3 Implementation setup/model training parameters

Using the previously mentioned two paddy disease datasets, the performance analysis of the four classification models has been conducted and all of these models have been implemented on Keras with the TensorFlow framework. An Intel Core i9 laptop with NVIDIA Geforce RTX 4060 portable GPU, 32 GB internal memory, and 2.2 GHz clock frequency has been used for training the models. The learning rate scheduler, ReduceLROnPlateau [76] has been used to improve convergence and overcome plateaus. To avoid the overfitting of the models, early stopping [80] has been used. Additionally, different batch sizes have been trialled to identify the batch size that makes the most of the models stable. The training parameters used for all the classification models are given in Table 3.2.

3.2.4 Performance evaluation metrics for classification

Accuracy, F1-Score, precision, and recall have been used for the performance analysis of the selected classification models and their equations have been presented in Equations 3.6, 3.7, 3.8, and 3.9 respectively. We have also utilised the confusion matrix for a better understanding of the misclassifications and the confusion matrix used for the classification task has been illustrated in Figure 3.13. A concise description of the performance evaluation metrics to be used for the classification task has been given below.

Table 3.2: Model training specifications for the classification task.

Training parameters	Parameter values
Optimizer	Adam
Loss function	Categorical cross-entropy
Learning rate scheduler: initial learning rate	0.001 or 0.0001
Learning rate scheduler: minimum learning rate	1e-9 = 0.000000001
Learning rate scheduler: learning rate reduction factor	0.5
Learning rate scheduler: cooldown epochs	2
Learning rate scheduler: learning rate reduction patience	3
Early stopping: patience	12 for the dataset [65] and 15 for the dataset [16]
Batch size	64 for the dataset [65] and 100 for the dataset [16]
Maximum number of epochs	200 (chose the model with the best validation F1-score)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.6}$$

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.7}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.8}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.9}$$

Where, TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives.

	Predicted label	Predicted label
True label	TP	FN
True label	FP	TN

Figure 3.13: Illustration of the confusion matrix to be used for disease classification.

- **Accuracy (Acc):** Accuracy measures the proportion of correct predictions made by the model out of the total predictions [4], [19], [21]–[24], [27], [40], [74], [77], [78].
- **Precision (Pr):** Precision measures the model's accuracy in predicting positive instances. It is calculated as the ratio of true positives to the total predicted positives [19], [22], [24], [40].
- **Recall (Rec):** Recall is calculated as the ratio of true positives to the sum of true positives and false negatives [4], [12], [21], [24], [40], [74], [78]. It measures the model's ability to correctly identify all positive instances.
- **F1-Score (F1):** The F1-score is a weighted harmonic mean of precision and recall which combines the insights of both precision and recall to assess prediction performance. It is particularly valued for its ability to balance the trade-off between precision and recall [4], [12], [19], [21], [22], [24], [40], [74], [77], [79].
- **Confusion matrix (CM) for classification:** The confusion matrix is a tabular or graphical performance measurement tool for classification models [12], [74]. It is an N x N arrays that evaluate classification model performance by comparing predicted outputs against actual outputs [12], [77]. It provides insight into the analysis of model performance, model credibility and error types [24], [74], [77].

4 Results

This chapter provides a brief analysis and illustration of the quantitative and qualitative results that we have obtained through the research methodology presented in chapter 3. Since our research belonged to two different tasks related to paddy disease identification: (1) segmentation and (2) classification, the result chapter also has two different sections covering the individual tasks. Results obtained for identifying paddy diseases through two computer vision-based tasks (segmentation and classification) have been given below.

4.1 Segmentation results

For the segmentation task of this study, we have used the paddy disease segmentation dataset represented in Section 3.1.1. Subsequently, we selected four recent segmentation models (Unet, VGG16 UNet, TransUNet and Deep Residual UNet) from the field of machine learning to analyse their applicability in paddy disease. As described in the research methodology, the original training set, which had 616 training samples, was enhanced through traditional data augmentations, increasing the total number of training images to 1500 and 2500. Finally, to assess the final performance, we used 23 unseen test images without any augmentation.

4.1.1 Test performance comparison of individual models across three train data sizes

Figure 4.1 shows that all the selected segmentation models have demonstrated equal or improved test performance with increased training data across metrics such as IoU, F1-score, precision, recall, and accuracy. Exceptions were observed in UNet's recall and TransUNet's precision.

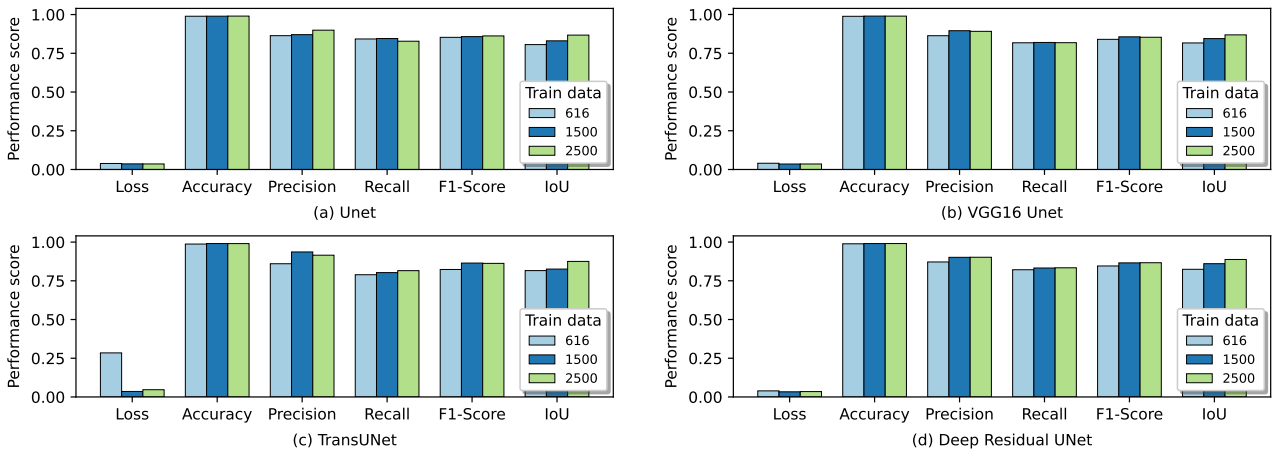


Figure 4.1: Visualisation of selected segmentation models' test performance comparison across three train data sizes.

4.1.2 Validation and test performance across three train data sizes

All models were trained to maximise the validation IoU with the same model training parameters specified in Table 3.1. To improve the convergence and overcome plateaus, the LR scheduler was utilised. For testing, instead of using models from the final epoch, model weights with the highest validation IoU were selected. This approach prevented overfitting and is supported by the nearly comparable performance between validation and test results. The performance of chosen four segmentation models with 616, 1500 and 2500 train data sizes is detailed in Table 4.1, Table 4.2, and Table 4.3 respectively and a visual representation of the same data with a bar-chart format has been presented in Figure 4.2, Figure 4.3, and Figure 4.4 respectively for better comprehension.

Validation results: Below are the validation results for the selected segmentation models using 616, 1500, and 2500 train data.

- **Validation performance with 616 train data:** From the validation losses from Figure 4.2 and Table 4.1, it has been observed that TransUNet had the highest validation loss with nearly 30%, while Deep Residual UNet had the lowest validation loss with nearly 6%. For accuracy, all the models have shown almost equal performance where Deep Residual UNet had the highest score of 98.24%. UNet has been observed to have the lowest score in validation accuracy, F1-score and IoU with 98.16%, 73.37%, and 81.41% respectively, while Deep Residual UNet has the highest with 98.24%, 75.23%, and 82.17% respectively. TransUNet took the longest to get the best validation IoU (50 epochs), while VGG16 UNet took the shortest with 15 epochs.

Table 4.1: Comparison of the selected segmentation models' performance for 616 train data.

Model name	Test loss	Test accuracy	Test precision	Test recall	Test F1-score	Test IoU	Val. loss	Best val. accuracy	Best Val. F1-score	Best val. IoU	Best val. IoU at epoch
UNet	**0.03887**	**0.98917**	0.86372	**0.84246**	**0.85296**	**0.80614**	0.0607	**0.9816**	**0.7337**	**0.8141**	17
VGG16 UNet	0.04029	0.98836	0.86315	0.81750	0.83970	0.81648	0.0622	0.9819	0.7429	0.8171	**15**
Trans UNet	**0.28453**	**0.98736**	**0.86023**	**0.78928**	**0.82323**	0.81589	**0.2969**	0.9818	0.7484	0.8158	**50**
Deep Residual UNet	0.03931	0.98882	**0.87158**	0.82134	0.84571	**0.82419**	**0.0582**	**0.9824**	**0.7523**	**0.8217**	41

- **Validation performance with 1500 train data:** From Figure 4.3 and Table 4.2, we can see that while Deep Residual UNet had the highest validation scores in accuracy, F1-score and IoU with 98.99, 80.69, and 86.06% respectively, TransUNet had the lowest with 98.65%, 74.26%, and 82.26% respectively. On the other hand, Deep Residual UNet had the lowest validation loss with under 3%, while TransUNet had around 4.8%. VGG16 UNet trained for the least number of epochs (43), while Deep Residual UNet and TransUNet required the most (73 epochs).

- **Validation performance with 2500 train data:** From Figure 4.4 and Table 4.3, it is observed that TransUNet had the highest validation loss with approximately 3.3%, Deep Residual UNet had the lowest with 2.25%. UNet and VGG16 UNet have equal and the lowest validation accuracy with 98.98% among the other models. Furthermore, VGG16 UNet achieved the lowest validation F1-score and IoU with 81.12% and 86.88% respectively, while Deep Residual UNet led with 84.58% and 88.82% respectively. To reach the highest validation IoU, Deep Residual UNet took the longest (70 epochs), while UNet needed the least number of epochs (60).

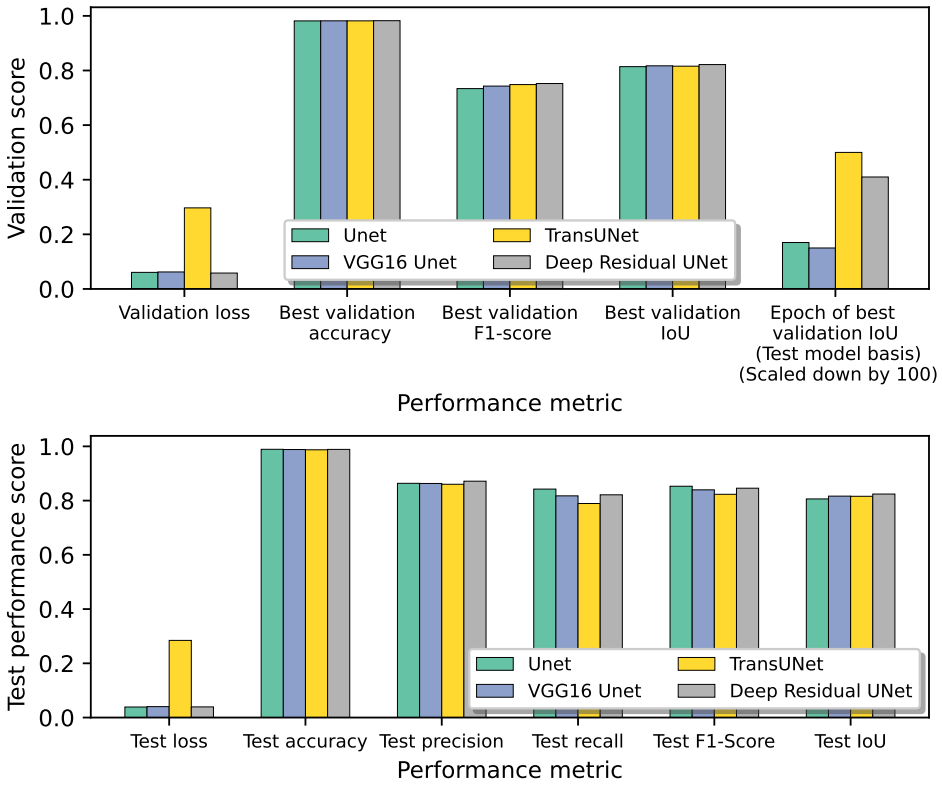


Figure 4.2: For 616 train data, visual representation of the performance comparison of the chosen segmentation models on validation dataset and test dataset.

Table 4.2: Comparison of the selected segmentation models' performance for 1500 train data.

Model name	Test loss	Test accuracy	Test precision	Test recall	Test F1-score	Test IoU	Val. loss	Best val. accuracy	Best val. F1-score	Best val. IoU	Best val. IoU at epoch
UNet	**0.03626**	**0.98953**	**0.87036**	**0.84521**	0.85760	0.83045	0.0402	0.9877	0.7576	0.8320	57
VGG16 UNet	0.03533	0.98970	0.89537	0.81965	**0.85584**	0.84447	0.0333	0.9884	0.7727	0.8432	**43**
Trans UNet	0.03556	**0.99062**	**0.93636**	**0.80298**	0.86456	**0.82580**	**0.0477**	**0.9865**	**0.7426**	**0.8226**	**73**
Deep Residual UNet	**0.03318**	0.99036	0.90152	0.83253	**0.86565**	**0.86042**	**0.0291**	**0.9899**	**0.8069**	**0.8606**	**73**

Table 4.3: Comparison of the selected segmentation models' performance for 2500 train data.

Model name	Test loss	Test accuracy	Test precision	Test recall	Test F1-score	Test IoU	Val. loss	Best val. accuracy	Best val. F1-score	Best val. IoU	Best val. IoU at epoch
UNet	0.03578	0.99012	0.89922	0.82791	0.86209	**0.86739**	0.0305	**0.9898**	0.8152	0.8685	**60**
VGG16 UNet	**0.03552**	**0.98950**	**0.89134**	0.81807	**0.85313**	0.86852	0.0282	**0.9898**	**0.8112**	**0.8688**	62
Trans UNet	**0.04673**	0.99031	**0.91538**	**0.81558**	0.86261	0.87531	**0.0332**	0.9904	0.8352	0.8757	66
Deep Residual UNet	0.03469	**0.99043**	0.90193	**0.83392**	**0.86659**	**0.88768**	**0.0235**	**0.9913**	**0.8458**	**0.8882**	**70**

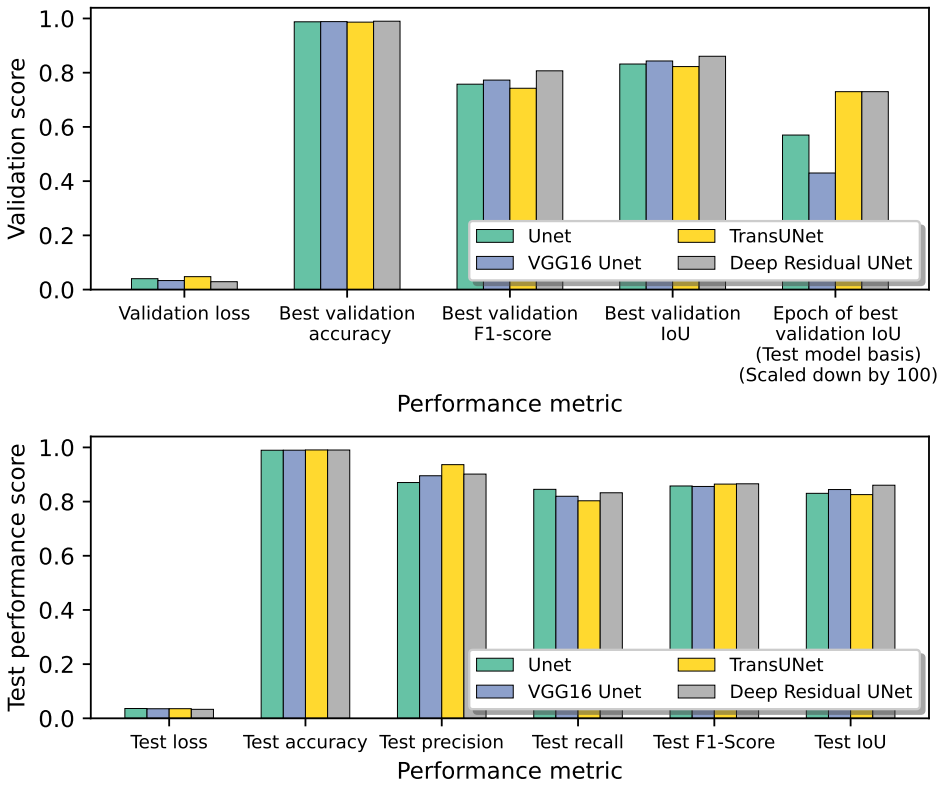


Figure 4.3: For 1500 train data, visual representation of the performance comparison of the chosen segmentation models on validation dataset and test dataset.

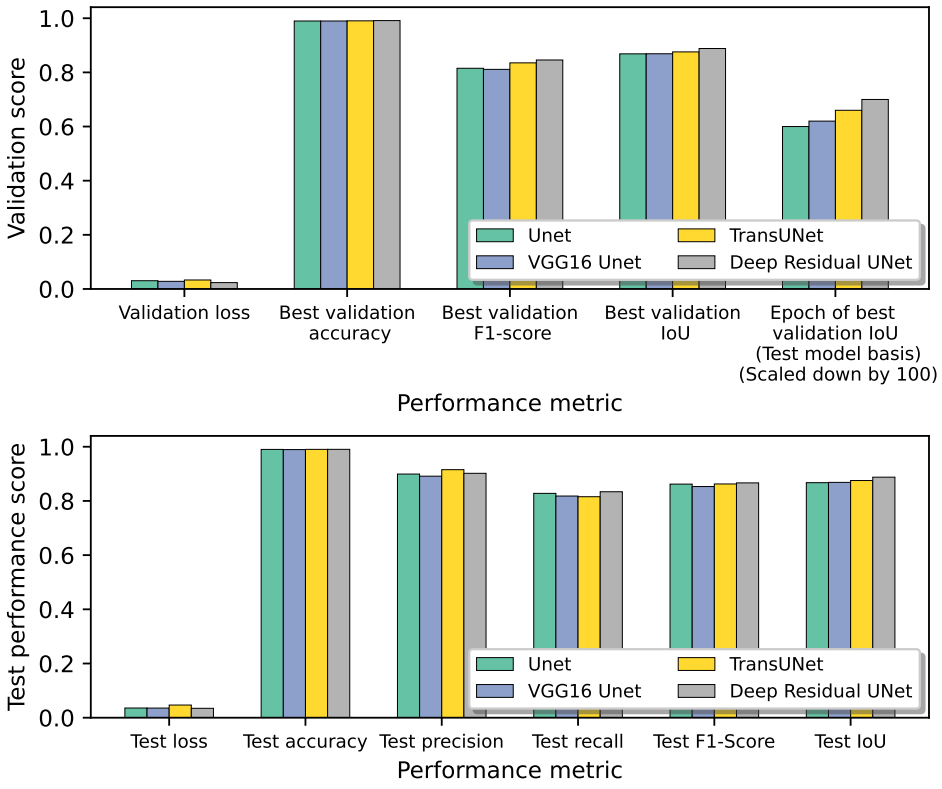


Figure 4.4: For 2500 train data, visual representation of the performance comparison of the chosen segmentation models on validation dataset and test dataset.

Model learning curves: Model learning curves for the selected segmentation models using the novel segmentation dataset with 616, 1500, and 2500 train data have been provided below.

- **Model learning curves with 616 train data:** The chosen models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.5. Figure 4.5 (a) and (b) indicate that the UNet model began to overfit after epoch 20 and its validation IoU plateaued around the same epoch. For testing, the UNet model from Epoch 17 was selected. Figure 4.5 (e) shows the TransUNet model exhibited good generalisation. Figure 4.5 (g) reveals that the Deep Residual UNet model initially experienced instability during the first 10 epochs which stabilised after this period.

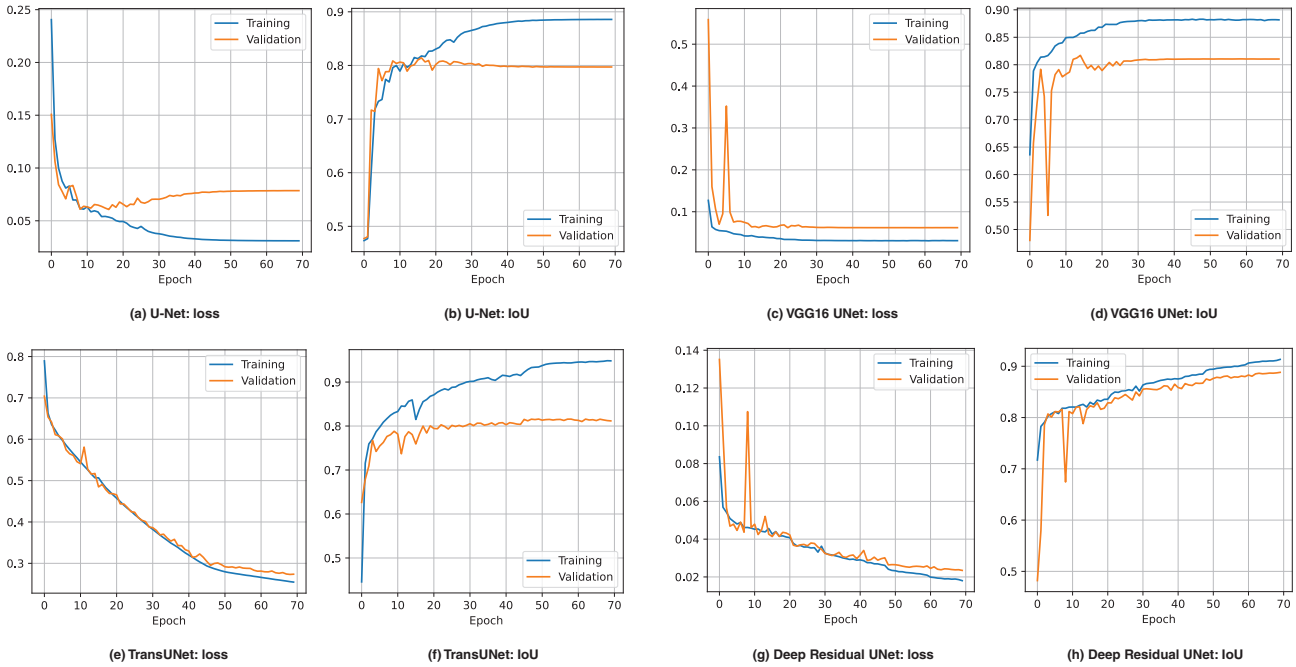


Figure 4.5: For 616 train data, learning curves for the selected segmentation models visualising the validation performance over all epochs.

- **Model learning curves with 1500 train data:** The chosen segmentation models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.6. Figure 4.6 (a) and (b) indicate that the UNet model began to overfit after epoch 57, where the validation IoU also plateaued. Therefore, the UNet model from epoch 57 was selected for testing. Figure 4.6 (c) and (g) show initial training instability in the VGG16 UNet and Deep Residual UNet models during the first 10 epochs, followed by stabilisation within 20 epochs. Figure 4.6 (e) shows the TransUNet model exhibited good generalisation.

- **Model learning curves with 2500 train data:** The chosen segmentation models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.7. Figure 4.7 (a) shows that the UNet started to overfit after epoch 60 and its validation IoU also plateaued around the mentioned epoch. Therefore, the model from epoch 60 was selected for testing. Figure 4.7 (c) and (g) show VGG16 UNet and Deep Residual UNet had instability in the first 5 epochs and 10 epochs respectively, which stabilised after epochs 5 and 10 respectively. Therefore, models from epoch 5 and epoch 10 have been used for testing respectively. Figure 4.7 (e) demonstrates that the TransUNet model exhibited good generalisation.

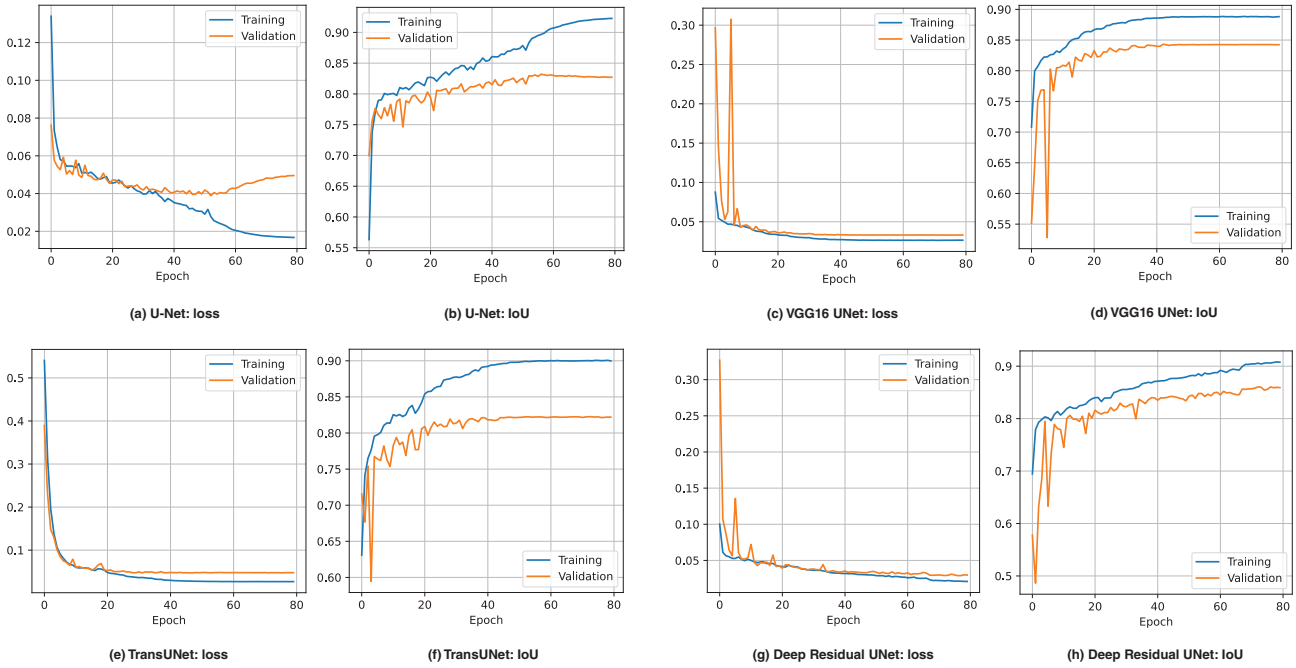


Figure 4.6: For 1500 train data, learning curves for the selected segmentation models visualising the validation performance over all epochs.

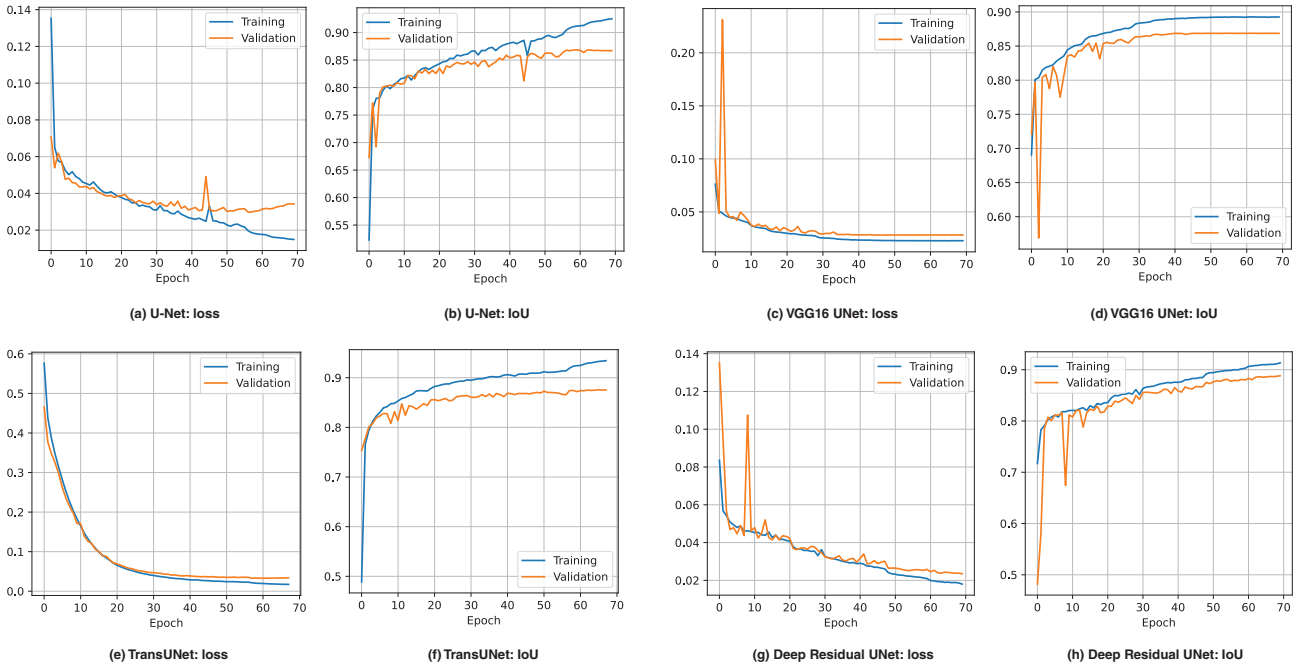


Figure 4.7: For 2500 train data, learning curves for the selected segmentation models visualising the validation performance over all epochs.

Test results: Below are the test results for the selected segmentation models using 616, 1500 and 2500 train data.

- **Test performance with 616 train data:** Figure 4.2 and Table 4.1 reveal that all the models had nearly equal accuracy with approximately 98%. Except for Deep Residual UNet, all the models had approximately equal performance in precision with around 86%. For precision, the lowest score was observed in TransUNet with nearly 86%, while the highest score was observed in Deep Residual UNet with approximately 87.16%. For the test loss, Unet had the lowest test loss with 3.89%, while TransUNet had the highest score with nearly 28.5%. Additionally, TransUNet had been observed to have the lowest scores in accuracy, recall, and F1-score with 98.74%, 78.93%, and 82.32% respectively, while UNet had the highest with 98.92%, 84.25% and 85.3% respectively. In terms of IoU, UNet had shown the lowest performance with slightly over 80%, while the Deep Residual UNet had the highest score with 82.42%.

- **Test performance with 1500 train data:** Similar to the results from train data 616, Figure 4.3 and Table 4.2 show that all the models had nearly equal accuracy ranging from 98% to 99%. For the test loss, Unet had the highest test loss at 3.6%, while Deep Residual UNet had the lowest with 3.33% approximately. For accuracy, precision and recall, Unet had the highest with 0.98.95%, 87.04% and 84.52% respectively and TransUNet had the highest with 99.06%, 93.64% and 0.80.3% respectively. Additionally, for F1-Score and IoU, Deep Residual UNet had the highest with approximately 86.57% and 86.04% respectively, while VGG16 UNet is observed to have the lowest F1-Score with around 85.58% and TransUNet had the lowest IoU with 82.58%.

- **Test performance with 2500 train data:** Similar to the results from train data 616 and 1500, Figure 4.4 and Table 4.3 show that all the models had nearly equal accuracy ranging from 98% to 99%. For test loss, TransUnet had the highest with 4.5%, compared to the other models having around 3.5%. For accuracy and F1-Score, VGG16 UNet had the lowest with nearly 98.95% and 85.31% respectively, while Deep Residual UNet had the highest with approximately 99.04% and 86.66% respectively. TransUNet achieved the highest score in precision with 91.54%, while VGG16 UNet secured the lowest with 89.13%. TransUNet earned the lowest recall score at just over 81%, whereas Deep Residual UNet scored 83.39%. UNet had the lowest score in IoU with 86.74%, while the Deep Residual UNet secured the best score with 88.77%.

Confusion matrices: The confusion matrices for individual models trained with 616, 1500, and 2500 train data are depicted in Figure 4.8, Figure 4.9 and Figure 4.10 respectively. These confusion matrices show high accuracy in correctly classifying non-disease and disease pixels, with minimal misclassifications.

- **Model trained with 616 train data:** The misclassified disease pixel counts for (a) UNet, (b) VGG16 UNet, (c) TransUNet, and (d) Deep Residual UNet, stand at 7472, 7286, 7209, and 6891 respectively. UNet exhibited the highest disease pixel misclassifications, whereas Deep Residual UNet showed the lowest.

- **Model trained with 1500 train data:** The misclassified disease pixel counts for (a) UNet, (b) VGG16 UNet, (c) TransUNet, and (d) Deep Residual UNet, stand at 7077, 5384, 3068, and 5112 respectively. UNet exhibited the highest disease pixel misclassifications, whereas TransUNet showed the lowest.

- **Model trained with 2500 train data:** The misclassified disease pixel counts for (a) UNet, (b) VGG16 UNet, (c) TransUNet, and (d) Deep Residual UNet, stand at 5216, 5606, 4238, and 5097 respectively. VGG16 UNet exhibited the highest disease pixel misclassifications, whereas TransUNet showed the lowest.

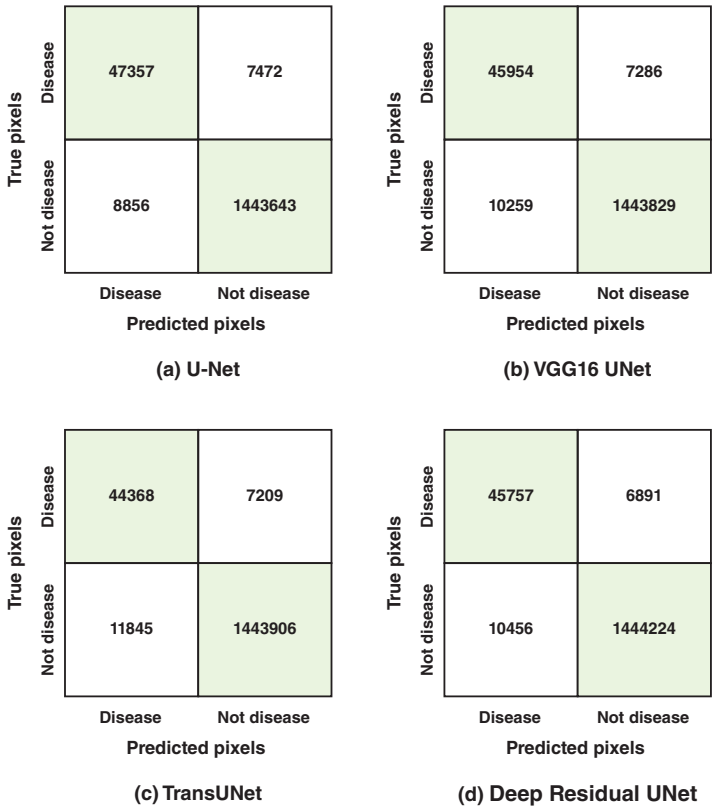


Figure 4.8: For 616 train data, confusion matrices of the chosen segmentation models on the test dataset.

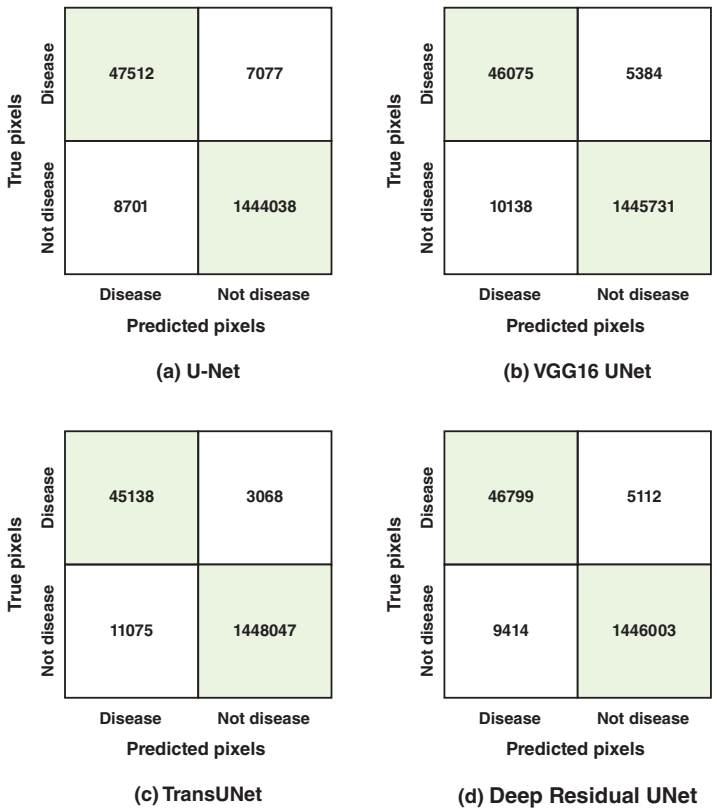


Figure 4.9: For 1500 train data, confusion matrices of the chosen segmentation models on the test dataset.

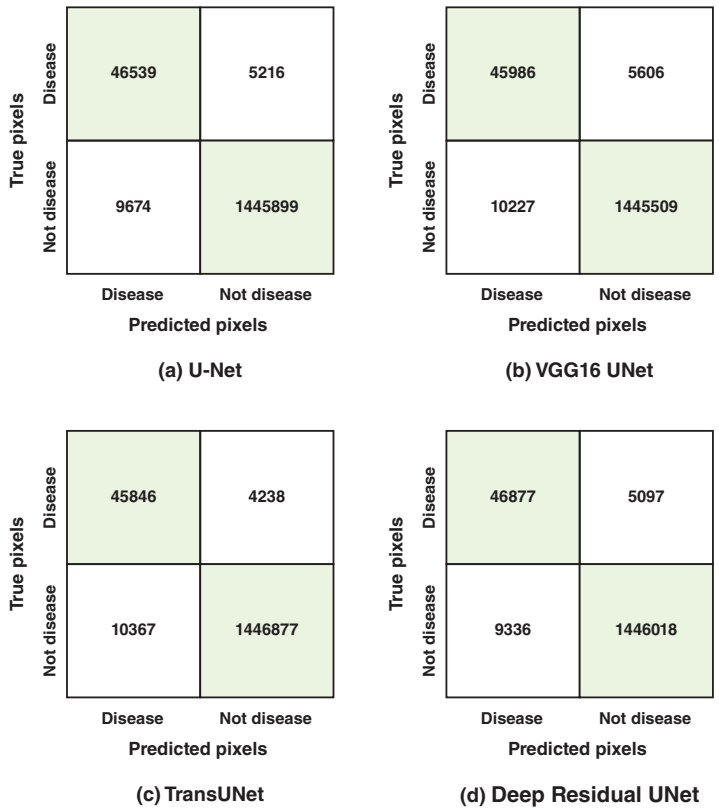


Figure 4.10: For 2500 train data, confusion matrices of the chosen segmentation models on the test dataset.

Qualitative test results: For the selected segmentation models, qualitative test results using 616, 1500, and 2500 train data have been provided below.

- **Model trained with 616 train data:** The masks predicted by the selected segmentation models have been presented in Figure 4.11 for qualitative performance analysis. For Sample 1, diseases at the leaf edges were accurately predicted by the VGG16 UNet, and TransUNet, while other models failed to do so. In Sample 3, the other models were unable to identify diseases at the leaf edges, while Deep Residual UNet did it successfully. In Sample 3, while Deep Residual UNet successfully identified diseases at the leaf edges, the other models were unable to do it. Additionally, none of the models were able to detect all small circular disease patterns in Sample 3. Figure 4.12 displays the mask predictions with associated probabilities for the selected models.

- **Model trained with 1500 train data:** The masks predicted by the selected models have been presented in Figure 4.13 for qualitative performance analysis. For sample 1, none of the models succeeded in properly detecting disease at the leaf edge. Compared to predictions using 616 train data for sample 3, all models detected most of the small circular disease patterns. Among them, TransUNet and Deep Residual UNet have shown the highest performance in detecting small patterns. However, for sample 3, only Deep Residual UNet and TransUNet accurately predicted diseases at the leaf edges. Figure 4.14 displays the mask predictions with associated probabilities for the selected models.

- **Model trained with 2500 train data:** The masks predicted by the selected models have been presented in Figure 4.15 for qualitative performance analysis. No model predicted disease at leaf edges in sample 1. For Sample 3, UNet detected fewer small circular disease patterns compared to other models and TransUNet detected the majority. For the same sample, except for UNet and VGG16 UNet, the rest of the models accurately predicted diseases at the leaf edges. Figure 4.16 displays the mask predictions with associated probabilities for the selected models.

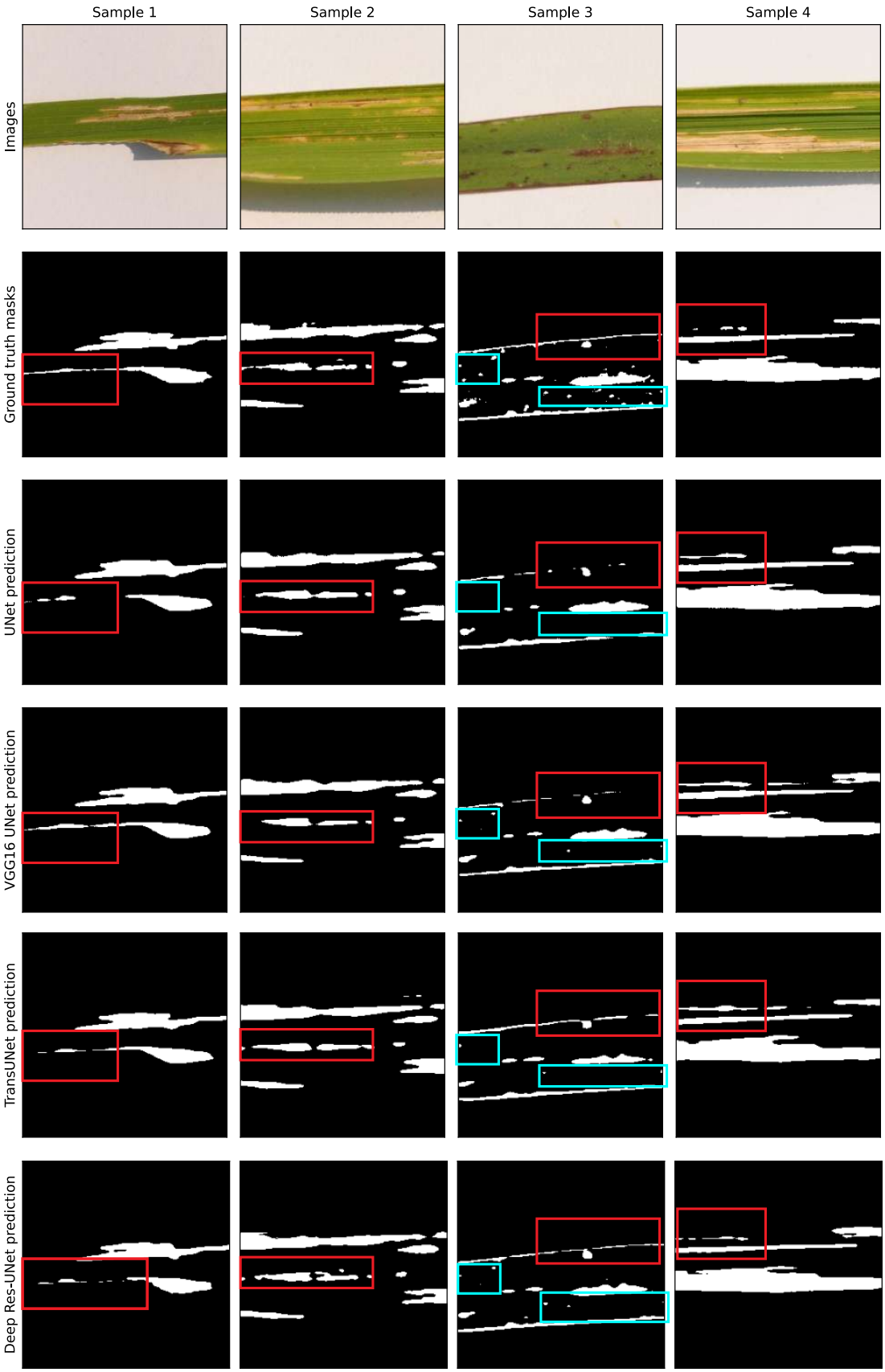


Figure 4.11: For 616 train data, comparison of mask predictions on the randomly selected test samples by the chosen segmentation models. The original image was compressed to meet file size limitation.

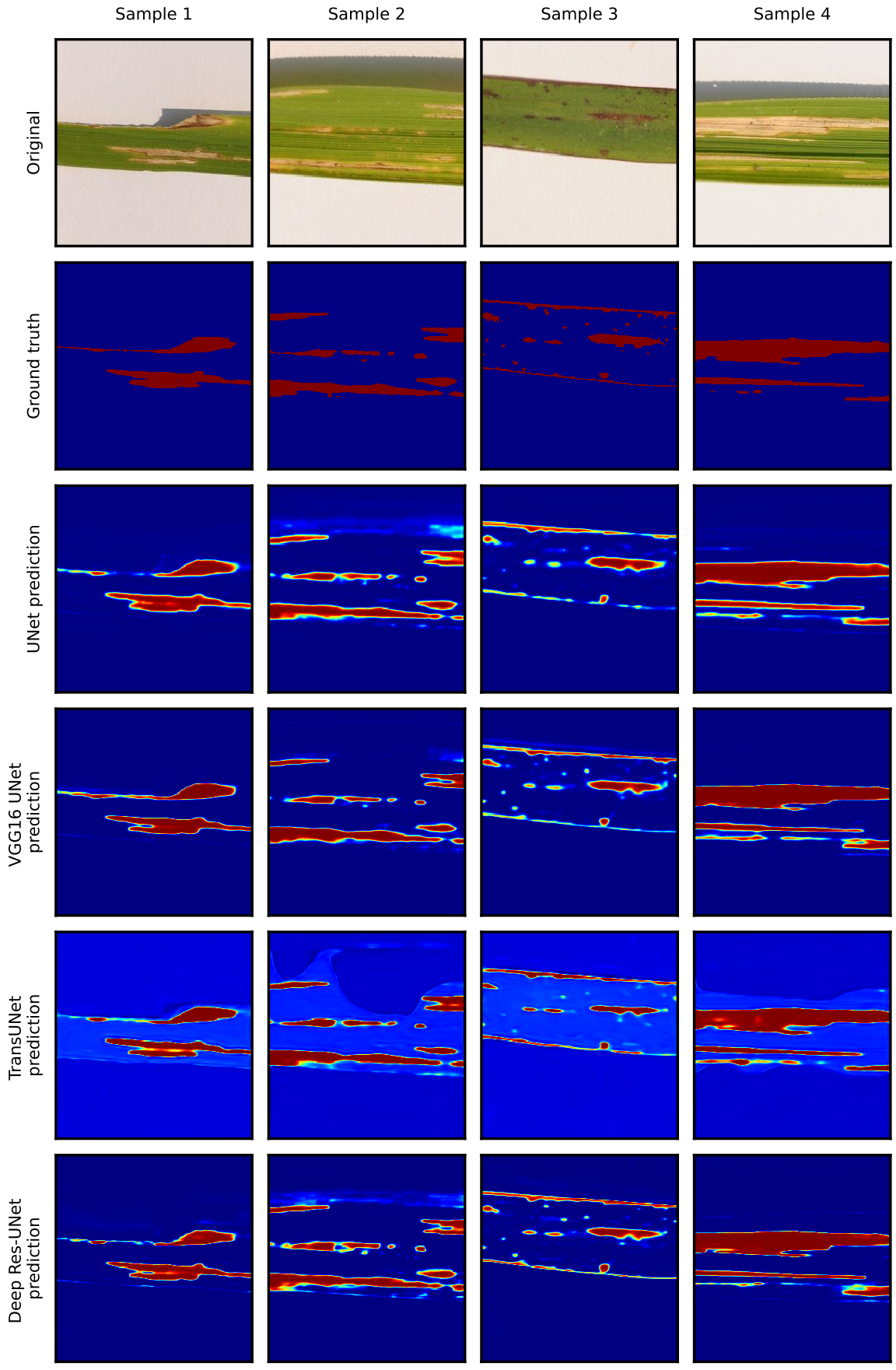


Figure 4.12: For 616 train data, comparison of mask prediction probabilities on the randomly selected test samples by the chosen segmentation models. The red colour represents higher probabilities. The original image was compressed to meet file size limitation.

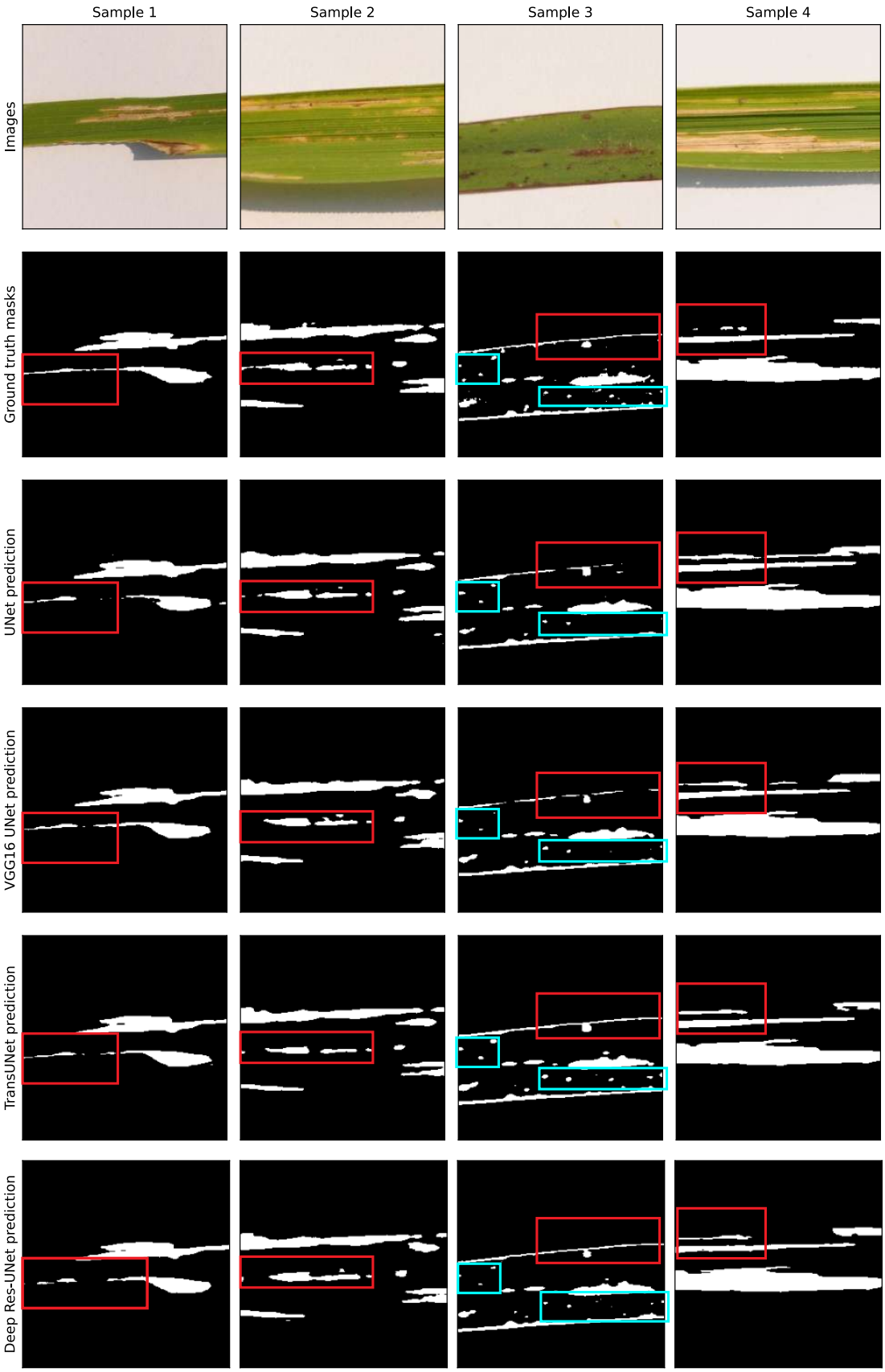


Figure 4.13: For 1500 train data, comparison of mask predictions on the randomly selected test samples by the chosen segmentation models. The original image was compressed to meet file size limitation.

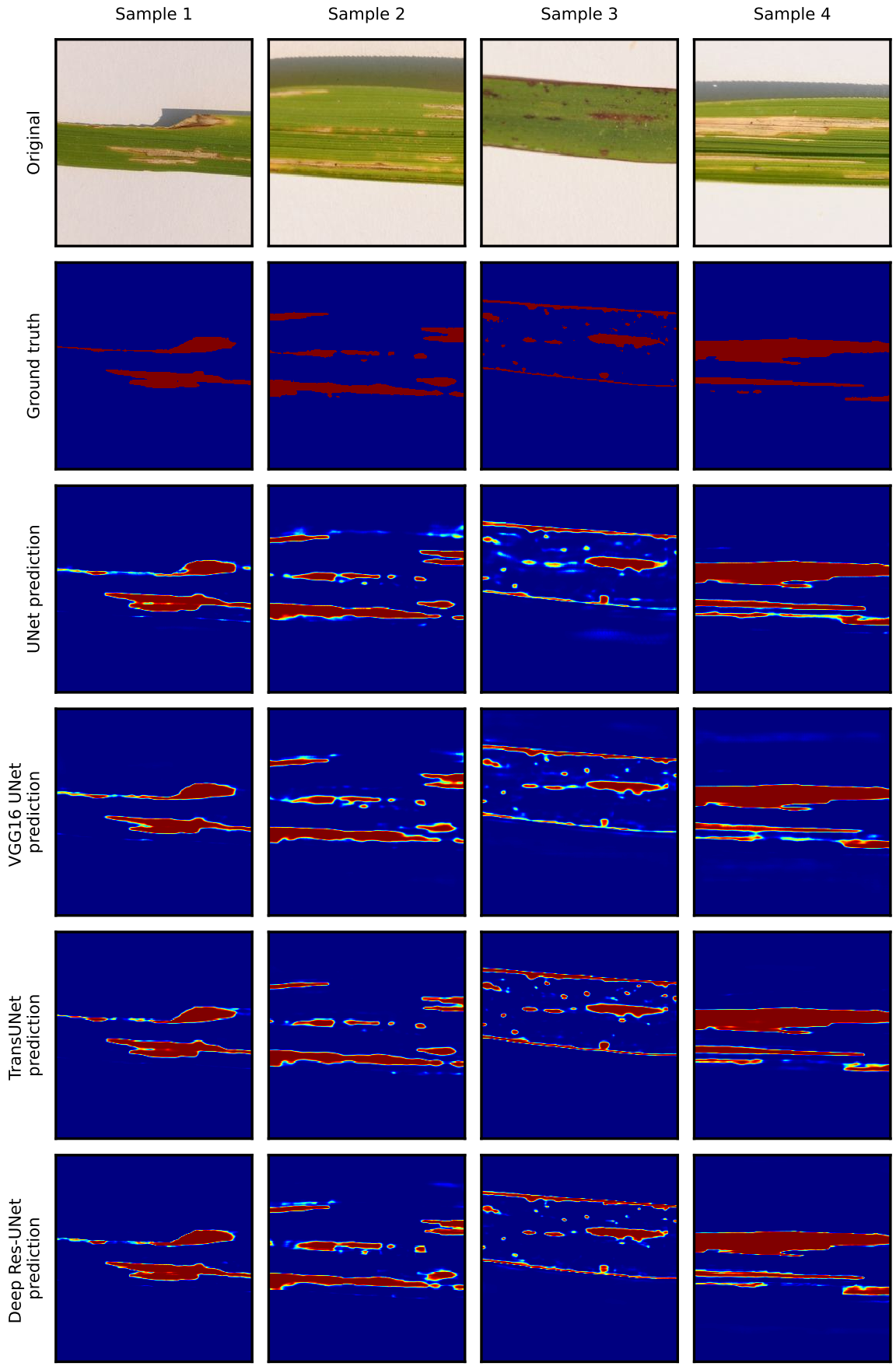


Figure 4.14: For 1500 train data, comparison of mask prediction probabilities on the randomly selected test samples by the chosen segmentation models. The red colour represents higher probabilities. The original image was compressed to meet file size limitation.

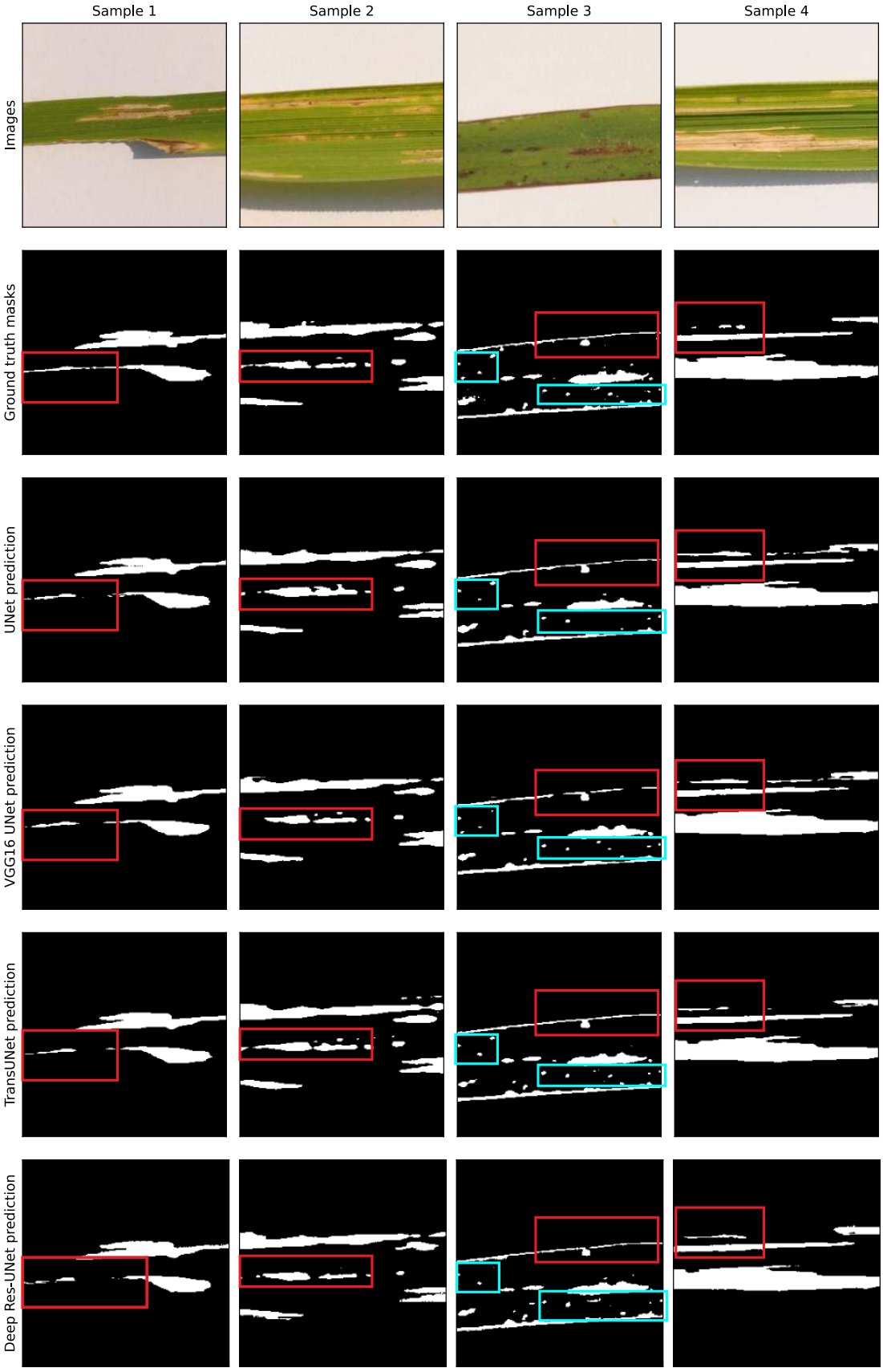


Figure 4.15: For 2500 train data, comparison of mask predictions on the randomly selected test samples by the chosen segmentation models. The original image was compressed to meet file size limitation.

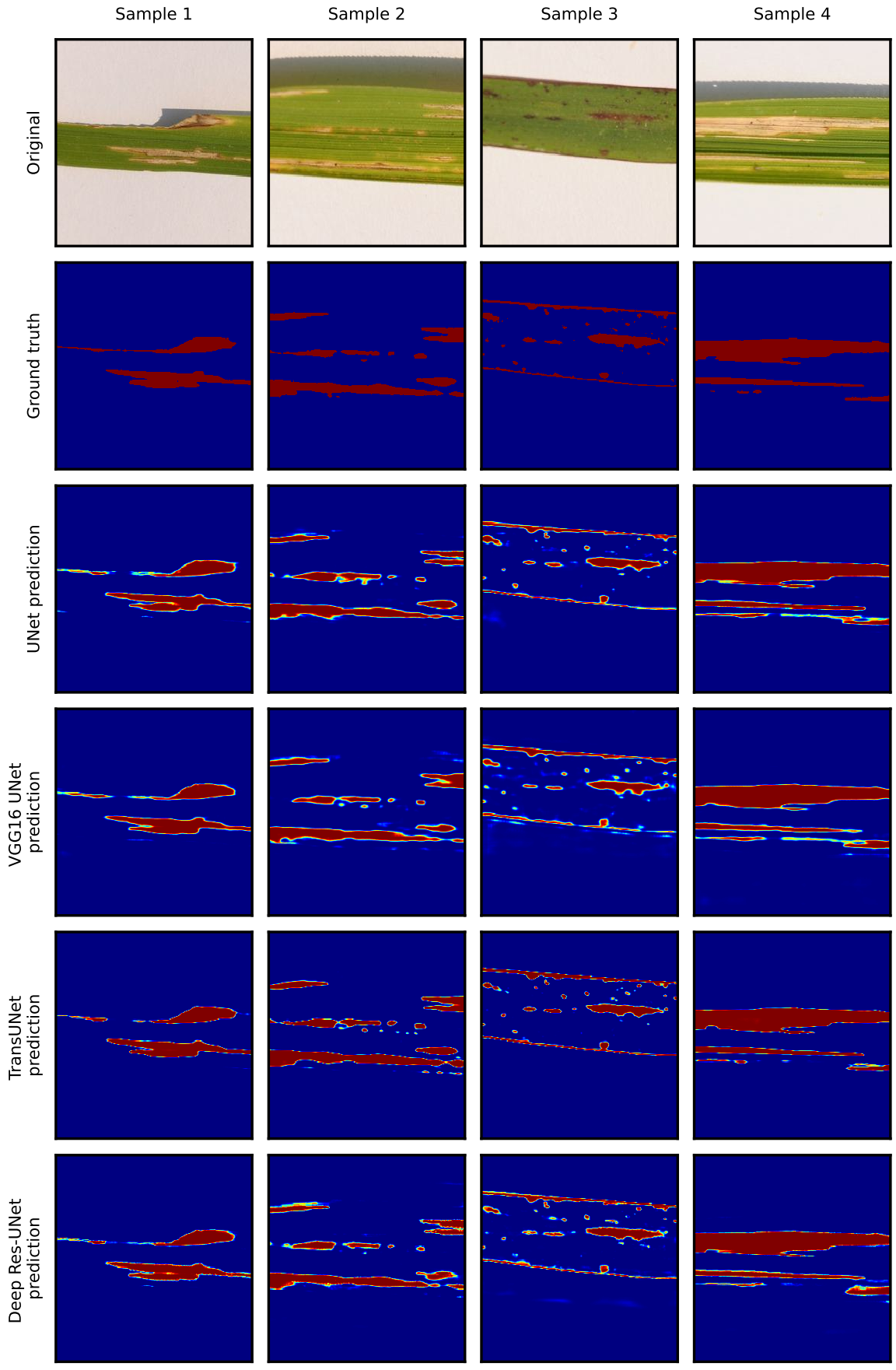


Figure 4.16: For 2500 train data, comparison of mask prediction probabilities on the randomly selected test samples by the chosen segmentation models. The red colour represents higher probabilities. The original image was compressed to meet file size limitation.

4.2 Classification results

We used four classification models on two paddy disease datasets, as described in 3.2.1, with various intensities of augmentation to measure performance on validation and test data. All the selected classification models were trained to maximise the validation F1-score with the same model training parameters specified in Table 3.2 and to improve convergence and overcome plateaus, LR scheduler was utilised. Additionally, to avoid overfitting, early stopping was used. For testing, instead of using models from the final epoch, model weights with the highest validation F1 score were selected. The approach taken to prevent overfitting is supported by the nearly comparable performance between validation and test results from Table 4.4, Table 4.5 and Table 4.6.

4.2.1 Results obtained for Paddy Doctor dataset

The obtained results for the dataset [16], have been presented in three sections: individual model test performance comparison, training time comparison, and validation and test performance comparison across three augmentation intensities.

Test performance comparison of individual models across three augmentation intensities

From Figure 4.17, we can observe that DenseNet121 and ViT had shown a low performance with basic augmentation, whereas Ensemble2 demonstrated the highest. However, in the case of MobileNet, its performance decreased as the level of augmentation increased. On the other hand, with extensive augmentation, DenseNet121 and ViT showed high performance.

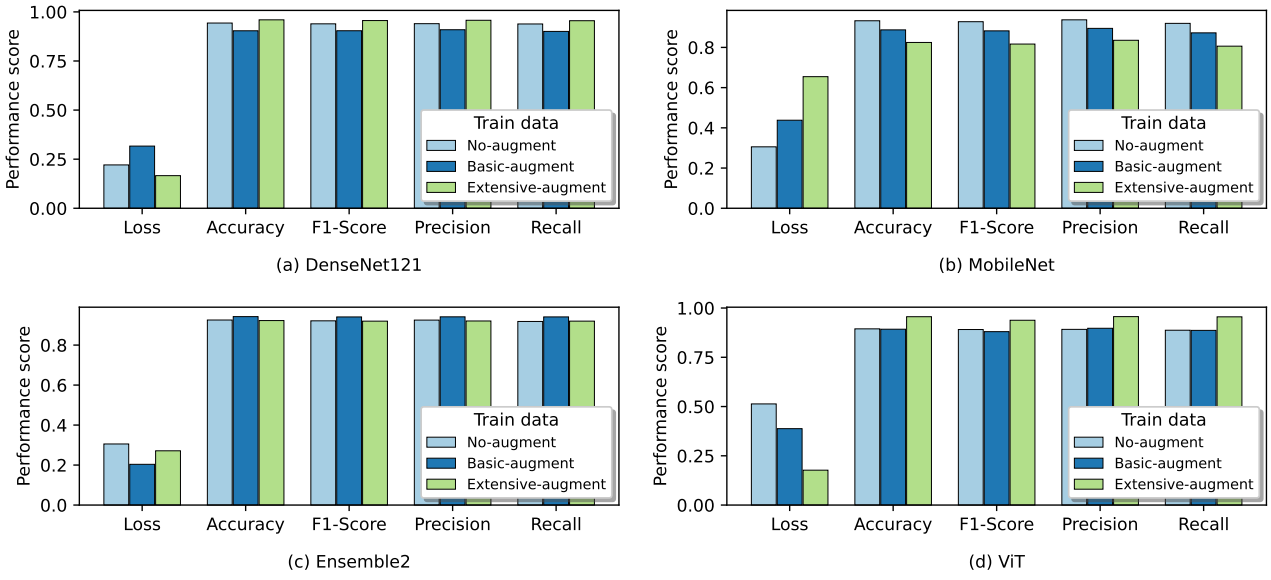


Figure 4.17: For the paddy doctor dataset, visualisation of the selected classification models' test performance comparison across three augmentation intensities.

Training time comparison across three augmentation intensities

Figure 4.18 shows that ViT required the longest training time across all augmentation variations. The longest training time required by ViT was for extensive augmentation and it was nearly 7 days, about 2-6 times higher than the rest of the models in this study. The MobileNet had the shortest training time for all augmentation intensities.

Validation and test performance comparison across three augmentation intensities

The performance of the chosen four classification models with three augmentation intensities (none, basic, and extensive) on the training samples is detailed in Table 4.4, Table 4.5, and Table 4.6 respectively and a visual representation of the same data with a bar-chart format has been presented in Figure 4.19, Figure 4.20 and Figure 4.21 for better comprehension.

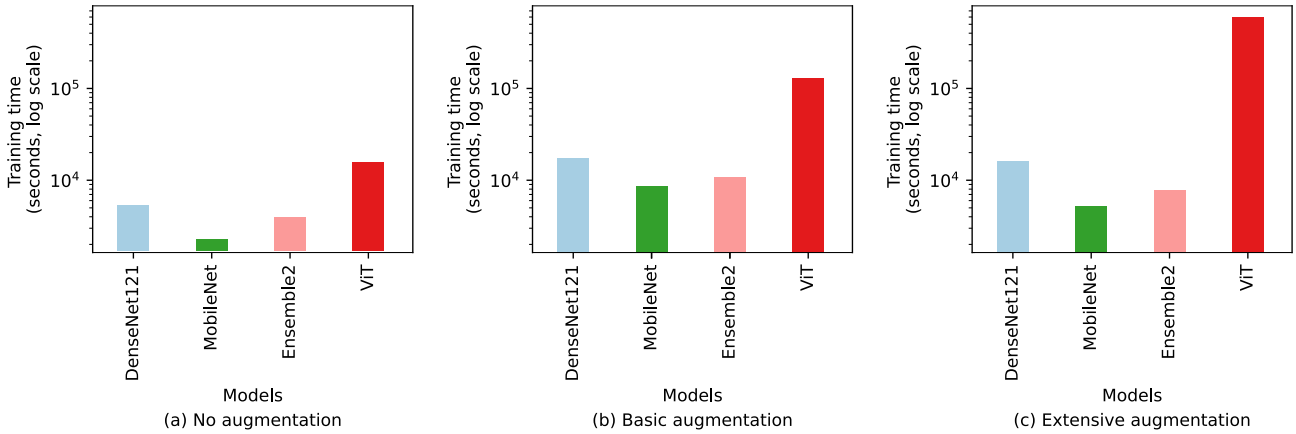


Figure 4.18: For the paddy doctor dataset, training time comparison of the selected classification models across three augmentation intensities. Due to high variation in time data, it was scaled logarithmically before plotting.

Table 4.4: For the Paddy Doctor dataset, performance comparison of the selected classification models without augmentation.

Model name	Test loss	Test accuracy	Test F1-score	Test precision	Test recall	Val. loss	Best val. accuracy	Best val. F1-score	Best val. F1-score at epoch	Train time (seconds)
Dense Net121	**0.22076**	**0.943297**	**0.939239**	**0.940036**	**0.938562**	**0.2419**	**0.939**	**0.9387**	37	5127.62
Mo-bileNet	0.30558	0.932512	0.92787	0.937177	0.919594	0.3099	0.9305	0.9277	48	**2159.53**
Ensem-ble2	0.305490	0.925424	0.921318	0.924991	0.918173	0.3359	0.9228	0.9209	**59**	3762.88
ViT	**0.513310**	**0.894299**	**0.890865**	**0.891685**	**0.887280**	**0.5049**	**0.8966**	**0.8931**	**34**	**14864.92**

Validation results: Validation results obtained using the Paddy Doctor [16] dataset over three augmentation variations have been presented below.

- **Without augmentation:** From Table 4.4 and Figure 4.19, it is observed that ViT had the highest validation losses with approximately 50%, while DenseNet121 had the lowest loss at 24.19%. Additionally, DenseNet121 had the highest performance in accuracy and F1-score with nearly 94%, while ViT had the lowest with nearly 90%. Additionally, Ensemble2 took the highest number of epochs (59) to achieve the best validation F1-score, while ViT required the lowest (54).
- **With basic augmentation:** From Table 4.5 and Figure 4.20, we can observe that similar to the results without augmentation, DenseNet121 had the highest performance in accuracy and F1-score with nearly 97%, while ViT had the lowest with nearly 87%. In the case of validation losses, DenseNet121 had the lowest with 13.44% and ViT had the highest with nearly 48%. Additionally, DenseNet121 took 135 epochs (longest) to achieve the best validation F1-score, while MobileNet required the shortest (83 epochs). We can observe that compared to training without augmentation, the overall training time has increased.
- **With extensive augmentation:** From Table 4.6 and Figure 4.21, we can observe that similar to the results without augmentation, DenseNet121 had the highest validation performance in accuracy and F1-score with nearly 97%, while ViT had the lowest with nearly 92%. In the case of validation losses, DenseNet121 had the lowest with just over 11% and ViT had the highest with nearly 28%. Additionally, DenseNet121 took the shortest (63 epochs) to achieve the best validation F1 score, while ViT required the longest (133 epochs).

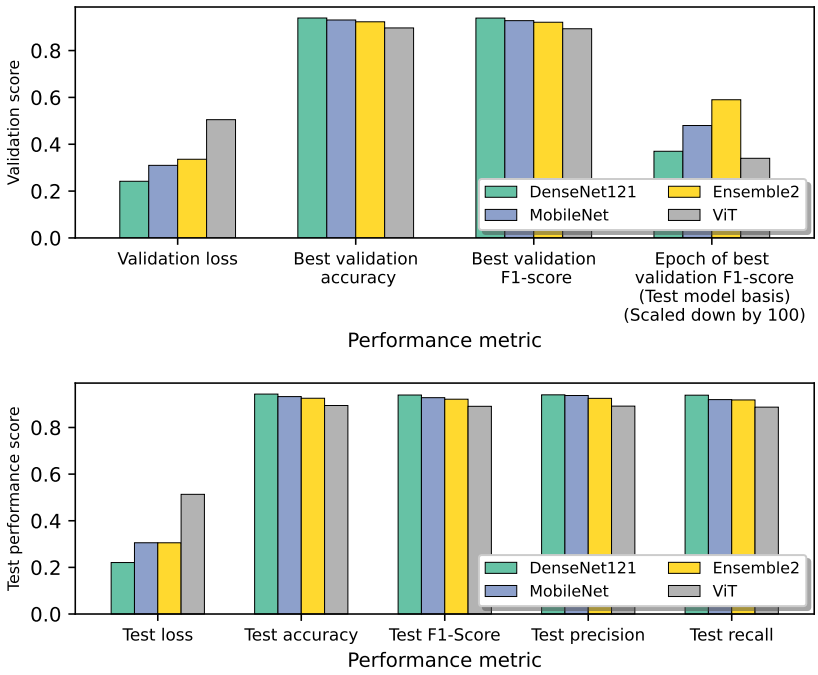


Figure 4.19: For the Paddy Doctor dataset, visual representation of the performance comparison of the selected classification models without augmentation.

Table 4.5: For the Paddy Doctor dataset, performance comparison of the selected classification models with basic augmentation.

Model name	Test loss	Test accuracy	Test F1-score	Test precision	Test recall	Val. loss	Best val. accuracy	Best val. F1-score	Best val. F1-score at epoch	Train time (seconds)
Dense Net121	0.316642	0.903852	0.904045	0.909190	0.900674	**0.1344**	**0.9660**	**0.9661**	**135**	17497.73
Mo-bileNet	**0.437932**	**0.887211**	0.882305	**0.894682**	**0.872376**	0.1973	0.9537	0.9529	**83**	**8658.68**
Ensem-ble2	**0.203745**	**0.942373**	**0.940769**	**0.941301**	**0.940903**	0.2176	0.9371	0.9402	103	10755.67
ViT	0.387732	0.892450	**0.880069**	0.897241	0.886579	**0.4768**	**0.8692**	**0.8679**	96	**130020.81**

Table 4.6: For the Paddy Doctor dataset, performance comparison of the selected classification models with extensive augmentation.

Model name	Test loss	Test accuracy	Test F1-score	Test precision	Test recall	Val. loss	Best val. accuracy	Best val. F1-score	Best val. F1-score at epoch	Train time (seconds)
Dense Net121	**0.166118**	**0.959631**	**0.955953**	**0.957332**	0.955131	**0.1139**	**0.9718**	**0.9723**	**63**	16013.36
Mo-bileNet	**0.654521**	**0.824653**	**0.816688**	**0.835446**	**0.806223**	0.1842	0.9541	0.9534	71	**5249.03**
Ensem-ble2	0.271355	0.922651	0.919742	0.920576	0.919810	0.2566	0.9294	0.9292	68	7896.46
ViT	0.177048	0.955932	0.937889	0.956371	**0.955309**	**0.2814**	**0.9220**	**0.9238**	**133**	**599582.55**

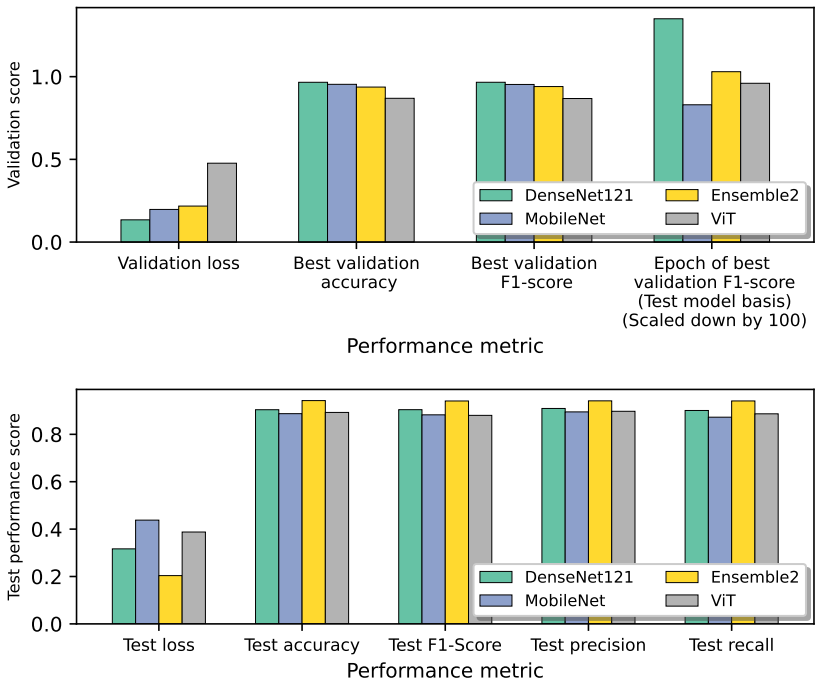


Figure 4.20: For the Paddy Doctor dataset, visual representation of the performance comparison of the selected classification models with basic augmentation.

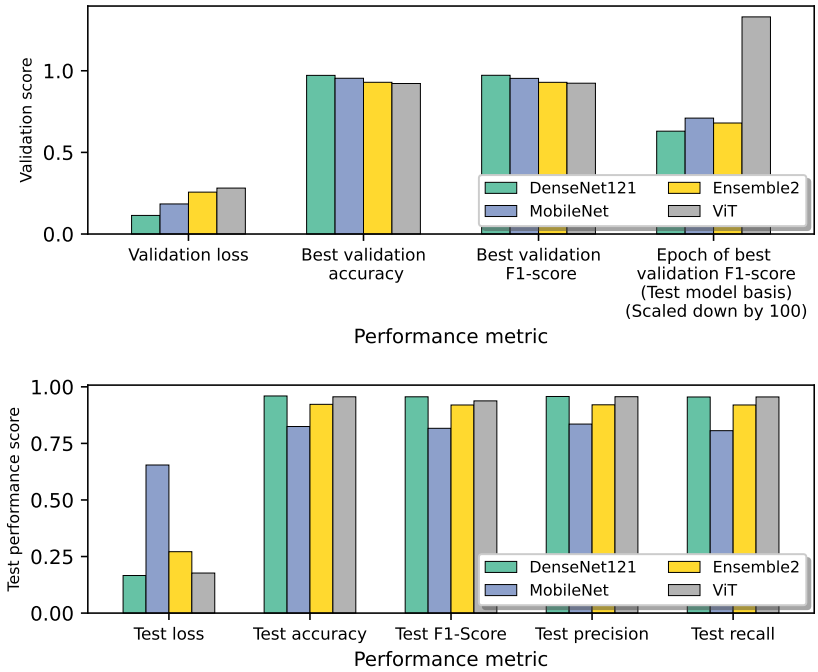


Figure 4.21: For the Paddy Doctor dataset, visual representation of the performance comparison of the selected classification models with extensive augmentation.

Model learning curves: Model learning curves for the selected classification models using the Paddy Doctor dataset over three augmentation variations have been provided below.

- **Without augmentation:** The chosen classification models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.22. From this figure, we can observe that none of the models had overfitting. Figure 4.22 (a) and (c) indicate that DenseNet121 and MobileNet had initial instability within the first 10 and 25 epochs respectively, but stabilised after this period. Figure 4.22 (a) and (c) show that DenseNet121 and MobileNet had a slight gap between the training and validation results, while Figures 4.22 (e) reveal that Ensembl2 had a modest gap.

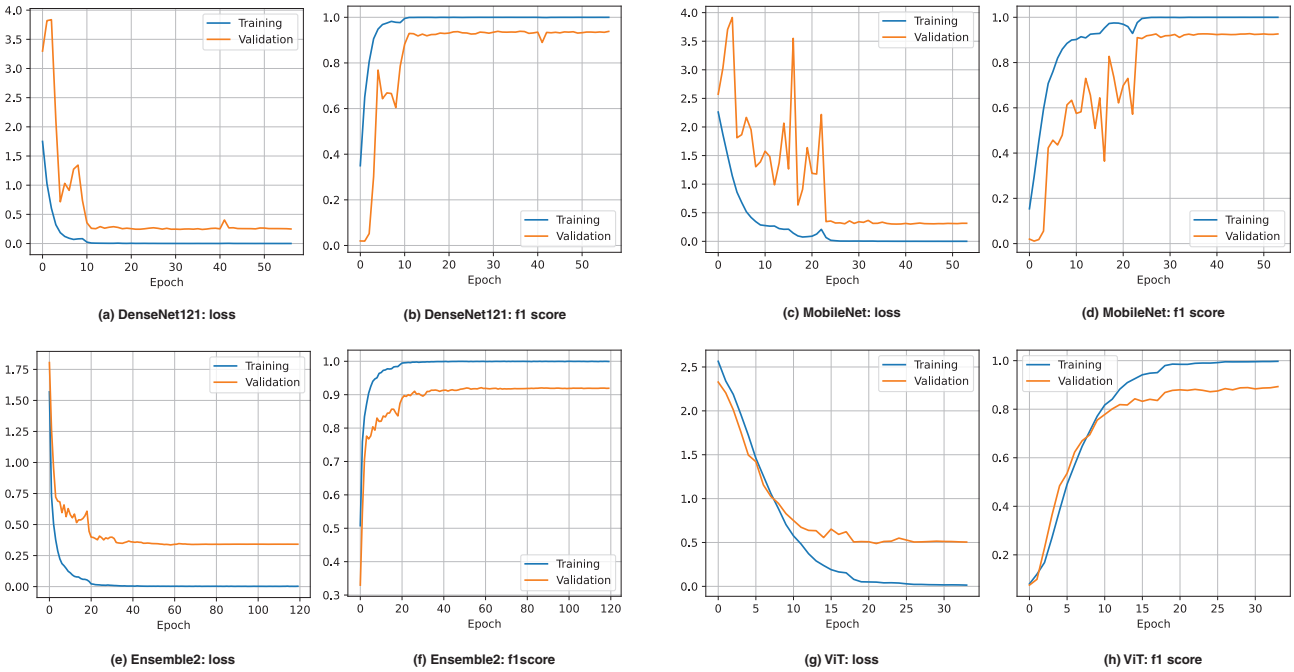


Figure 4.22: For the paddy doctor dataset without augmentation on train data, visualising the learning curves for the selected classification models.

- **With basic augmentation:** The chosen classification models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.23. From Figure 4.23 (g), we can observe that ViT started to overfit around epoch 90. Figure 4.23 (c) indicates that MobileNet had initial instability within the first 25 epochs, but stabilised after this period. Figure 4.23 (c) and (e) show that the MobileNet and Ensemble2 models exhibited smaller gaps between training and validation results compared to those without augmentation.

- **With extensive augmentation:** The chosen classification models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.24. Figure 4.24 demonstrates that none of the models had overfitting. Figure 4.24 (a) and (c) indicate that DenseNet121 and MobileNet had initial instability within the first 20 epochs, but stabilised after this period. Figure 4.24 (c), (e), and (g) show that the MobileNet, Ensemble2, and ViT models had smaller gaps between training and validation results compared to those without or with basic augmentation. Figure 4.24 (a) demonstrates that DenseNet121 exhibited good generalisation.

Test results: Test results obtained using the Paddy Doctor [16] dataset over three augmentation variations have been presented below.

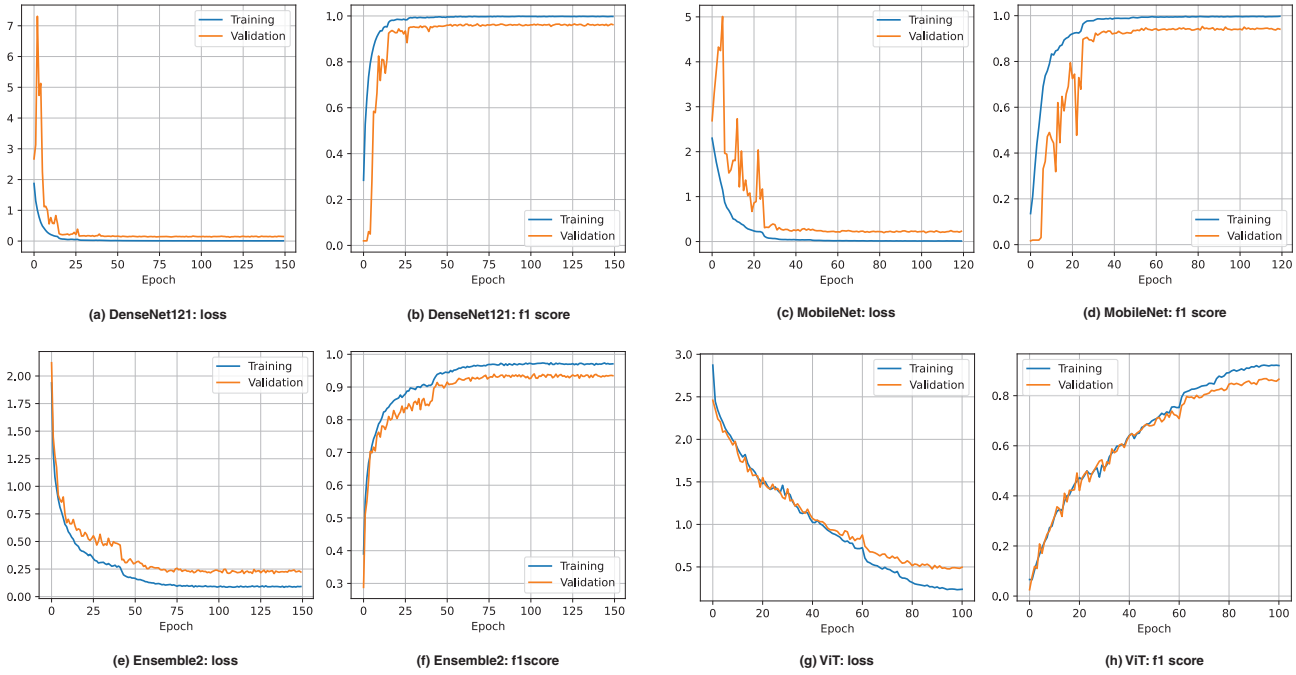


Figure 4.23: For the paddy doctor dataset with basic augmentation on train data, visualising the learning curves for the selected classification models.

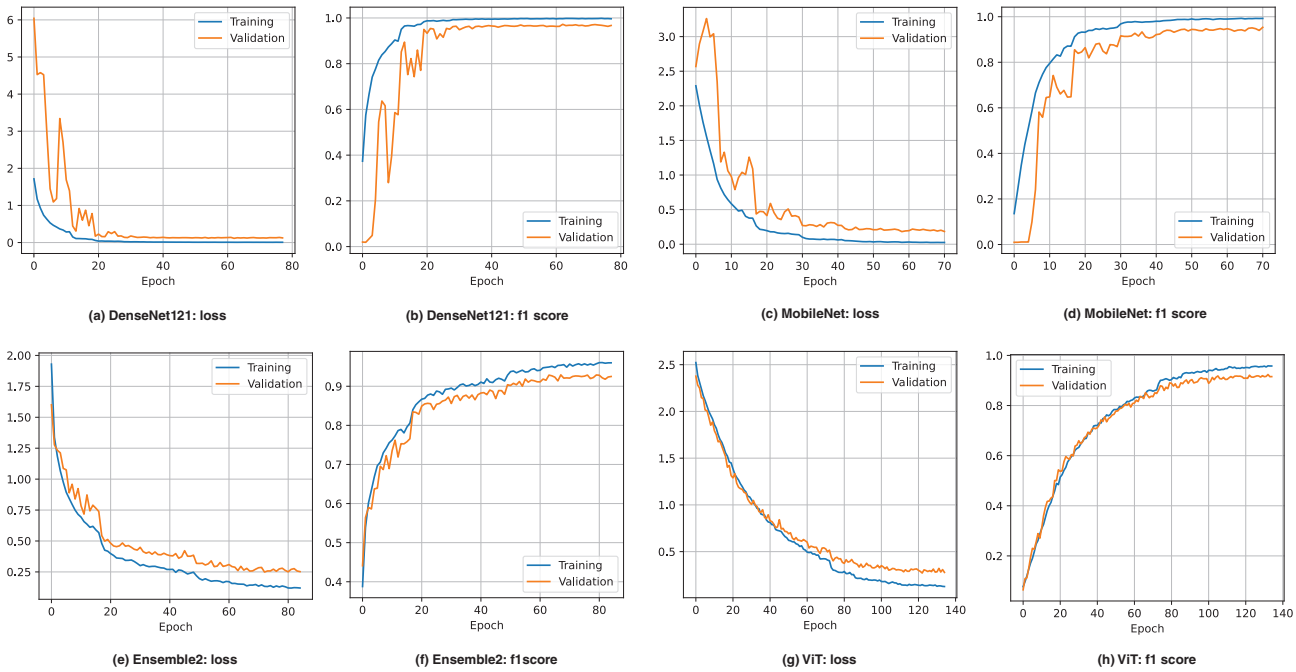


Figure 4.24: For the paddy doctor dataset with extensive augmentation on train data, visualising the learning curves for the selected classification models.

- **Without augmentation:** From Table 4.4 and Figure 4.19, we can observe that similar to the validation loss without augmentation for the test loss, DenseNet121 had the lowest approximately 22% and ViT had the highest with nearly 51%. Additionally, DenseNet121 achieved the highest performance in accuracy, F1-score, precision and recall with 94.33%, 93.92%, 94%, and 93.86% respectively, while ViT exhibited the lowest with 89.43%, 89.09%, 89.16%, and 88.73% respectively. The rest of the models achieved good performance, ranging from 91% to 93% for all the performance metrics.
- **With basic augmentation:** From Table 4.5 and Figure 4.20, it is observed that Ensemble2 had the lowest test loss with approximately 21% and MobileNet had the highest with 43.79%. Meanwhile, MobileNet exhibited the lowest performance in accuracy, precision, and recall with 88.72%, 89.47%, and 87.24% respectively, whereas Ensemble2 achieved the highest with 94.24%, 94.13%, and 94.1% respectively. The rest of the models achieved good performance within the range of 88% to 90% for all the performance metrics. Additionally, ViT had shown almost comparable performance to MobileNet.
- **With extensive augmentation:** From Table 4.6 and Figure 4.21, we can see that MobileNet had the highest test losses at nearly 65%, while DenseNet121 had the lowest at approximately 16%. DenseNet121 had shown superior performance in all the metrics except for recall, while ViT achieved the highest in recall with 95.53%. MobileNet exhibited the lowest performance in accuracy, F1-score, precision, and recall with 82.47%, 81.67%, 83.54%, and 80.62% respectively. Additionally, ViT had shown almost comparable performance to DenseNet121.

Confusion matrices: Confusion matrices for the selected classification models using the Paddy Doctor dataset over three augmentation variations have been presented below to help understand the reason behind misclassifications and give an overview of the models' performance.

- **Without augmentation:** Figure 4.25 shows confusion matrices for classification models trained without augmentation. From these confusion matrices, it is observed that all the classification models often misclassified normal plants to be Hispa. Additionally, Ensemble2 frequently misclassified Hispa as normal plants. Moreover, ViT often confused Blast with Tungro and White stem borer with Yellow stem borer.
- **With basic augmentation:** Figure 4.26 displays confusion matrices for classification models trained with basic augmentation. These matrices reveal that except for DenseNet121, the rest of the models frequently misclassified normal plants as Hispa. Additionally, DenseNet121 and ViT often misidentified Hispa as normal plants. DenseNet121 and MobileNet commonly confused Blast with Tungro. Moreover, Ensemble2 and ViT typically misidentified Hispa as Leaf roller.
- **With extensive augmentation:** Figure 4.27 shows confusion matrices for classification models trained with extensive augmentation. DenseNet121 and ViT often mistook Hispa for normal plants, while Ensemble2 and ViT commonly misclassified normal plants for Hispa. Notably, MobileNet regularly confused several diseases—Blast, Brown spot, Downy mildew, Hispa, Leaf roller, Tungro, and White stem borer—as normal plants.

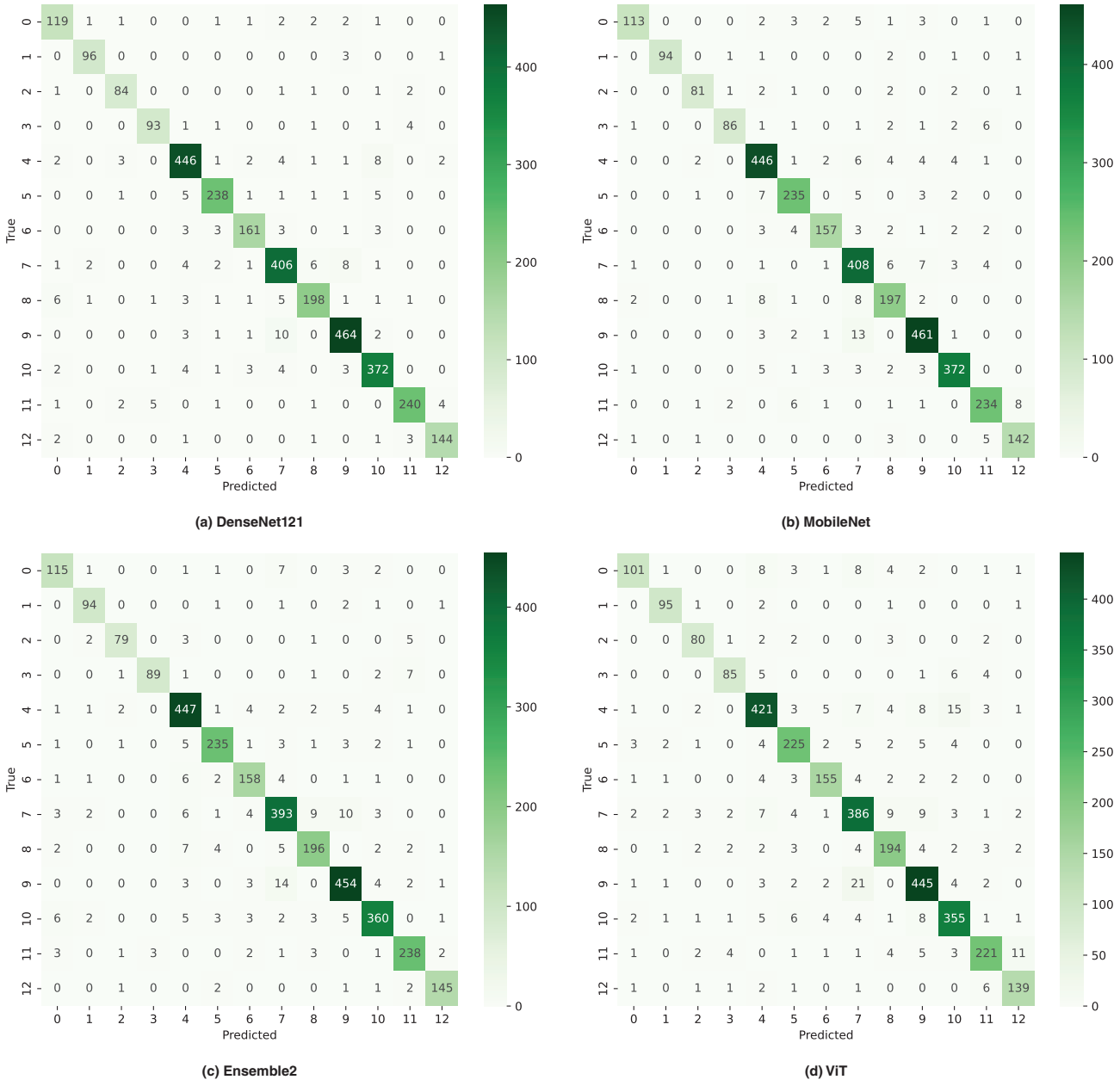


Figure 4.25: Visualising the confusion matrices for the selected classification models on the Paddy doctor test dataset without augmentation. For the label of the confusion matrices, 0 represents Bacterial leaf blight, 1 represents Bacterial leaf streak, 2 represents Bacterial panicle blight, 3 represents Black stem borer, 4 represents Blast, 5 represents Brown spot, 6 represents Downy mildew, 7 represents Hispa, 8 represents Leaf roller, 9 represents normal, 10 represents Tungro, 11 represents White stem borer, and 12 represents Yellow stem borer.

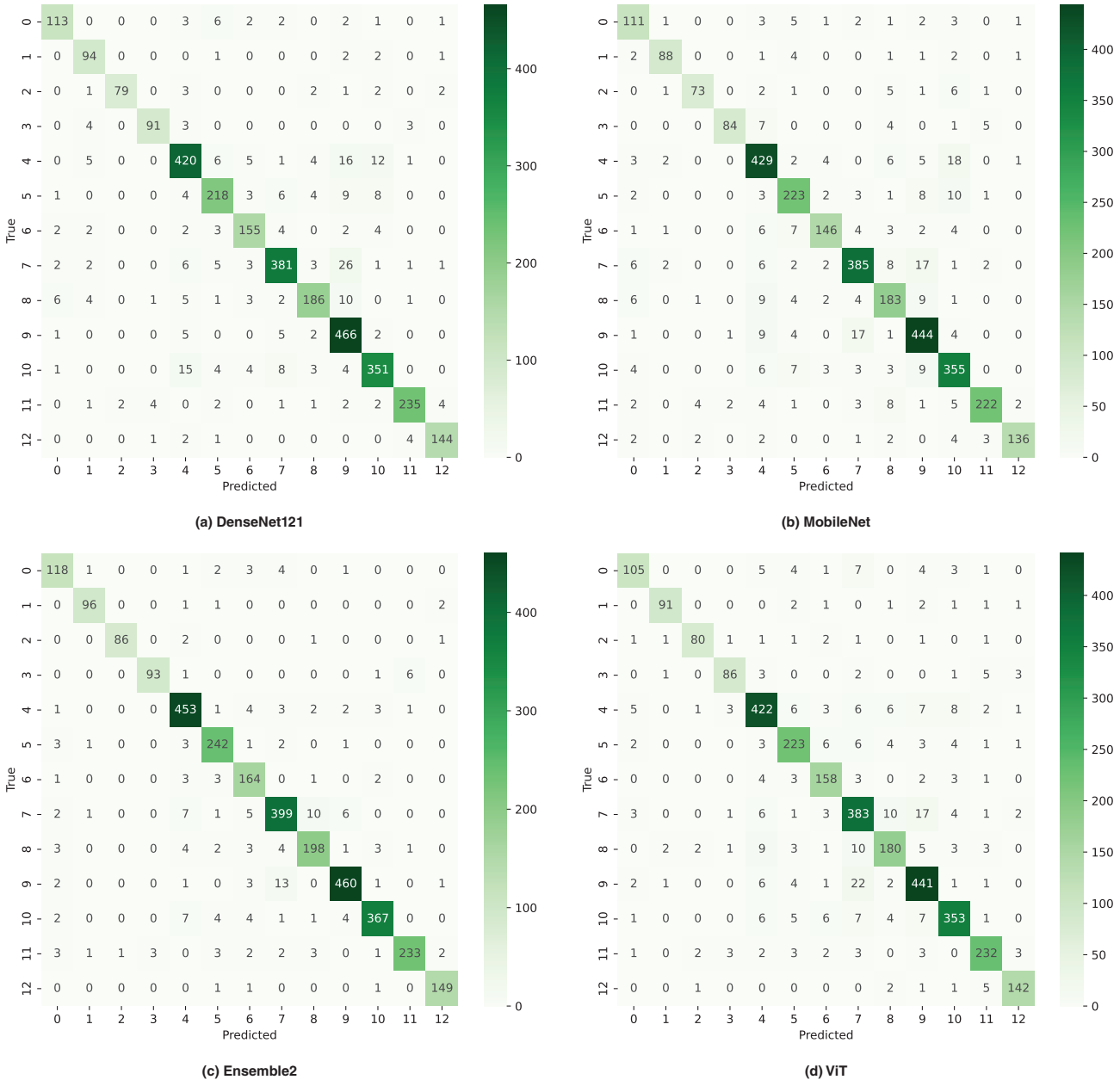


Figure 4.26: Visualising the confusion matrices for the selected classification models on the Paddy doctor test dataset with basic augmentation. For the label of the confusion matrices, 0 represents Bacterial leaf blight, 1 represents Bacterial leaf streak, 2 represents Bacterial panicle blight, 3 represents Black stem borer, 4 represents Blast, 5 represents Brown spot, 6 represents Downy mildew, 7 represents Hispa, 8 represents Leaf roller, 9 represents normal, 10 represents Tungro, 11 represents White stem borer, and 12 represents Yellow stem borer.

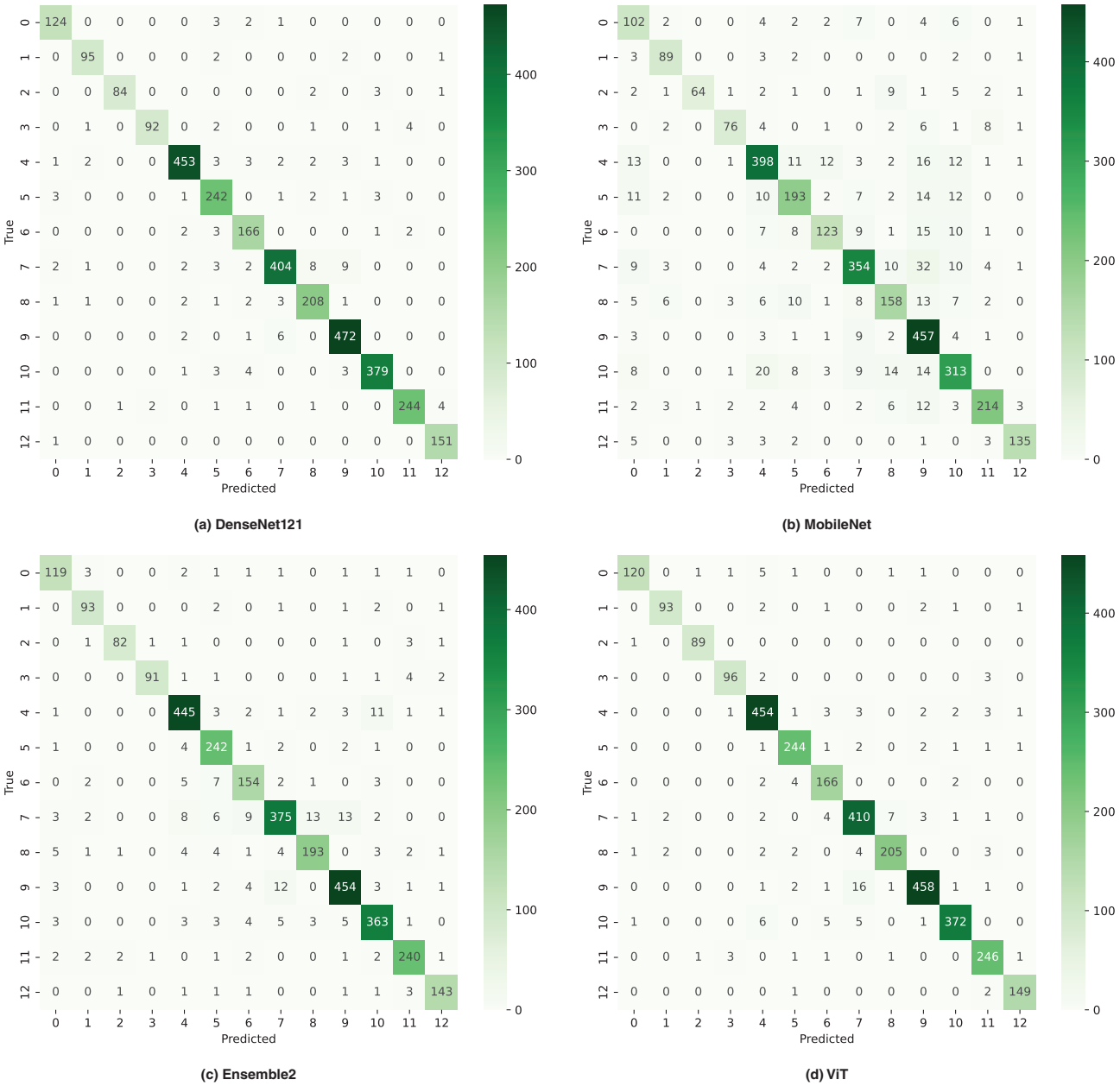


Figure 4.27: Visualising the confusion matrices for the selected classification models on the Paddy doctor test dataset with extensive augmentation. For the label of the confusion matrices, 0 represents Bacterial leaf blight, 1 represents Bacterial leaf streak, 2 represents Bacterial panicle blight, 3 represents Black stem borer, 4 represents Blast, 5 represents Brown spot, 6 represents Downy mildew, 7 represents Hispa, 8 represents Leaf roller, 9 represents normal, 10 represents Tungro, 11 represents White stem borer, and 12 represents Yellow stem borer.

4.2.2 Results obtained for Rice Leaf Disease dataset

The obtained results for the dataset [65], have been presented in three sections: individual model test performance comparison, training time comparison, and validation/test performance comparison across different augmentation intensities.

Test performance comparison of individual models across three augmentation intensities

From Figure 4.28, we can observe that DenseNet121 and Ensemble2 have nearly equal performance for all augmentation intensities. On the other hand, MobileNet's performance decreased as the level of augmentation increased, while ViT's performance increased as the level of augmentation increased.

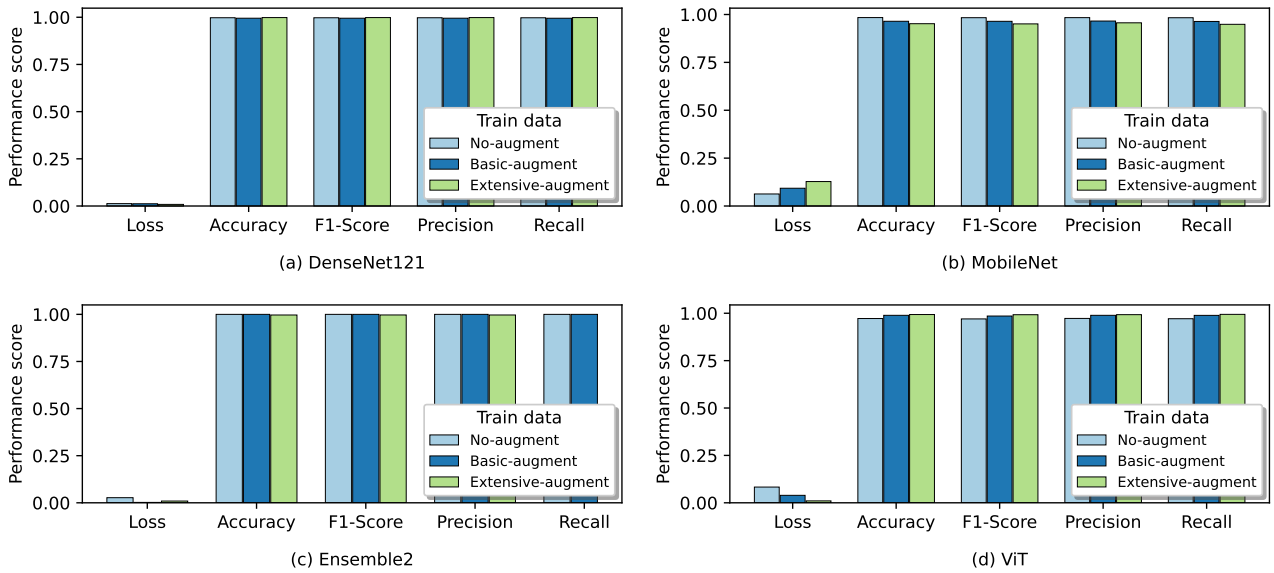


Figure 4.28: For the Rice Leaf Disease dataset, visualisation of the selected classification models' test performance comparison three augmentation intensities.

Training time comparison across three augmentation intensities

Figure 4.29 demonstrates that ViT required the longest training time across all augmentation variations. Notably, with extensive augmentations, ViT took about 6 hours which is around 3 to 10 times longer than the other models in the study. On the other hand, DenseNet121 exhibited the shortest training time, when trained without augmentation, while Ensemble2 needed the least training time for both basic and extensive augmentations.

Validation and test performance comparison across three augmentation intensities

The performance of the chosen four classification models with three augmentation intensities (none, basic, and extensive) on the training samples is detailed in Table 4.7, Table 4.8, and Table 4.9 respectively and a visual representation of the same data with a bar-chart format has been presented in Figure 4.30, Figure 4.31, and Figure 4.32 for better comprehension.

Validation results: Validation results obtained using the Rice Leaf Disease dataset [65] over three augmentation variations have been presented below.

- **Without augmentation:** From Table 4.7 and Figure 4.30, we can observe that ViT had the highest validation loss at approximately 13%, while DenseNet121 had the lowest with nearly 1.2%. Ensemble2 had the highest validation performance in accuracy and F1-score with perfect scores, while ViT had the lowest with approximately 97%. MobileNet required the longest time (23 epochs) to reach the best validation F1-score, while Ensemble2 needed the shortest (4 epochs).

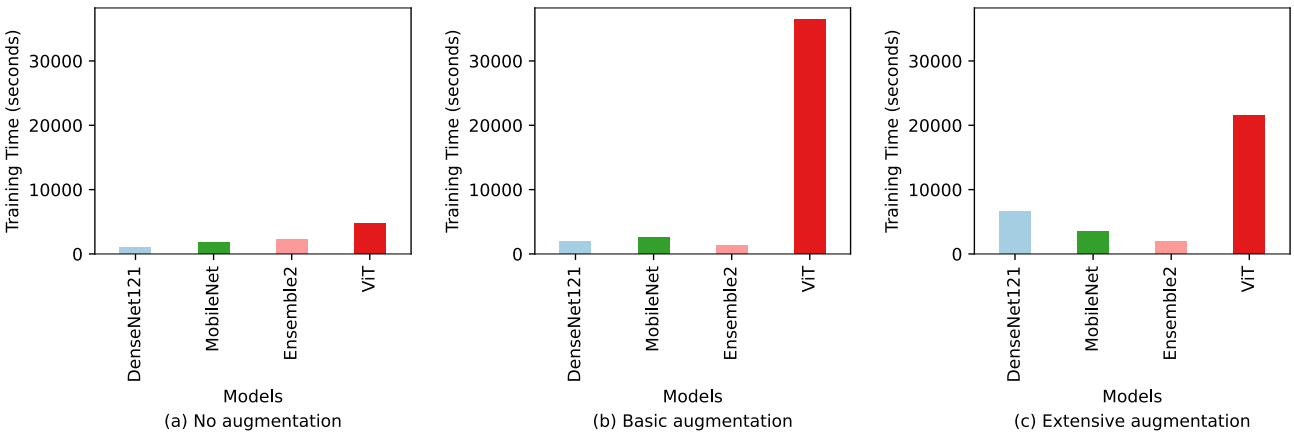


Figure 4.29: For the Rice Leaf Disease dataset, training time comparison of the selected classification models across three augmentation intensities.

Table 4.7: For the Rice Leaf Disease dataset, performance comparison of the selected classification models without augmentation.

Model name	Test loss	Test accuracy	Test F1-score	Test precision	Test recall	Val. loss	Best val. accuracy	Best val. F1-score	Best val. F1-score at epoch	Train time (seconds)
Dense Net121	**0.012620**	0.997473	0.997348	0.997417	0.997296	**0.0131**	0.9972	0.9972	18	**951.78814**
Mo- bileNet	0.062872	0.983993	0.983354	0.983660	0.983150	0.0498	0.9894	0.9890	**23**	1788.90537
Ensem- ble2	0.027141	**1**	**1**	**1**	**1**	0.0224	**1**	**1**	**4**	2187.42996
ViT	**0.083006**	**0.972199**	**0.970084**	**0.972711**	**0.971163**	**0.1344**	**0.9704**	**0.9694**	22	**4695.49538**

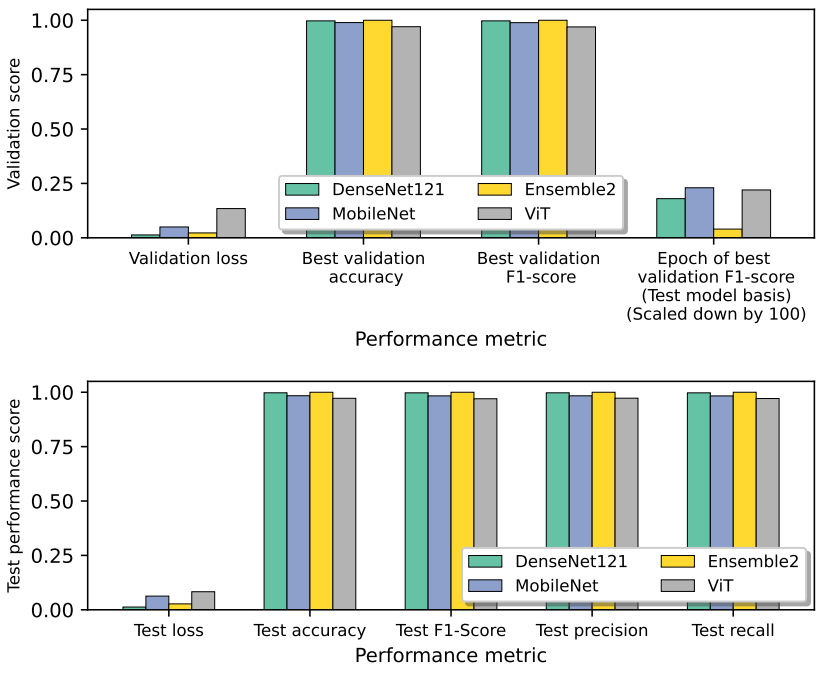


Figure 4.30: For the Rice Leaf Disease dataset, visual representation of the performance comparison of the selected classification models without augmentation.

Table 4.8: For the Rice Leaf Disease dataset, performance comparison of the selected classification models with basic augmentation.

Model name	Test loss	Test accuracy	Test F1-score	Test precision	Test recall	Val. loss	Best val. accuracy	Best val. F1-score	Best val. F1-score at epoch	Train time (seconds)
Dense Net121	0.011892	0.994945	0.99491	0.99503	0.99481	**0.0013**	**1**	**1**	21	1933.472600
Mo-bileNet	**0.092780**	**0.964617**	**0.964420**	**0.965840**	**0.963690**	**0.0505**	**0.9838**	**0.9841**	78	2619.94800
Ensem-ble2	**0.002129**	**1**	**1**	**1**	**1**	0.0032	**1**	**1**	**14**	**1251.50090**
ViT	0.039388	0.989048	0.985290	0.989050	0.988790	0.0447	0.9859	0.9861	**108**	**36412.30030**

- **With basic augmentation:** From Table 4.8 and Figure 4.31, we can observe that MobileNet had the highest validation loss with approximately 5%, while DenseNet121 had the lowest with nearly 0%. Regarding validation accuracy and F1-score, Ensemble2 and DenseNet121 achieved the highest with a perfect score (1), whereas MobileNet had the lowest with approximately 98%. Additionally, ViT took the highest number of epochs (108) to achieve the best validation F1-score, while Ensemble2 required the lowest (14).

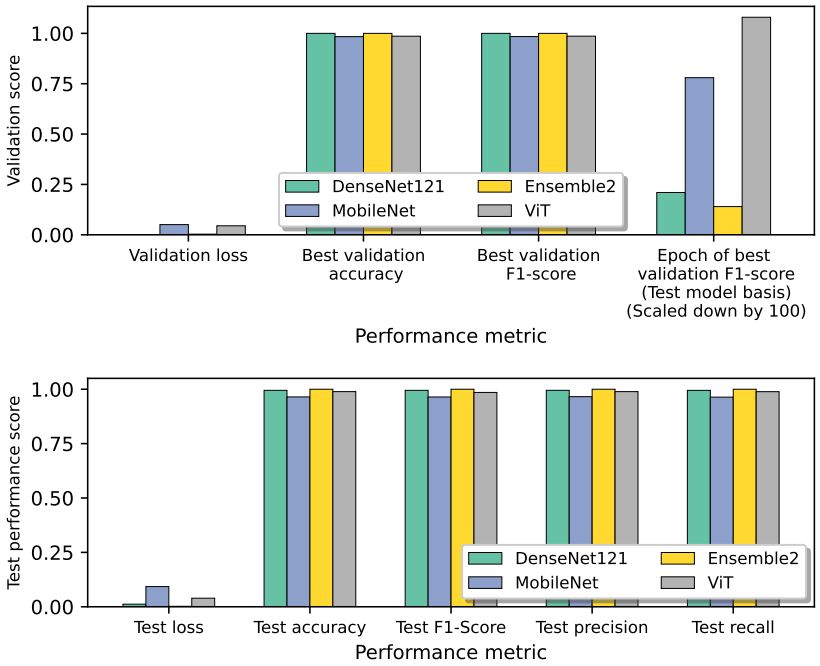


Figure 4.31: For the Rice Leaf Disease dataset, visual representation of the performance comparison of the selected classification models with basic augmentation.

- **With extensive augmentation:** From Table 4.9 and Figure 4.32, we can observe that MobileNet had the highest validation loss with approximately 5.19%, while ViT had the lowest with around 0.53%. Regarding validation accuracy and F1-score, DenseNet121 and Ensemble2 achieved the highest with a perfect score (1), while MobileNet had the lowest with approximately 98%. Additionally, MobileNet took the highest number of epochs (111) to achieve the best validation F1-score, while DenseNet121 required the lowest (27).

Model learning curves: Model learning curves for the selected classification models using the Rice Leaf Disease dataset over three augmentation variations have been provided below.

- **Without augmentation:** The chosen classification models' learning curves visualising the validation

Table 4.9: For the Rice Leaf Disease dataset, performance comparison of the selected classification models with extensive augmentation.

Model name	Test loss	Test accuracy	Test F1-score	Test precision	Test recall	Val. loss	Best val. accuracy	Best val. F1-score	Best val. F1-score at epoch	Train time (seconds)
Dense Net121	**0.008551**	**0.998315**	**0.998322**	**0.998313**	**0.998334**	0.0057	**1**	**1**	**27**	6633.77760
Mo-bileNet	**0.127715**	**0.951980**	**0.951000**	**0.956646**	**0.948761**	**0.0519**	**0.9817**	**0.9821**	**111**	3444.94039
Ensem-ble2	0.009337	0.996630	0.996813	0.996682	0.996956	0.0066	**1**	**1**	31	**1976.32826**
ViT	0.010119	0.993260	0.992290	0.992565	0.994135	**0.0053**	0.9993	0.9993	68	**21475.89383**

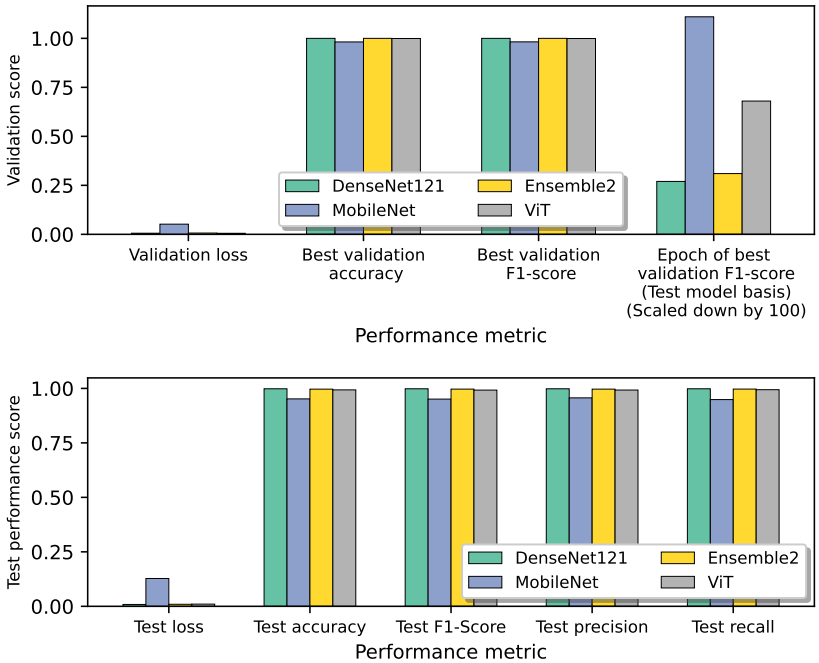


Figure 4.32: For the Rice Leaf Disease dataset, visual representation of the performance comparison of the selected classification models with extensive augmentation.

performance progressing throughout the total number of epochs has been illustrated in Figure 4.33. From Figure 4.33, we can observe that none of the models had overfitting. Figure 4.33 (a) and (c) indicate that DenseNet121 and MobileNet had initial instability within the first 10 and 15 epochs respectively, but stabilised after this period. Figure 4.33 (g) reveals that ViT had a slight gap between the training and validation results. Figure 4.33 (e) demonstrates that Emsemble2 exhibited good generalisation.

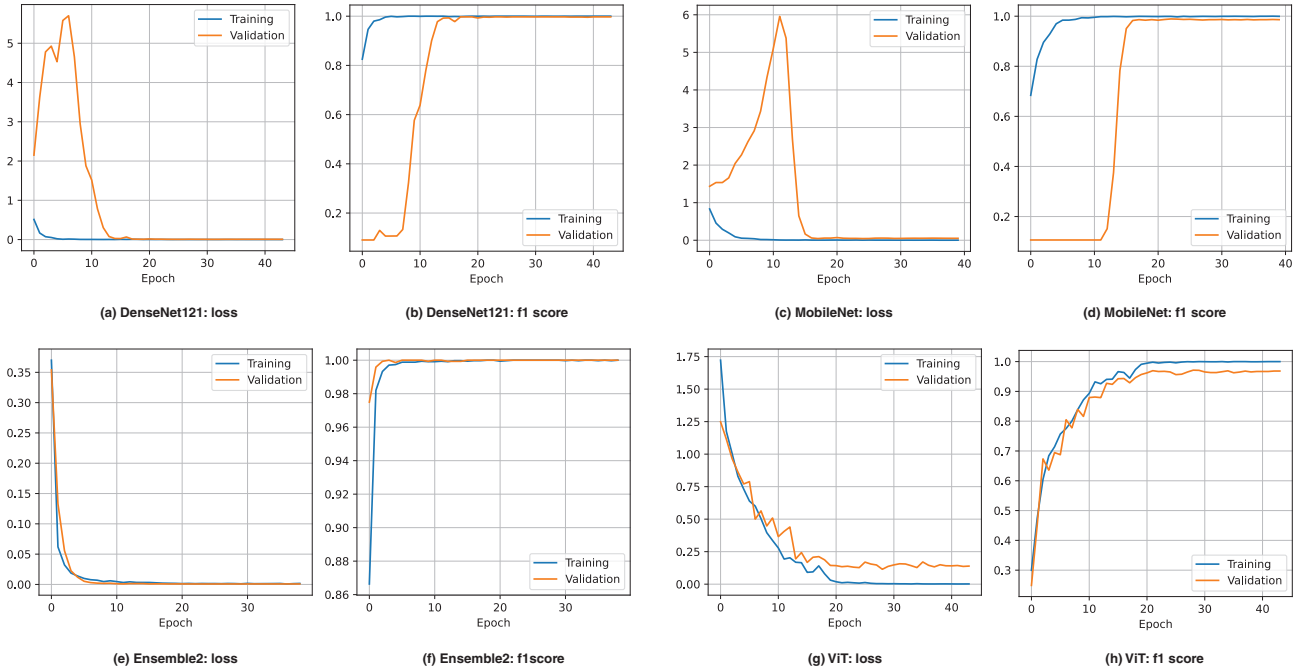


Figure 4.33: For the Rice Leaf Disease dataset without augmentation on train data, visualising the learning curves for the selected classification models visualising the validation performance over all epochs.

- **With basic augmentation:** The chosen classification models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.34. We can observe that none of the models had overfitting. Similar to training without augmentation, Figure 4.34 (a) and (c) indicate that DenseNet121 and MobileNet had initial instability within the first 12 and 10 epochs respectively, but stabilised after this period. Figure 4.34 (e) demonstrates that Emsemble2 exhibited good generalisation.

- **With extensive augmentation:** The chosen classification models' learning curves visualising the validation performance progressing throughout the total number of epochs have been illustrated in Figure 4.35. From Figure 4.35, we can observe that none of the models had overfitting. Figure 4.35 (a) and (c) indicate that DenseNet121 and MobileNet had initial instability within the first 20 and 30 epochs respectively, but stabilised after this period. Figures 4.35 (e) and (g) demonstrate that the Emsemble2 and ViT exhibited good generalisation.

Test results: Test results obtained using the Rice Leaf Disease dataset [65] dataset have been presented below.

- **Without augmentation:** From Table 4.7 and Figure 4.30, we can observe that for the test loss, DenseNet121 had the lowest with 1.2% and ViT had the highest with approximately 8%. Additionally, Ensemble2 achieved the highest and perfect performance scores (1) in all the performance measuring metrics, while ViT exhibited the lowest in accuracy, F1-score, precision, and recall with 97.22%, 97.01%, 97.27%, and 97.12% respectively. The rest of the models achieved good performance, ranging from 98% to 99% for all the performance metrics.

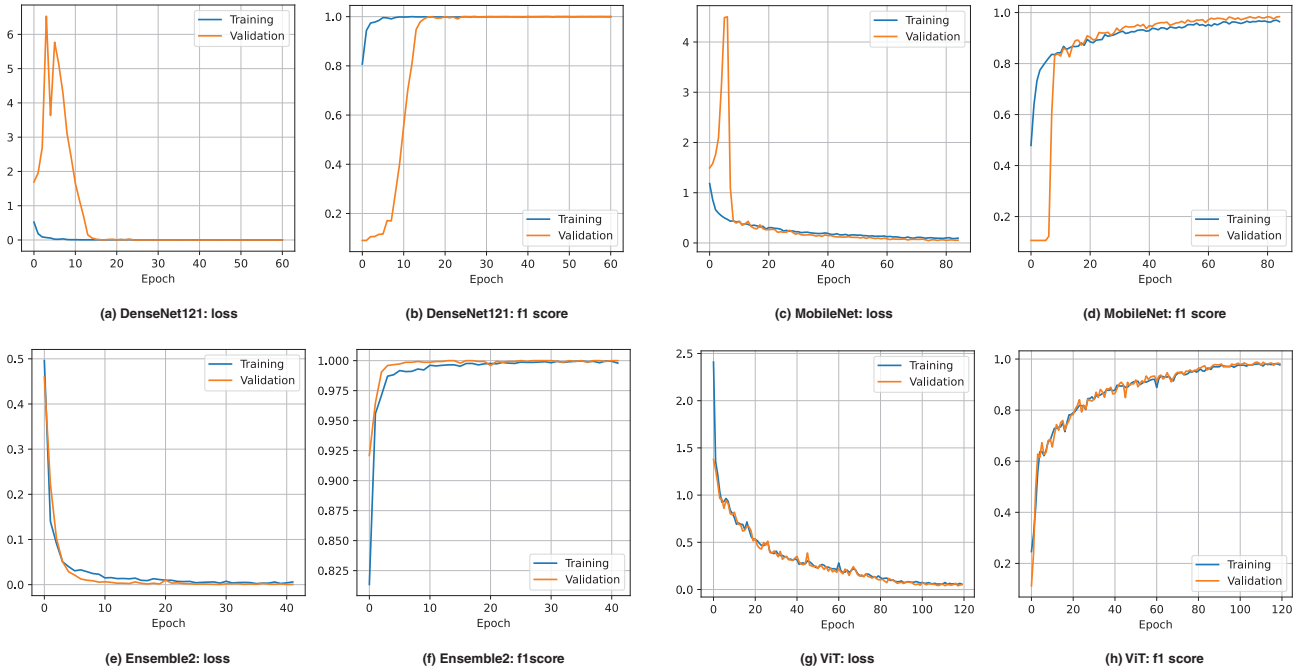


Figure 4.34: For the Rice Leaf Disease dataset with basic augmentation on train data, visualising the learning curves for the selected classification models visualising the validation performance over all epochs.

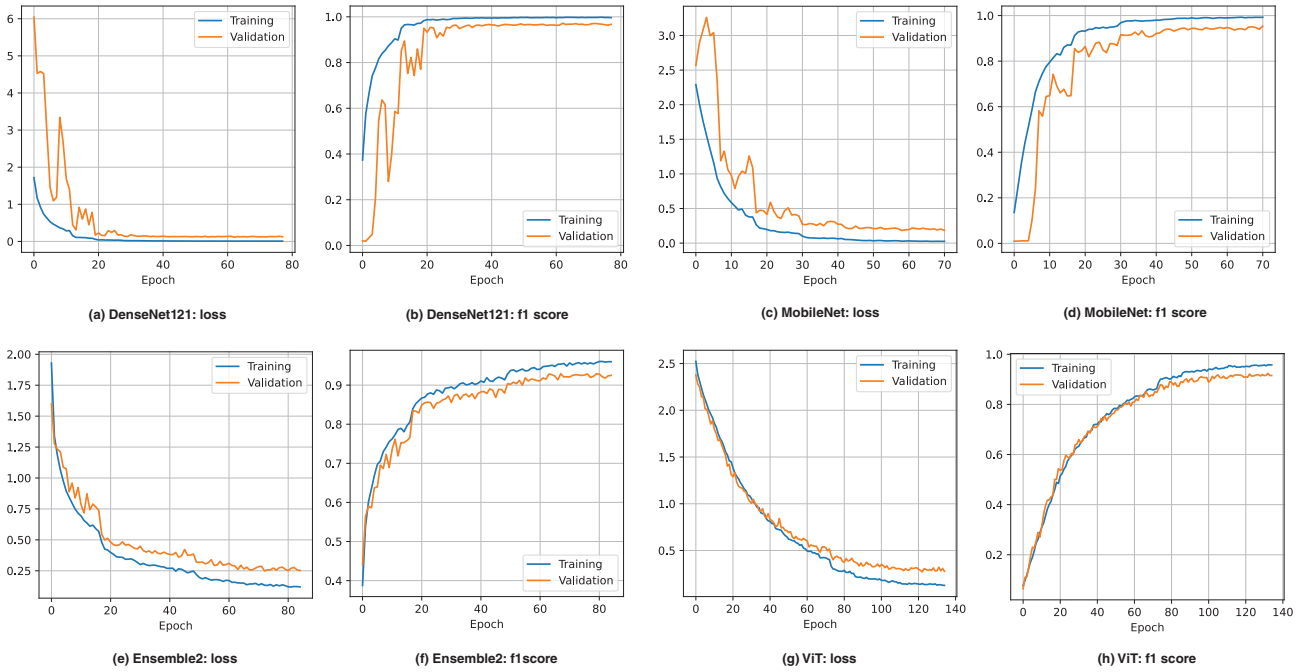


Figure 4.35: For the Rice Leaf Disease dataset without augmentation on train data, visualising the learning curves for the selected classification models visualising the validation performance over all epochs.

- **With basic augmentation:** From Table 4.8 and Figure 4.31, we can observe that for the test loss, Ensemble2 had the lowest with approximately 0.2% and MobileNet had the highest with just over 9%. Additionally, Ensemble2 achieved the highest and perfect performance scores (1) in all the performance measuring metrics, while MobileNet exhibited the lowest in accuracy, F1-score, precision, and recall with 96.46%, 96.44%, 96.58%, and 96.37% respectively. The rest of the models achieved good performance, ranging from 98% to 99% for all the performance metrics.
- **With extensive augmentation:** From Table 4.9 and Figure 4.32, we can observe that for the test loss, MobileNet had the highest with nearly 13% and DenseNet121 had the lowest with approximately 0.8%. Additionally, similar to the performance with basic augmentation, MobileNet exhibited the lowest in accuracy, F1-score, precision, and recall with 95.2%, 95.1%, 95.66%, and 94.88% respectively, while DenseNet121 achieved the highest with nearly perfect performance scores (1) in all the performance measuring metrics. The rest of the models achieved good performance with approximately 99% for all the performance metrics. Moreover, ViT had a nearly comparable performance to DenseNet121 and Ensemble2.

Confusion matrices: Confusion matrices for the selected classification models using the Rice Leaf Disease dataset over three augmentation variations have been presented below to help understand the reason behind misclassifications and give an overview of the models' performance.

- **Without augmentation:** Figure 4.36 presents confusion matrices for classification models trained without augmentation. This figure demonstrates that while most classes were accurately classified by all models, there were notable exceptions. Specifically, MobileNet frequently confused Blast with Tungro.
- **With basic augmentation:** Figure 4.37 presents confusion matrices for classification models trained with basic augmentation. Similar to results without augmentation, all the models accurately classified most classes with a few exceptions. Specifically, MobileNet often misidentified Brown spot and Bacterial blight as Blast.
- **With extensive augmentation:** Figure 4.38 presents confusion matrices for models trained with extensive augmentation. Similar to results without augmentation and basic augmentations, all the models accurately classified most classes with a few exceptions. Notably, MobileNet frequently misidentified Blast as Tungro and Brown spot as Blast.

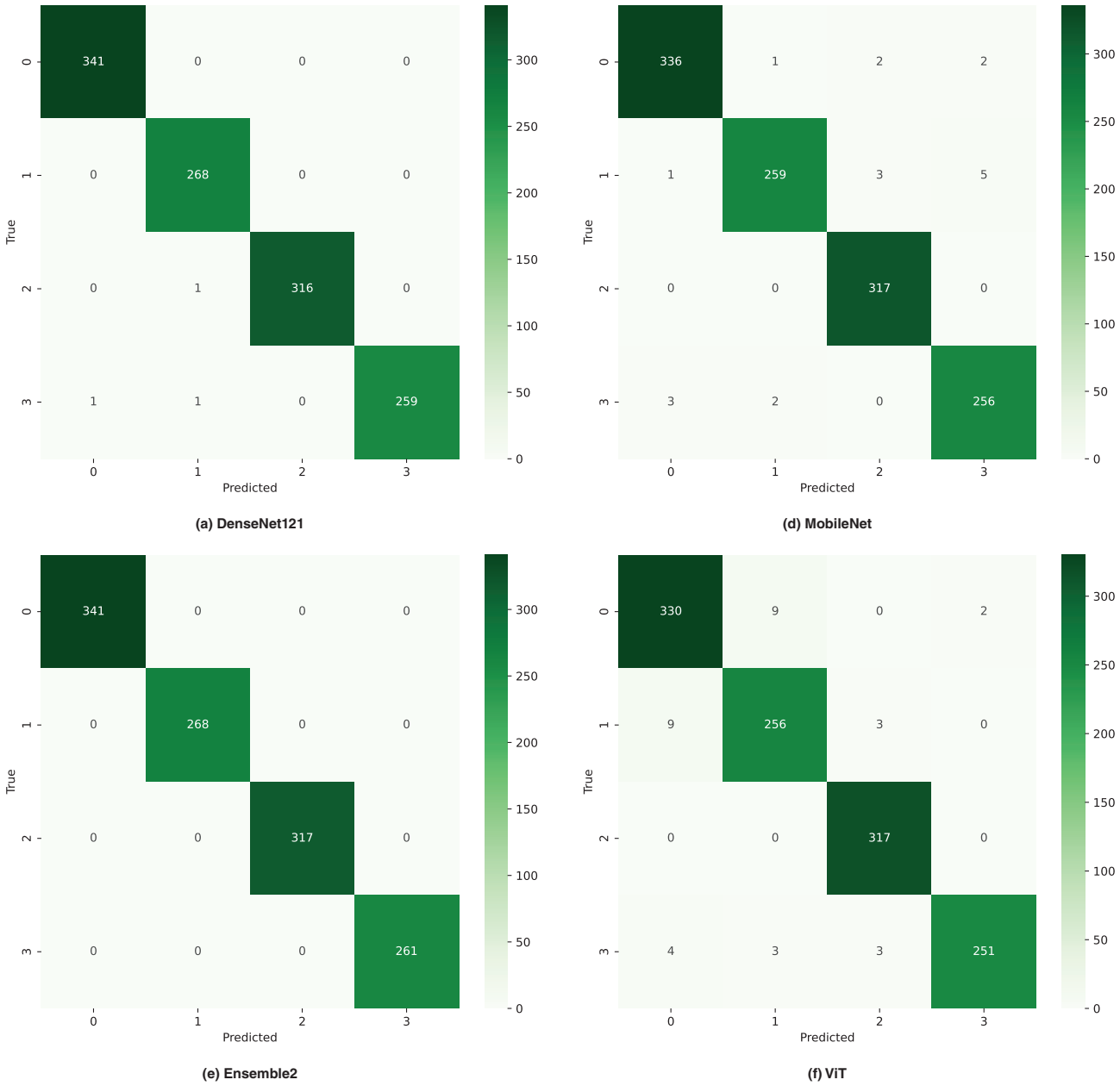


Figure 4.36: Visualising the confusion matrices for the selected classification models on the Rice Leaf Disease test dataset without augmentation. For the label of the confusion matrices, 0 represents Bacterial blight, 1 represents Blast, 2 represents Brown spot, and 3 represents Tungro.

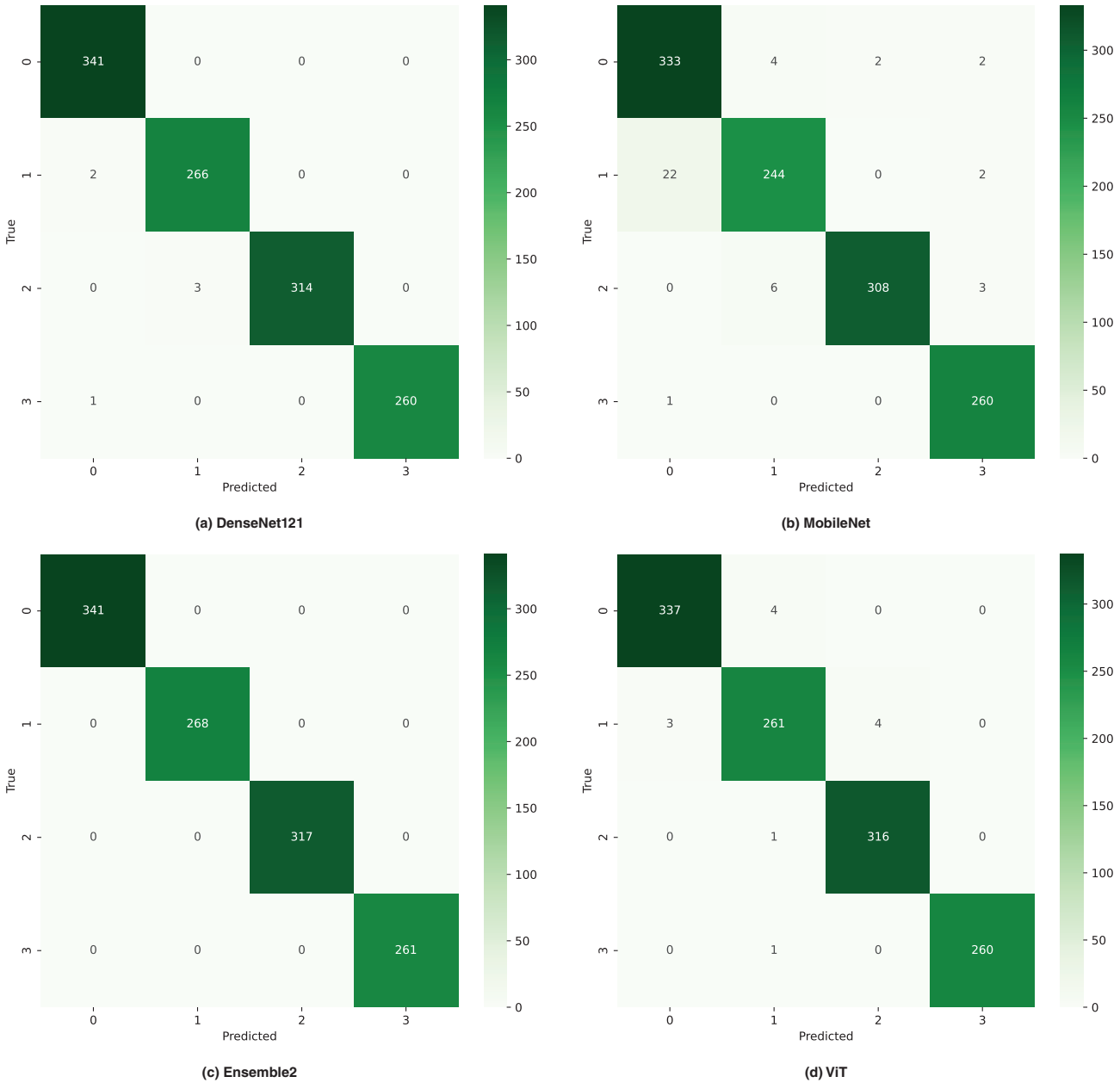


Figure 4.37: Visualising the confusion matrices for the selected classification models on the Rice Leaf Disease test dataset with basic augmentation. For the label of the confusion matrices, 0 represents Bacterial blight, 1 represents Blast, 2 represents Brown spot, and 3 represents Tungro.

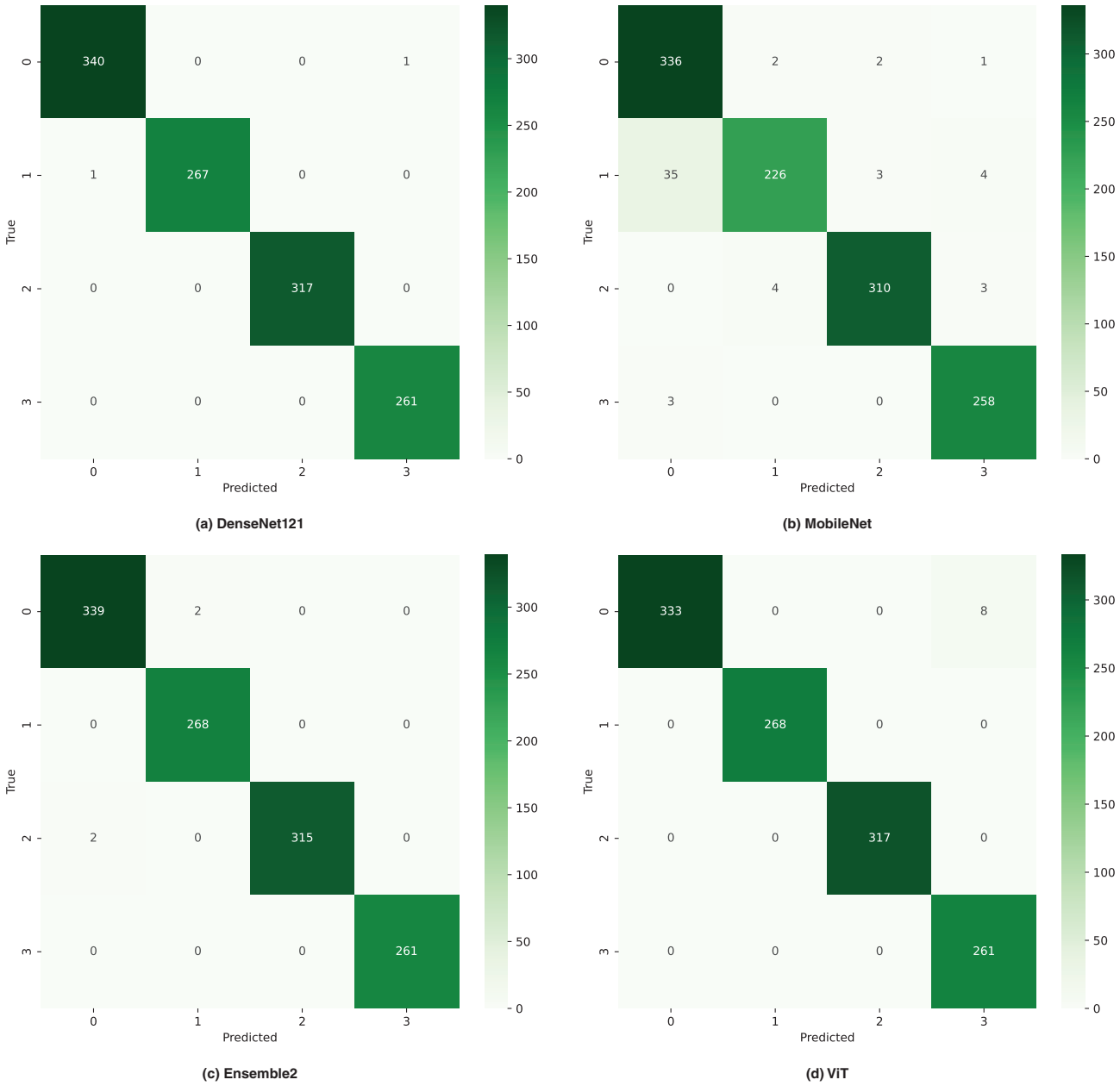


Figure 4.38: Visualising the confusion matrices for the selected classification models on the Rice Leaf Disease test dataset with extensive augmentation. For the label of the confusion matrices, 0 represents Bacterial blight, 1 represents Blast, 2 represents Brown spot, and 3 represents Tungro.

5 Discussions

This chapter provides discussions focused on answering the research questions and briefly covers our observations from the results. Since our research belonged to two different tasks related to paddy disease diagnosis: (1) segmentation and (2) classification, this discussion chapter also has two sections covering the individual tasks. A detailed discussion to answer the research questions related to the paddy disease diagnosis is given below. This chapter also includes a detailed description of the limitations faced by this research work.

5.1 Discussion of segmentation task

Classifying paddy diseases presents significant challenges due to the complex field environment and similarities in disease patterns. The usage of segmentation resolves this issue by isolating the diseased areas/ROI from healthy leaf parts [51]. To increase the efficiency of paddy disease diagnosis, the segmentation can be used as a pre-processing step or incorporated into the classification model which can be observed in various crop studies [21], [28], [41], [51]–[54]. To address the lack of an open-access paddy disease segmentation dataset, we developed one to evaluate the effectiveness of existing segmentation models in paddy disease. The research questions for the segmentation task have been restated below.

- RQ1: Which mask-based segmentation model achieves the highest and lowest performance scores in intersection over union (IoU)?
- RQ2: Considering all parameters including model size, and qualitative and quantitative test performance, which mask-based segmentation model is the most suitable to be used as a preprocessing step or incorporated within the classification process?
- RQ3: If the number of augmented images is higher (more than double) than the original number of samples, does it have any negative or positive effect on the test performance using unmodified test samples (both qualitative and quantitative)?

5.1.1 Analysing research question 1 (RQ1)

IoU compares pixel-level similarity between the ground truth and predicted masks [21], [32]. Therefore, it is significantly important when assessing a model's performance. Below is a performance comparison of the selected models across three train data sizes, which will help to identify the models with the highest and lowest IoU scores.

- **Highest IoU test performance:** When trained with 616 samples, Deep Residual UNet and VGG16 UNet achieved the highest and the second-highest IoU scores at 82.42% and 81.65% respectively. With 1500 training samples, the highest IoU was achieved by Deep Residual UNet at 86.04% and the second highest IoU score was recorded by VGG16 UNet at 84.45%. With 2500 training samples, Deep Residual UNet and TransUNet showed the highest and second-highest IoU scores at 88.77% and 87.53% respectively. Overall,

Deep Residual UNet consistently demonstrated superior IoU performance across varying train data sizes. The combined usage of residual connections, skip connections, and encoder-decoder architecture helped the Deep Residual UNet to effectively identify complex paddy disease patterns [29]. Hence, Deep Residual UNet has shown the highest suitability to be used in the disease diagnosis system based on IoU.

- **Lowest IoU test performance:** When trained with 616 samples, UNet and TransUNet recorded the lowest and second-lowest IoU scores, at 80.61 and 81.59% respectively. With 1500 train samples, TransUNet and UNet again showed the lowest and second-lowest scores, at 82.58% and 83.05% respectively. With 2500 train samples, UNet and VGG16 UNet scored the lowest and second-lowest IoU scores, at 86.74% and 86.85% respectively. Hence, based on the results, it can be concluded that in most cases, the UNet and the TransUnet exhibited the lowest IoU performance. Due to the incorporation of ViT, TransUNet is a large model and needs significantly more training data to perform effectively [57]. However, our maximum train dataset size was 2500, which is insufficient compared to larger datasets such as ImageNet or JFT-300M. Thus, TransUNet suffers from overfitting and has high test loss. On the other hand, UNet struggles with capturing long-range dependencies and global context [55] and is deprived of advanced features such as attention heads [30], residual connections [29], [32], Respaths [32], Resblocks [32], and deep residual blocks [29] or pre-trained knowledge. Therefore, UNet has shown low performance in IoU. Lastly, we can say that UNet and TransUNet showed the lowest suitability to be used in the disease diagnosis system based on IoU.

5.1.2 Analysing research question 2 (RQ2)

Selecting a segmentation model for preprocessing or integration within a classification system involves consideration of several factors, assessing only the quantitative performance (test and validation) is not enough. It is crucial to assess the quality of segmentation results and consider the model size, and its demand for computational resources, especially for applications intended for resource-constraint environments such as mobile applications used by farmers. Despite the high quantitative performance, some models might fail to capture essential fine details such as the small brown spots which are crucial for identifying paddy Brown spot disease [1], [4]. Thus, selecting a segmentation model should be a comprehensive process that considers all the relevant factors, including quantitative and qualitative performance, model size, and computational resource requirements. Structural analysis and the qualitative and quantitative performance analyses of the selected segmentation models are given below.

- **Structural analysis:** From the analysis presented in Table 2.2, it is evident that UNet has the smallest model size at only 8.23 MB, which is 3 to 47 times smaller than other models used in this study. It also has the least amount of trainable and total parameters among the models reviewed. In comparison, the other models possess approximately 4 to 44 times more trainable parameters than UNet. Therefore, in scenarios with strict resource constraints, performance might need to be compromised to ensure compatibility. Additionally, Deep Residual UNet, an UNet variant is enhanced with deep residual connections and exhibited lower total parameters and smaller model sizes after UNet in this study. Therefore, when resource constraints are minimal, Deep Residual UNet could be a suitable choice.

- **Qualitative performance analysis:**

  - **616 train data:** From the qualitative assessment using 616 training samples in the results chapter 4.1.2, it is clear that Deep Residual UNet showed superior performance, as it was the only model successful in detecting small circular disease patterns, which are very important for paddy disease diagnosis. The usage of multiple residual connections in Deep Residual UNet overcame the vanishing gradient issue common in deep networks and the usage of skip connections facilitated the transfer of fine details from earlier layers to the later layers which enhanced the model performance in disease detection [29].

  - **1500 train data:** From the qualitative assessment using 1500 training samples in results chapter 4.1.2, we can see that both TransUNet and Deep Residual UNet demonstrated excellent performance in detecting small scattered circular disease patterns and diseases at the leaf edges. TransUNet's integration of CNN and ViT extracts high-resolution spatial information and global context respectively and their combination effectively captures the paddy disease patterns [55]. On the other hand, Deep Residual UNet succeeded due to the aforementioned reasons.
  - **2500 train data:** From the qualitative analysis with 2500 training data in results chapter 4.1.2, we can observe that similar to the qualitative assessment using 1500 training samples, TransUNet and Deep Residual UNet demonstrated superior performance in detecting small circular disease patterns and diseases on leaf edges. Their enhanced performance has been possible due to the above-mentioned reasons.

  From the above analysis, we can conclude that across various train data sizes, Deep Residual UNet performs the best considering only the qualitative results and is a suitable candidate for selection.

- **Quantitative performance analysis:** In the segmentation tasks, the ROI usually covers a small number of pixels compared to the total number of pixels in the entire image. Hence precision, recall, F1-score, and accuracy might not always provide the correct perception of the performance and show superior performance due to correctly classifying background pixels [32]. Therefore, for this study, only the IoU should be considered when selecting models based on quantitative performance. From the analysis of research question 01 (RQ1) presented in section 5.1.1, we have concluded that in terms of IoU performance across varying train data sizes, Deep Residual UNet consistently demonstrated superiority and can be considered as a candidate for selection.

With a more holistic view, considering quantitative and qualitative performance and architectural benefits and limitations of the models in this study, we can conclude that Deep Residual UNet is the most ideal choice to be used as a preprocessing step or included within the disease classification model. Especially for scenarios where slightly better computational resources are available and higher performance metrics are needed without significant compromises on efficiency. It has a moderate model size of 31.41 MB and enhanced disease-detection capabilities through the usage of both residual and skip connectors, potentially leading to more refined segmentation outputs when trained with sufficient data. The residual connectors utilised in their custom deep residual blocks enable deep network training without gradient vanishing and the passing of detailed spatial information in the decoder from the encoder through skip connections.

5.1.3 Analysing research question 3 (RQ3)

According to Wang et al. [81], data augmentation significantly enhances the performance of classification models. In this study, we want to observe the impact of excessive data augmentation for the pixel-level classification/segmentation task. For this, we have expanded the training data size through data augmentation from 616 to 1500 and 2500, which correspond to approximately 2.5 and 4 times the original dataset respectively.

- **Quantitative performance analysis:** From Figure 4.1, we can observe that these increased train data have significantly improved the quantitative test performance in all performance evaluation metrics across most models. Among them, drastic improvements were mostly noticeable in IoU and precision. However, when trained with 2500 samples, which is about four times the original dataset size, performance declination was observed in UNet's recall, and VGG16 UNet and TransUNet's precision. These exceptions might be due to the models' sensitivity to augmentation noise. For the test loss, as the number of training data increased and the model's performance improved, the test loss for all the models decreased. Exceptions were seen in TransUNet's test loss when trained with 2500 train data, which might be again due to the model's sensitivity to augmentation noise.

- **Qualitative performance analysis:** Figures 4.11, 4.13, and 4.15 demonstrate that the quality of circular disease pattern detection improves as the training data size increases, as seen in sample 3. However, the detection of diseases at the edges showed degradation with increased train data in a few samples. For increased train data (1500 and 2500), the diseases at the edges are visible in sample 3, but missing in sample 1.

Overall, a higher number of augmented train images has a positive impact on both the qualitative and quantitative results.

5.1.4 Additional observations from segmentation results

From the segmentation results, additional observations have been made in terms of test loss, disease misclassification count, and the influence of the LR scheduler on the overall model performances and the details are given below.

- **Test loss:** From Table 4.1 and Figure 4.1, it is evident that TransUNet exhibited a significantly larger test loss when trained with 616 data samples compared to other models in this study. When trained with 616 train data, the test loss for this model is approximately 28%, which is approximately 7% higher than the other models. The excessive data demand of the ViT model within TransUNet architecture is responsible for its high test loss [55], [57]. As the training data increased (1500 and 2500), the test loss significantly reduced from 28% to between 3-5%. Moreover, a test loss of approximately 5% with larger train data sizes is considered reasonable.

- **Disease pixel misclassification:** Deep Residual UNet achieved the lowest disease pixel misclassification when trained with 616 data and TransUNet achieved the lowest disease pixel misclassification when trained with 1500 and 2500 data. TransUNet's hybrid CNN and transformer structure helped it to capture fine disease details efficiently, which led to the lowest misclassification [55]. On the other hand, the presence of deep residual blocks present in the Deep Residual UNet might be the reason for its enhanced performance and low disease misclassification.

  UNet achieved the highest disease pixel misclassification when trained with 616 and 1500 data and VGG16 UNet achieved the highest disease pixel misclassification when trained with 2500 data. The limitations of Unet in capturing long-range dependencies and global context [55] and absence of advanced features such as residual connections [29] and deep residual blocks [29], or pre-trained knowledge compared to the other models in this study, might be the reason for its high disease misclassification. Additionally, the high disease misclassifications of VGG16 UNet also might be due to the augmentation noise sensitivity of the VGG16-based encoder.

- **Learning rate scheduler impact:** To get insight into the robustness and scalability of each model, identical training parameters were maintained across three train dataset sizes. VGG16 UNet and Deep Residual UNet showed initial instability during the first few epochs, which was quickly stabilised due to the LR scheduler. LR scheduler also aided in reaching convergence and overcoming plateaus without leading to overfitting which can be validated by observing the nearly equal validation and test performance presented in Table 4.1, Table 4.2, and Table 4.3. Among all the selected models, UNet displayed early overfitting signs, particularly with smaller train datasets. Its lack of advanced features compared to other models in the research and its simpler architecture are probably the causes of this.

This comprehensive analysis evaluates the performance of four segmentation models across three train data sizes and assesses their effectiveness in classifying paddy diseases at the pixel level. This study emphasises the need for a comprehensive evaluation process when selecting segmentation models for preprocessing or integration into classification systems. It highlights the importance of considering both quantitative and qualitative metrics as well as computational demands to ensure optimal performance in practical applications.

5.2 Discussion of classification task

Paddy disease diagnosis is a complex task, whether conducted manually or with software. Manual identification of paddy diseases is limited by time constraints, inaccuracies, and biases from the assessor, while software-based methods face challenges such as identifying small symptoms in noisy images, handling disease pattern similarities, creating masks through image processing, generalising across datasets (both same-dataset and cross-dataset), and managing high computational requirements and costs, and occlusion and visibility issues [3]–[9], [11], [12], [27], [41], [73], [82], [83]. The research questions for the classification task are restated below.

- RQ4: For paddy disease classification with complex field data, can the performance of ViT be achieved with traditional CNN models (MobileNet, and DensNet121) or an ensemble of the traditional models?
- RQ5: How does the level of data augmentation (none, basic, and extensive) impact the performance of models classifying paddy diseases using datasets with complex field environments?

5.2.1 Analysing research question 4 (RQ4)

The results obtained from two paddy disease datasets, as described in Chapter 4.2, will be used for this discussion. Between these two datasets, Paddy Doctor [16] is comparatively larger and has 16225 samples belonging to 13 classes and the Rice Leaf Disease Dataset [65] has 5932 samples belonging to 4 classes. An analysis of each dataset is given below.

- **Analysis of Paddy Doctor dataset results:** For model training without augmentation, ViT performed poorly compared to the other models in all performance measuring metrics (accuracy, precision, recall and F1-score). This has happened due to the inherent data-hungry nature of ViTs and the used datasets without augmentation are insufficient for training [57]. Therefore, ViT exhibited the largest test loss with nearly 50%. For the basic augmentation, ViT had almost comparable performance to DenseNet121 and MobileNet in all performance measuring metrics, while DenseNet121 had the best performance in all metrics. The basic augmentation mitigated the issues related to a small dataset to some extent, which resulted in a decrease in the test loss from approximately 50% to almost 38%. For extensive augmentation, the test loss of ViT further decreased to approximately 17% and gave a comparable performance to DenseNet121 in accuracy, precision and recall, which is the best-performing model for this augmentation. The increase in augmentation intensity expanded the training dataset size, which improved the ViT performance. On the other hand, with extensive augmentation, the performance of MobileNet and Ensemble2 declined, which might be due to the high augmentation noise present in the data and the sensitivity of these models to augmentation noise.
- **Analysis of Rice Leaf Disease dataset results:** Similar to the results of the Paddy Doctor dataset, for model training without augmentation, ViT had the largest test loss with nearly 8% and the lowest performance compared to the other models in all performance measuring metrics (accuracy, precision, recall, and F1-score) for the previously mentioned reasons. Furthermore, for the basic augmentation, ViT had almost comparable performance to DenseNet121 in all performance measuring metrics, while Emsemble2 had the best performance in all metrics. The basic augmentation mitigated the issues related to a small dataset to some extent, which resulted in a decrease in the test loss from approximately 8% to almost 4%. Additionally, in the case of extensive augmentation, the test loss of ViT further decreased to approximately 1% and gave a comparable performance to DenseNet121 and Ensemble2. For extensive augmentation, DenseNet121 was the best-performing model. ViT's improved performance is due to the same reasons previously mentioned for the Paddy Doctor dataset.

From the above analysis of the results for the two datasets, we can conclude that the performance of ViT is usually poor when it is trained without any augmentation. With basic augmentation, ViT demonstrated

increased and comparable performance with traditional models, DenseNet121 (Paddy Doctor and Rice Leaf Disease Dataset) and MobileNet (Paddy Doctor). With extensive augmentation, the performance of ViT improved further and showed comparable performance with traditional models, DenseNet121 (Paddy Doctor and Rice Leaf Disease Dataset) and Ensemble2 (Rice Leaf Disease Dataset). Considering the high training time needed for ViT models as seen in Table 4.4, Table 4.5, Table 4.6, Table 4.7, Table 4.8, and Table 4.9, it is better to use traditional models and focus on improving the model performance with additional features rather than relying on computation heavy ViTs.

5.2.2 Analysing research question 5 (RQ5)

From Figure 4.17 and Figure 4.28, we can observe that as the augmentation intensity increased, the performance of ViT improved and the test loss decreased. ViT's dependency on large datasets for achieving good performance explains this behaviour [57]. On the other hand, MobileNet performed well without any augmentation and as the augmentation intensity increased, its performance decreased. Since MobileNet is a small model with a less complex structure, it is highly susceptible to overfitting and this low performance might be due to this reason. But Figures 4.22, 4.23, 4.24, 4.33, 4.34, and 4.35 confirm that there was no overfitting. Therefore, the low performance could be due to MobileNet's simplistic network architecture and its sensitivity to the augmentation noise.

For the Rice Leaf Disease Dataset, Densenet121 and Ensemble2 demonstrated nearly equal performance for all augmentation intensities. But for the Paddy doctor dataset with basic augmentation, DenseNet121 had the lowest performance, while Ensemble2 had the highest compared to the other models in this study. Since we were unable to determine the cause of this behaviour, we decided to include it as part of our future work.

5.2.3 Additional observations from classification results

From the results, additional observations have been made in terms of disease misclassifications and the influence of the LR scheduler on the overall model performances and the details are given below.

- **Disease misclassifications:** From the confusion matrices presented in section 4.2.1, it has been observed that several diseases have been misclassified as normal plants. It might be because, as observed in Figure 3.10, the Paddy Doctor dataset is heavily imbalanced, with normal plants having the largest portion of samples (14%). Therefore, the models frequently misclassified diseases as normal plants. Additionally, White stem borer and Yellow stem borer are both from the same disease family and have overlapping and similar symptoms as described in section 2.1.4. Hence, ViT often misclassified White stem borer as Yellow stem borer.

  Even though the Rice Leaf Disease Dataset having four classes is nearly balanced as seen in Figure 3.11, Tungro (22%) and Blast (24%) are the minority classes compared to the other two classes (27%). Furthermore, the confusion matrices presented in section 4.2.2 show that MobileNet had the highest amount of misclassifications among the other classification models and it frequently misclassified Brown spot and Bacterial blight as Blast, and Blast as Tungro. According to the analysis presented in Table 2.1, Brown spots, Bacterial blight, and Blast show similar disease patterns such as dark brown lesions, olive spots, and dark brownish-black spots on leaves respectively, while Blast and Tungro show diamond-shaped lesions and stripes in on leaves respectively. Due to the aforementioned similarities, these diseases were often misclassified.

- **Learning rate scheduler impact:** Identical model training parameters were maintained across all augmentation intensities to get insight into the robustness and scalability of each model. For both the datasets, DenseNet121 and MobileNet had shown initial instability during the first few epochs, which was quickly stabilised due to the LR scheduler. LR scheduler also aided in reaching convergence and overcoming plateaus without leading to overfitting which can be validated by observing the nearly equal validation and test performance presented in Table 4.19, Table 4.5, Table 4.6, Table 4.7, Table 4.8, and Table 4.9.

5.3 Thesis limitations

For this study, we were allocated approximately four months, during which we encountered several issues and the current results also have limitations. The most noteworthy of these are mentioned below.

- **Time limitation:** This thesis faced several limitations and among these, the time limitation constraint is the most significant one. It prevented the implementation and analysis of various image processing and augmentation techniques. Most notably, our work on generative modelling-based data augmentation with VAE remained incomplete. Additionally, we used larger computer vision-based models such as ViT and TransUNet which required significant training time. For larger datasets, even with the support of TPUs and GPUs, training times for the ViT model exceeded six days as seen in Table 4.6.

- **Memory efficiency shortcomings:** Developing a memory-efficient and lightweight model with high accuracy for disease diagnosis is a complex and challenging task. The models used in this study, including our proposed ones, remain large and require further optimisation to improve memory efficiency and reduce size. Our models are not yet optimised for minimal resource use. However, they can serve as a baseline for future efforts aimed at developing models suitable for deployment on both mobile devices and web applications. These objectives have been included in our future work.

- **Imbalanced data challenges:** The Paddy Doctor dataset used for classification was heavily imbalanced. When we attempted to balance this dataset using data augmentation strategies, the models started learning the augmentation noises instead of the actual disease patterns which resulted in poor model generalisation. We also tried creating a balanced dataset by selecting an equal number of samples from each class. Although this approach showed improved generalisation on the validation data, it performed poorly on test data. Due to time constraints, we had to give up this approach and include it in future work.

- **Challenges in paddy segmentation data:** To discard environmental biases and background noises present in the samples, integrating a segmentation model within the paddy diagnosis system or as a preprocessing step is essential. However, due to the absence of an open-access paddy disease segmentation dataset and lack of analysis on the existing segmentation models' applicability for paddy disease diagnosis, this was not possible. In this thesis, we have addressed this research gap by creating a segmentation dataset. But, due to the time constraint, we could not choose a dataset with a complex background and a large number of samples. Despite choosing a dataset with a simple background, this dataset still presented challenges associated with background shadows, minor noises and very similar disease patterns. Even though the selected dataset was small in size, it accurately represented small disease patterns and textures because the samples were taken at close camera angles. This new segmentation dataset and our analysis of the existing segmentation model's applicability in paddy disease detection will work as an initial step towards developing a segmentation-based paddy diagnosis system. Moreover, disease patterns often appear similar and lack proper visual symptoms which necessitate multispectral or hyperspectral imaging as previously mentioned in the research gaps. Additionally, for ground truth mask creation, disease identification can vary significantly based on individual perspectives and disease understanding. Therefore, misclassified disease pixels in the ground truth images might lead to lower prediction performance.

- **Potential bias in assessing model performance:** Despite maintaining consistent training parameters, the usage of a learning rate scheduler might have selectively boosted or hindered the performance of certain models. The learning rate scheduler decreases the learning rate to enhance convergence and escape performance plateaus. Even though the LR scheduler introduces a potential bias, the performance disparity among the models remains minimal within just 2% and suggests that the effect of any bias is minimal. Overall, the learning rate scheduler had a beneficial impact on training efficiency and model performance.

Despite existing limitations, this research significantly contributes to the research community and the field of precision agriculture. It provides a foundation for the researchers to address the identified gaps and effectively tackle associated challenges and issues within this domain.

6 Conclusion and Future Work

This chapter concludes our thesis work, highlighting our contributions to precision agriculture through our exploratory analysis of paddy disease classification and segmentation tasks for automated paddy disease management. We have also detailed the potential scope of extension of this thesis work to address the identified research gaps, and tackle the open issues and challenges associated with this domain.

6.1 Conclusion

Early diagnosis through a computer vision-based system is crucial for having an effective paddy disease management system. This system will help to ensure food security on a global scale and address the challenges resulting from the ever-rising population, changes in the climate, and issues associated with the traditional manual assessment of paddy diseases [1], [3], [8], [27]. For this study, we have investigated the prevalent paddy diseases, and identified their causing agents and visible manifestations. Additionally, an overview of four segmentation models (UNet, VGG16 UNet, TransUNet and Deep residual UNet) and four classification models (DenseNet, MobileNet, Vision transformer and a custom ensemble model of DenseNet121, and Xception) used for this study have been provided with a tabular analysis of their structural differences, advantages and limitations. Moreover, we have identified the key research gaps in this field and addressed the lack of open-access paddy disease segmentation datasets by creating a novel dataset using image processing. The segmentation dataset is available in this link and using this dataset, we have assessed the four existing segmentation models' applicability in paddy disease. From our analysis, we have concluded that Deep Residual UNet is the most suitable model to be used as a preprocessing step or included within the classification model for mitigating the problems associated with backgrounds and overlapping disease symptoms. This decision was made considering its quantitative and qualitative performance, model size, and structural advantages and limitations. Furthermore, we have analysed the impact of training these models with a significantly higher number of augmented images—more than double the original dataset and found that, although the quantitative performance improved with additional data, the qualitative performance occasionally declined. On the other hand, since ViT necessitates significant computational resources and large datasets to provide adequate results, we evaluated whether its comparable performance could achieved by the traditional models or their ensembles and observed it to be feasible. Additionally, the impact of augmentation intensity on the classification models has been explored.

6.2 Future work

Paddy diseases have a serious detrimental effect on plant health and yield production. Our current study finds the optimal machine learning models for the classification and segmentation of paddy diseases. An overview of our future work plans for enhancing precision agriculture has been provided below.

- **Creating a paddy disease segmentation dataset with a complex field environment:** Our research has been conducted using a controlled environment dataset for analysing the utility of segmentation models in the paddy disease segmentation, in broader-scale precision agriculture which had previously shown their utility in remote sensing [29], and biomedical imaging [30]–[32]. Datasets covering samples in complex field

environments need to be created which might require a substantial amount of funding, time and manual labour for conducting the pixel-level annotations. It will help in creating applications for disease severity analysis which will significantly improve the automatic crop monitoring system and disease management process and mitigate the need for manual labour, especially in very large fields which require a lot of time for manual inspection and analysis.

- **Creating a segmentation-based classification model:** Due to the absence of a paddy segmentation dataset and uncertainty about the existing segmentation models' applicability to paddy disease, we were unable to develop a paddy diagnosis model that includes segmentation as either a preprocessing step or within the classification process. We plan to develop a segmentation-based paddy disease classification model.

- **Optimise performance through additional hyperparameter tuning:** In this study, to analyse the performance of the segmentation and classification models, we maintained uniform hyperparameters for finding the most effective model suitable for the paddy disease segmentation and classification tasks. Since we have identified the most effective model, the next step will be to tune the hyper-parameters including the number of attention heads, dropout rates, learning rate, batch size, patch size, number of skip connections, number of hidden layers etc. to get the most optimal results. It is important to note that, for the construction of the model architecture of the ensemble model for the classification task, optimal parameters such as dropout rate, the number of hidden layers, and the type of pooling layers (global or average) were determined utilising Keras Tuner's Random Search method [84].

- **Analyse the impact of image resolution on model performance:** From the study by Chen et al. [55], we know that images with higher resolution give better results. Hence, our future work will also incorporate conducting experimental analysis with different image resolutions and observing its effect on the results for the paddy disease segmentation and classification.

- **Creating weed segmentation dataset in a complex field environment:** Weeds have one of the most prominent adversarial effects on plant health after various plant diseases. They take away the essential water, nutrients and sunlight which would have been otherwise utilised by the crop. Thus, weeds in the crop fields have a detrimental effect on crop health and the overall crop yield [38]. An autonomous system which is capable of identifying weeds from fields is a necessity for improving the overall crop yield. Weed segmentation tasks have been conducted in the utilising dataset collected from different crops such as maize [85], brinjal/eggplant [86], sunflower [38] and sorghum [87]. But so far, no work has been done for paddy fields and no dataset is available for the paddy field weed segmentation. We can address this research gap by providing a dataset for paddy weed segmentation using open-access paddy datasets with field environment samples [16], [65], [67]–[71].

- **Develop a memory-efficient classification model:** Our future goals encompass implementing a small-scale model for the effective classification of paddy diseases utilising publicly available datasets [6], [16], [65]–[71]. Synthetic data will be created using traditional and generative model-based image augmentation techniques to enhance the dataset diversity, mitigate overfitting, and minimise the need for additional data acquisition. We will evaluate the model's capacity to manage disease symptoms and pattern changes caused by the variations in geolocation. We also aim to ensure the model's versatility, allowing it to accurately detect and classify diseases across crops, not limited to paddy.

- **Develop a model integrating optimal segmentation and classification models:** Since early disease identification is essential for efficient disease management and improving the overall crop yield, we plan to develop an autonomous disease identification and classification system. For this, we will use our identified optimal segmentation model as a pre-processing step to tackle complex backgrounds and use the developed memory-efficient classification model for classifying the disease.

- **Develop a web or mobile-based application:** A mobile or web application will be developed using a memory-efficient model. In addition to guaranteeing the timely and accurate classification of contaminated diseases, this model and application will help to maximise resource usage, reduce environmental damage and ensure enhanced agricultural yield.
- **Use generative models for data augmentation:** The data collection process is costly and time-consuming and obtaining samples for rare diseases is even more difficult. Hence, to solve this problem, we have to rely on data augmentation techniques. In this study, for data augmentation, we have utilised only the traditional methods. We are planning to analyse the generative model-based augmentations and their effect on the performance of the models.

By addressing the tasks outlined in this section, improvements in the paddy disease diagnosis system can be achieved which aligns with our goal of contributing to precision agriculture.

A Usage of AI

Due to the nature of language models/AI tools to provide inaccurate and misleading information related to literature and references, these were not used to find literature and references. Reliable sources such as Clarivate (Web of Science), Google Scholar, Elsevier, Springer, and IEEE Xplore were used to find relevant literature and references. Considering the ethical use of AI and the guidelines provided by the NMBU, ChatGPT was utilised only in the following ways for this thesis:

- **Locate synonyms and connecting words:** While writing the report, we addressed the word repetition issue by using ChatGPT to identify synonyms and linking words which enhanced the report's quality. For instance, to avoid using "demonstrates" repeatedly in the sentence "Figure W demonstrates that ViT-based models required the most training time across all augmentation variations", we employed AI tools to find suitable replacements.
  - **Prompt for ChatGPT:** "Figure W demonstrates that ViT-based models required the most training time across all augmentation variations"–give suitable words to replace "demonstrates"
- **Latex assistance:** ChatGPT was utilised for addressing queries regarding report formatting, such as creating new pages, printing all contents before a new section starts and adjusting spaces between sections and figures. Additionally, it significantly accelerated the creation of numerous tables related to the literature review, methodology, and results by providing LaTeX structures. For instance, the prompt used to generate the latex table structure for the model analysis was as follows:
  - **Prompt for ChatGPT:** Create a latex table that has 4 columns (Properties, DenseNet121, Inceptionv3, Xception) and rows (Properties, Model category, Model size (MB), Activation function, Trainable parameters, Non-trainable parameters, Total parameters, Advantages, Limitations) and fill the table with W to be replaced with real values.
- **Coding assistance:** In addition to Stack Overflow, ChatGPT was used for diagnosing code errors and exploring possible debugging techniques. It was also utilised to customise various features of Matplotlib and Seaborn for making the figures/plots. For instance, in Matplotlib, titles are usually placed at the top of figures using "title". But for the report, titles were required to be put below each subplot. ChatGPT was prompted in the following way for assistance.
  - **Prompt for ChatGPT:** When using Matplotlib, how to put the model names below the plot as a text instead of putting them on the title?

The usage of AI for repetitive and time-consuming tasks unrelated to core research such as LaTeX formatting and synonym finding has accelerated the thesis process, allowing greater focus on reviewing current literature, identifying research gaps, and addressing these gaps through our research.

Bibliography

[1] P. Sobiyaa, K. Jayareka, K. Maheshkumar, S. Naveena, and K. S. Rao, "Paddy disease classification using machine learning technique," *Materials Today: Proceedings*, vol. 64, pp. 883–887, 2022, International Conference on Advanced Materials for Innovation and Sustainability, ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2022.05.398. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785322036975.

[2] V. K. Shrivastava and M. K. Pradhan, "Rice plant disease classification using color features: A machine learning paradigm," *Journal of Plant Pathology*, vol. 103, no. 1, pp. 17–26, 2021, ISSN: 2239-7264. DOI: 10.1007/s42161-020-00683-3. [Online]. Available: https://doi.org/10.1007/s42161-020-00683-3.

[3] J. Chen, D. Zhang, A. Zeb, and Y. A. Nanehkaran, "Identification of rice plant diseases using lightweight attention networks," *Expert Systems with Applications*, vol. 169, p. 114 514, 2021, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.114514. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420311581.

[4] G. Latif, S. E. Abdelhamid, R. E. Mallouhy, J. Alghazo, and Z. A. Kazimi, "Deep learning utilization in agriculture: Detection of rice plant diseases using an improved cnn model," *Plants*, vol. 11, no. 17, 2022, ISSN: 2223-7747. DOI: 10.3390/plants11172230. [Online]. Available: https://www.mdpi.com/2223-7747/11/17/2230.

[5] M. Aggarwal, V. Khullar, and N. Goyal, "Contemporary and futuristic intelligent technologies for rice leaf disease detection," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2022, pp. 1–6. DOI: 10.1109/ICRITO56286.2022.9965113.

[6] C. R. Rahman, P. S. Arko, M. E. Ali, *et al.*, "Identification and recognition of rice diseases and pests using convolutional neural networks," *Biosystems Engineering*, vol. 194, pp. 112–120, 2020, ISSN: 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2020.03.020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1537511020300830.

[7] S. Lamba, A. Baliyan, and V. Kukreja, "A novel gcl hybrid classification model for paddy diseases," *International Journal of Information Technology*, vol. 15, no. 2, pp. 1127–1136, 2023, ISSN: 2511-2112. DOI: 10.1007/s41870-022-01094-6. [Online]. Available: https://doi.org/10.1007/s41870-022-01094-6.

[8] A. A. Salamai, N. Ajabnoor, W. E. Khalid, M. M. Ali, and A. A. Murayr, "Lesion-aware visual transformer network for paddy diseases detection in precision agriculture," *European Journal of Agronomy*, vol. 148, p. 126 884, 2023, ISSN: 1161-0301. DOI: https://doi.org/10.1016/j.eja.2023.126884. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1161030123001521.

[9] J. Chen, J. Chen, D. Zhang, Y. Sun, and Y. Nanehkaran, "Using deep transfer learning for image-based plant disease identification," *Computers and Electronics in Agriculture*, vol. 173, p. 105 393, 2020, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2020.105393. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169919322422.

[10] A. Haridasan, J. Thomas, and E. D. Raj, "Deep learning system for paddy plant disease detection and classification," *Environmental Monitoring and Assessment*, vol. 195, no. 1, p. 120, 2022, ISSN: 1573-2959. DOI: 10.1007/s10661-022-10656-x. [Online]. Available: https://doi.org/10.1007/s10661-022-10656-x.

[11] N. Senan, M. Aamir, R. Ibrahim, N. S. A. M. Taujuddin, and W. W. Muda, "An efficient convolutional neural network for paddy leaf disease and pest classification," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, 2020. DOI: 10.14569/IJACSA.2020.0110716. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0110716.

[12] H. Andrianto, Suhardi, A. Faizal, and F. Armandika, "Smartphone application for deep learning-based rice plant disease detection," in *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, 2020, pp. 387–392. DOI: 10.1109/ICITSI50517.2020.9264942.

[13] S. P. Singh, K. Pritamdas, K. J. Devi, and S. D. Devi, "Custom convolutional neural network for detection and classification of rice plant diseases," *Procedia Computer Science*, vol. 218, pp. 2026–2040, 2023, International Conference on Machine Learning and Data Engineering, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2023.01.179. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923001795.

[14] S. S. Hari, M. Sivakumar, P. Renuga, S. karthikeyan, and S. Suriya, "Detection of plant disease by leaf image using convolutional neural network," in *2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN)*, 2019, pp. 1–5. DOI: 10.1109/ViTECoN.2019.8899748.

[15] Y. Wang, H. Wang, and Z. Peng, "Rice diseases detection and classification using attention based neural network and bayesian optimization," *Expert Systems with Applications*, vol. 178, p. 114 770, 2021, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114770. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417421002116.

[16] P. A, B. K. S, M. D, and P. Arjunan, "Paddy doctor: A visual image dataset for automated paddy disease classification and benchmarking," 2022. DOI: 10.21227/hz4v-af08. [Online]. Available: https://dx.doi.org/10.21227/hz4v-af08.

[17] P. A, M. D, and B. S, "Paddynet: An improved deep convolutional neural network for automated disease identification on visual paddy leaf images," *International Journal of Advanced Computer Science and Applications*, vol. 14, Jan. 2023. DOI: 10.14569/IJACSA.2023.01406122.

[18] Y. Borhani, J. Khoramdel, and E. Najafi, "A deep learning based approach for automated plant disease classification using vision transformer," *Scientific Reports*, vol. 12, no. 1, 2022. DOI: 10.1038/s41598-022-15163-0. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133584864&doi=10.1038%2fs41598-022-15163-0&partnerID=40&md5=b26c80fe07d4d90147e73567d01f5e63.

[19] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, *Explainable vision transformer enabled convolutional neural network for plant disease identification: Plantxvit*, 2022. DOI: https://doi.org/10.48550/arXiv.2207.07919. arXiv: 2207.07919 [cs.CV].

[20] H.-T. Thai, K.-H. Le, and N. L.-T. Nguyen, "Formerleaf: An efficient vision transformer for cassava leaf disease detection," *Computers and Electronics in Agriculture*, vol. 204, p. 107 518, 2023, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2022.107518. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169922008262.

[21] F. Arshad, M. Mateen, S. Hayat, *et al.*, "Pldpnet: End-to-end hybrid deep learning framework for potato leaf disease prediction," *Alexandria Engineering Journal*, vol. 78, pp. 406–418, 2023, ISSN: 1110-0168. DOI: https://doi.org/10.1016/j.aej.2023.07.076. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S111001682300666X.

[22] Q. Zeng, L. Niu, S. Wang, and W. Ni, "Sevit: A large-scale and fine-grained plant disease classification model based on transformer and attention convolution," *Multimedia Systems*, vol. 29, no. 3, pp. 1001–1010, 2023. DOI: 10.1007/s00530-022-01034-1. [Online]. Available: https://doi.org/10.1007/s00530-022-01034-1.

[23] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, "Vision transformer for plant disease detection: Plantvit," in *Computer Vision and Image Processing*, B. Raman, S. Murala, A. Chowdhury, A. Dhall, and P. Goyal, Eds., Cham: Springer International Publishing, 2022, pp. 501–511, ISBN: 978-3-031-11346-8. DOI: https://doi.org/10.1007/978-3-031-11346-8_43.

[24] P. S. Thakur, S. Chaturvedi, P. Khanna, T. Sheorey, and A. Ojha, "Vision transformer meets convolutional neural network for plant disease classification," *Ecological Informatics*, vol. 77, p. 102 245, 2023, ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2023.102245. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954123002741.

[25] M. Mustak Un Nobi, M. Rifat, M. F. Mridha, S. Alfarhood, M. Safran, and D. Che, "Gld-det: Guava leaf disease detection in real-time using lightweight deep learning approach based on mobilenet," *Agronomy*, vol. 13, no. 9, 2023, ISSN: 2073-4395. DOI: 10.3390/agronomy13092240. [Online]. Available: https://www.mdpi.com/2073-4395/13/9/2240.

[26] Y. Yang, G. Jiao, J. Liu, W. Zhao, and J. Zheng, "A lightweight rice disease identification network based on attention mechanism and dynamic convolution," *Ecological Informatics*, vol. 78, p. 102 320, 2023, ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2023.102320. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954123003497.

[27] S. Radhakrishnan, "An improved machine learning algorithm for predicting blast disease in paddy crop," *Materials Today: Proceedings*, vol. 33, pp. 682–686, 2020, International Conference on Future Generation Functional Materials and Research 2020, ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2020.05.802. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785320344151.

[28] T. Daniya and S. Vigneshwari, "Rider water wave-enabled deep learning for disease detection in rice plant," *Advances in Engineering Software*, vol. 182, p. 103 472, 2023, ISSN: 0965-9978. DOI: https://doi.org/10.1016/j.advengsoft.2023.103472. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0965997823000649.

[29] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, May 2018, ISSN: 1558-0571. DOI: 10.1109/lgrs.2018.2802944. [Online]. Available: http://dx.doi.org/10.1109/LGRS.2018.2802944.

[30] O. Oktay, J. Schlemper, L. L. Folgoc, *et al.*, *Attention u-net: Learning where to look for the pancreas*, 2018. arXiv: 1804.03999 [cs.CV].

[31] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, 2015. arXiv: 1505.04597 [cs.CV].

[32] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, Jan. 2020, ISSN: 0893-6080. DOI: 10.1016/j.neunet.2019.08.025. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2019.08.025.

[33] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, *Encoder-decoder with atrous separable convolution for semantic image segmentation*, 2018. arXiv: 1802.02611 [cs.CV].

[34] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, *Rethinking atrous convolution for semantic image segmentation*, 2017. arXiv: 1706.05587 [cs.CV].

[35] S. Zhu, W. Ma, J. Lu, B. Ren, C. Wang, and J. Wang, "A novel approach for apple leaf disease image segmentation in complex scenes based on two-stage deeplabv3+ with adaptive loss," *Computers and Electronics in Agriculture*, vol. 204, p. 107 539, 2023, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2022.107539. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016816992200847X.

[36] S. Zhang and C. Zhang, "Modified u-net for plant diseased leaf image segmentation," *Computers and Electronics in Agriculture*, vol. 204, p. 107 511, 2023, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2022.107511. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169922008195.

[37] K. Perveen, S. Debnath, B. Pandey, *et al.*, "Deep learning-based multiscale cnn-based u network model for leaf disease diagnosis and segmentation of lesions in tomato," *Physiological and Molecular Plant Pathology*, vol. 128, p. 102 148, 2023, ISSN: 0885-5765. DOI: https://doi.org/10.1016/j.pmpp.2023.102148. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885576523002035.

[38] H. M. Sahin, T. Miftahushudur, B. Grieve, and H. Yin, "Segmentation of weeds and crops using multispectral imaging and crf-enhanced u-net," *Computers and Electronics in Agriculture*, vol. 211, p. 107 956, 2023, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2023.107956. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169923003447.

[39] P. K. Sethy, N. K. Barpanda, A. K. Rath, and S. K. Behera, "Image processing techniques for diagnosing rice plant disease: A survey," *Procedia Computer Science*, vol. 167, pp. 516–530, 2020, International Conference on Computational Intelligence and Data Science, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.03.308. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920307742.

[40] M. Deb, K. G. Dhal, R. Mondal, and J. Gálvez, "Paddy disease classification study: A deep convolutional neural network approach," *Optical Memory and Neural Networks*, vol. 30, no. 4, pp. 338–357, 2021, International Conference on Future Generation Functional Materials and Research 2021, ISSN: 1934-7898. DOI: 10.3103/S1060992X2104007X. [Online]. Available: https://doi.org/10.3103/S1060992X2104007X.

[41] A. Nigam, A. K. Tiwari, and A. Pandey, "Paddy leaf diseases recognition and classification using pca and bfo-dnn algorithm by image processing," *Materials Today: Proceedings*, vol. 33, pp. 4856–4862, 2020, International Conference on Nanotechnology: Ideas, Innovation and Industries, ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2020.08.397. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785320362696.

[42] *Rice stem rot*, https://ipm.ucanr.edu/agriculture/rice/stem-rot-of-rice/, Accessed on 7th February, 2024, Apr. 2004.

[43] *Rice stem rot*, http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/diseases/item/stem-rot, Accessed on 7th February, 2024, n.d.

[44] *Bakanae*, http://www.agritech.tnau.ac.in/expert_system/paddy/cpdisbakanae.html, Accessed on 7th February, 2024, n.d.

[45] A. Sparks, N. Castilla, and C. Vera Cruz, *Bakanae*, http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/diseases/item/bakanae, Published by Rice Knowledge Bank. Accessed on 7th February, 2024, n.d.

[46] *Bakanae*, https://ipm.ucanr.edu/agriculture/rice/bakanae/, Published by UC IPM. Accessed on 7th February, 2024, Apr. 2004.

[47] *Tungro*, http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/diseases/item/tungro, Published by Rice Knowledge Bank. Accessed on 29th October, 2023, n.d.

[48] *Stem borer*, http://www.knowledgebank.irri.org/training/fact-sheets/pest-management/insects/item/stem-borer, Accessed on 7th February, 2024, n.d.

[49] *Stem borer*, http://www.agritech.tnau.ac.in/expert_system/paddy/cppests_SB.html, Accessed on 7th February, 2024, n.d.

[50] A. K. Chaitanya, H. V. R. Jamedar, A. Shanmugam, *et al.*, "Chapter 7 - advances in qtl mapping for biotic stress tolerance in wheat," in *QTL Mapping in Crop Improvement*, S. H. Wani, D. Wang, and G. Pratap Singh, Eds., Boston: Academic Press, 2023, pp. 119–148, ISBN: 978-0-323-85243-2. DOI: https://doi.org/10.1016/B978-0-323-85243-2.00025-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323852432000258.

[51] H. R. Kappali, S. K.M., and P. S.K., "Parametric evaluation of segmentation techniques for paddy diseases analysis," *Journal of Agricultural Engineering*, Aug. 2023. DOI: 10.4081/jae.2023.1532. [Online]. Available: https://www.agroengineering.org/index.php/jae/article/view/1532.

[52] M. Ramkumar Raja, J. V, F. H. Shajin, and E. Roopa Devi, "Radial basis function neural network optimized with salp swarm algorithm espoused paddy leaf disease classification," *Biomedical Signal Processing and Control*, vol. 86, p. 105 038, 2023, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2023.105038. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809423004718.

[53] N. Raj, S. Perumal, S. Singla, G. K. Sharma, S. Qamar, and A. P. Chakkaravarthy, "Computer aided agriculture development for crop disease detection by segmentation and classification using deep learning architectures," *Computers and Electrical Engineering*, vol. 103, p. 108 357, 2022, ISSN: 0045-7906. DOI: https://doi.org/10.1016/j.compeleceng.2022.108357. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0045790622005742.

[54] S. Ramesh and D. Vydeki, "Recognition and classification of paddy leaf diseases using optimized deep neural network with jaya algorithm," *Information Processing in Agriculture*, vol. 7, no. 2, pp. 249–260, 2020, ISSN: 2214-3173. DOI: https://doi.org/10.1016/j.inpa.2019.09.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214317319300769.

[55] J. Chen, Y. Lu, Q. Yu, *et al.*, *Transunet: Transformers make strong encoders for medical image segmentation*, 2021. arXiv: 2102.04306 [cs.CV].

[56] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556 [cs.CV].

[57] Y. Li and J. Yang, "Meta-learning baselines and database for few-shot classification in agriculture," *Computers and Electronics in Agriculture*, vol. 182, p. 106 055, 2021, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2021.106055. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169921000739.

[58] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

[59] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, 2018. arXiv: 1608.06993 [cs.CV].

[60] A. G. Howard, M. Zhu, B. Chen, *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. DOI: 10.48550/arXiv.1704.04861. [Online]. Available: https://arxiv.org/abs/1704.04861.

[61] C. Zhang, P. Benz, D. M. Argaw, *et al.*, *Resnet or densenet? introducing dense shortcuts to resnet*, 2020. arXiv: 2010.12496 [cs.CV].

[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. DOI: https://doi.org/10.48550/arXiv.2010.11929. arXiv: 2010.11929 [cs.CV].

[63] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, *Understanding robustness of transformers for image classification*, 2021. arXiv: 2103.14586 [cs.CV].

[64] F. Chollet, *Xception: Deep learning with depthwise separable convolutions*, 2017. arXiv: 1610.02357 [cs.CV].

[65] P. K. Sethy, *Rice leaf disease image samples*, https://doi.org/10.17632/fwcj7stb8r.1, 2020. DOI: 10.17632/fwcj7stb8r.1.

[66] J. Shah, H. Prajapati, and V. Dabhi, *Rice leaf diseases*, UCI Machine Learning Repository, 2019. DOI: 10.24432/C5R013.

[67] M. F. Hossain, S. Abujar, S. R. H. Noori, and S. A. Hossain, *Dhan-shomadhan: A dataset of rice leaf disease classification for bangladeshi local rice*, https://doi.org/10.17632/znsxdctwtt.1, 2021. DOI: 10.17632/znsxdctwtt.1.

[68] *Microsoft rice disease classification challenge*, https://zindi.africa/competitions/microsoft-rice-disease-classification-challenge, Published by Unknown. Accessed on 29th October, 2023., n.d.

[69] J.-H. Ou and C.-y. Chen, *Rice leaf diseases in taiwan*, 2022. DOI: 10.34740/KAGGLE/DS/2012808. [Online]. Available: https://www.kaggle.com/ds/2012808.

[70] S. Agarwa, *Philippines rice diseases*, https://www.kaggle.com/datasets/shrupyag001/philippines-rice-diseases, Published by Shruti Agarwa. Accessed on 29th October, 2023., n.d.

[71] T. Setiady, *Leaf rice disease*, https://www.kaggle.com/datasets/tedisetiady/leaf-rice-disease-indonesia, Published by Tedi Setiady. Accessed on 29th October, 2023., n.d.

[72] H. M. DO, *Rice diseases image dataset*, 2022. DOI: Notgiven. [Online]. Available: https://www.kaggle.com/datasets/minhhuy2810/rice-diseases-image-dataset.

[73] A. Ahmad, D. Saraswat, and A. El Gamal, "A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools," *Smart Agricultural Technology*, vol. 3, p. 100 083, 2023, ISSN: 2772-3755. DOI: https://doi.org/10.1016/j.atech.2022.100083. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S277237552200048X.

[74] S. Lamba, V. Kukreja, A. Baliyan, S. Rani, and S. H. Ahmed, "A novel hybrid severity prediction model for blast paddy disease using machine learning," *Sustainability*, vol. 15, no. 2, 2023, ISSN: 2071-1050. DOI: 10.3390/su15021502. [Online]. Available: https://www.mdpi.com/2071-1050/15/2/1502.

[75] A. Pal and V. Kumar, "Agridet: Plant leaf disease severity classification using agriculture detection framework," *Engineering Applications of Artificial Intelligence*, vol. 119, p. 105 754, 2023, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2022.105754. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197622007448.

[76] *Reducelronplateau*, https://keras.io/api/callbacks/reduce_lr_on_plateau/, Published by Keras. Accessed on 20th April, 2024, n.d.

[77] M. S. H. Shovon, S. J. Mozumder, O. K. Pal, M. F. Mridha, N. Asai, and J. Shin, "Plantdet: A robust multi-model ensemble method based on deep learning for plant disease detection," *IEEE Access*, vol. 11, pp. 34 846–34 859, 2023. DOI: 10.1109/ACCESS.2023.3264835.

[78] M. M. Islam, M. A. A. Adil, M. A. Talukder, *et al.*, "Deepcrop: Deep learning-based crop disease prediction with web application," *Journal of Agriculture and Food Research*, vol. 14, p. 100 764, 2023, ISSN: 2666-1543. DOI: https://doi.org/10.1016/j.jafr.2023.100764. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154323002715.

[79] Y. Guo, Y. Lan, and X. Chen, "Cst: Convolutional swin transformer for detecting the degree and types of plant diseases," *Computers and Electronics in Agriculture*, vol. 202, p. 107 407, 2022, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2022.107407. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169922007153.

[80] *Earlystopping*, https://keras.io/api/callbacks/early_stopping/, Published by Keras. Accessed on 23rd April, 2024, n.d.

[81] L. Perez and J. Wang, *The effectiveness of data augmentation in image classification using deep learning*, 2017. arXiv: 1712.04621 [cs.CV].

[82] J. Kotwal, D. Kashyap, and D. Pathan, "Agricultural plant diseases identification: From traditional approach to deep learning," *Materials Today: Proceedings*, vol. 80, pp. 344–356, 2023, 3rd International Congress on Mechanical and Systems Engineering (CAMSE 2022), ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2023.02.370. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785323009343.

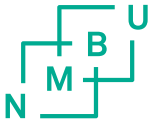
[83] N. N. Kurniawati, S. N. H. S. Abdullah, S. Abdullah, and S. Abdullah, "Texture analysis for diagnosing paddy disease," in *2009 International Conference on Electrical Engineering and Informatics*, vol. 01, 2009, pp. 23–27. DOI: 10.1109/ICEEI.2009.5254824.

[84] *Randomsearch tuner*, https://keras.io/api/keras_tuner/tuners/random/, Published by Keras. Accessed on 26th March, 2024, n.d.

[85] J. Gao, W. Liao, D. Nuyttens, *et al.*, "Cross-domain transfer learning for weed segmentation and mapping in precision farming using ground and uav images," *Expert Systems with Applications*, vol. 246, p. 122 980, 2024, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.122980. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423034826.

[86] S. K. Gupta, S. K. Yadav, S. K. Soni, U. Shanker, and P. K. Singh, "Multiclass weed identification using semantic segmentation: An automated approach for precision agriculture," *Ecological Informatics*, vol. 78, p. 102 366, 2023, ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2023.102366. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954123003953.

[87] M. J. Justina and M. Thenmozhi, "Sorghumweed dataset classification and sorghumweed dataset segmentation datasets for classification, detection, and segmentation in deep learning," *Data in Brief*, vol. 52, p. 109 935, 2024, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2023.109935. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340923009678.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway