Multimodal Peach Tree Disease Detection

Israa Bashir

Master of Science

School of Informatics

University of Edinburgh

2024

Abstract

This study explores the use of multimodal fusion approaches to combine information from many sources, such as ground RGB photos, multispectral imagery, and tabular data, to improve the classification performance of peach tree health and pest infestation detection. This strategy is especially pertinent to precision agriculture, where it is essential to have accurate crop health monitoring and diagnosis capabilities. Using a variety of deep learning models, such as InceptionV3, ResNet152, VGG19, ViT, and AttentionAugmented architectures, we carried out a thorough examination across two distinct fusion strategies: late fusion and intermediate fusion. Our results show that multimodal fusion, especially with attention-augmented and transformer-based models, considerably increases classification accuracy, precision, recall, and F1-scores. Notwithstanding these advancements, the study also notes many drawbacks, including the requirement for cautious model selection and sensitivity to class imbalance. The findings demonstrate how multimodal fusion may be used to combine complimentary data from several data modalities, opening up exciting new directions for further study on how to best implement these tactics to improve agricultural monitoring and decision-making.

Research Ethics Approval

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Israa Bashir*)

Acknowledgements

I would want to sincerely thank everyone for their support during this difficult journey. This research paper is not only the result of my academic work, but it also symbolises the daily struggle I have been going through since I started this program.

This work belongs as much to my family as it does to me, who are the most significant individuals in my life. It has been an unfathomable nightmare to be absent from you all this period, particularly since the crisis in Sudan. I had hoped that I could support you virtually, but being apart has been nearly too much to take. It has hurt to watch you struggle to survive from a distance and know that I am unable to help you.

Moving overseas was supposed to be a chance for personal development and novel experiences, but it has turned into a lonely and hopeless experience. I struggle with the realisation that our house has been destroyed and that the people I care about are in danger every day. The sadness that has replaced the joy I previously experienced from following my ambition has made it extremely difficult for me to concentrate on my schoolwork because of these realities.

I would especially like to thank James Garforth, my supervisor. Your advice, tolerance, and comprehension have been very helpful, particularly during the most trying times. Without your mentorship, I would not have progressed to this stage. To Fay and AlJenan I appreciate your belief in me, even though I found it difficult to recognise it in myself. I am incredibly grateful for the intellectual and personal assistance you gave me; your encouragement has kept me going forward.

To my google deep mind mentor, I sincerely appreciate your support and understanding. Your advice has been really helpful, particularly at the moments when I felt like my situation was too much to handle. I appreciate you giving me the room I needed to think through and deal with all that has been going on in my life and for your unwavering support in pushing through the difficult times.

Originally intended to be a celebration of my academic career, this research has evolved into a mirror of the inner resilience I've had to discover. Even though this chance has come at such a difficult time, words cannot begin to describe how much it means to me. I'm immensely grateful for the opportunity to develop, and I hope to return home someday to aid in the reconstruction of what has been destroyed.

Finally, this work is dedicated to my family in Egypt and Sudan. I wrote everything with love and a strong sense of purpose for our future. I commit to keeping up my efforts for all of us, not just for myself. Every day, your tenacity and courage inspire me. Thank you.

Table of Contents

Chapter 1

Introduction

1.1 Problem Statement

Agriculture is the main industry in emerging nations, employing a large percentage of the labour force and being vital to their economies. For example, in Sub-Saharan Africa, agriculture accounts for over 23% of GDP and employs over 60% of the labour force; cash crops, such as cocoa and coffee, are especially important. Likewise, rice is quite important in Asian nations like Vietnam and India.[60] Despite the vital role that agriculture plays, plant diseases account for up to 40% of crop losses worldwide each year.[22] This exacerbates food poverty and economic instability in these places.

In order to address these issues and improve crop output, sustainability, and the financial gains associated with cash crops, this project will create a precision agriculture system that uses deep learning and machine learning to identify and control complex diseases.

1.2 Objectives

The primary objectives of this study are:

- To create a multimodal deep learning model to detect illnesses in important cash crops by combining environmental data, multispectral imagery, UAV imagery, and ground imagery.
- To look into how the accuracy of disease detection is affected when deep learning models based on images are combined with vegetation indices.
- To assess how well multimodal models work in comparison to single-modality.

1.3 Scope of the Study

The integration of multimodal deep learning algorithms for the precise and prompt diagnosis of illnesses in cash crops is the main emphasis of this study. In order to improve the precision and effectiveness of disease diagnosis, the project will investigate the advantages of merging several data sources, including pictures, symptom descriptions, and environmental variables. The creation of transportable and easily available solutions for farmers in isolated regions is also included in the scope, with the goal of enhancing agricultural methods and promoting stable economies.

1.4 Significance of the Study

This study is important from an academic standpoint because it sheds light on how cutting-edge technology can be incorporated with conventional farming methods. In terms of industry, it provides scalable approaches to crop disease management, which is essential for preserving agricultural output and monetary stability in underdeveloped nations. Socially, the study is in line with the Sustainable Development Goals of the UN, which include ensuring food security and eradicating poverty. This research has the potential to improve agricultural resilience, productivity, and livelihoods in vulnerable regions by advancing agricultural technology and fostering economic growth, especially in light of growing threats from diseases and climate change.[1][5][7][39]

1.5 Structure of the Thesis

This thesis is structured as follows:

- **Chapter 1: Introduction** - Gives an overview of the problem, the motivation for the study, objectives, scope, and significance.
- **Chapter 2: Background** - Introduces key concepts and reviews existing literature on single modal and multimodal approaches in relation to agricultural disease detection.
- **Chapter 3: Methodology** - Describes the research design, data collection methods, using bounding boxes, and the development of the multimodal DL model.
- **Chapter 4: Results and Discussion** - Explains the results and its implications.
- **Chapter 5: Conclusion and Recommendations** - Summarizes the study, its contributions and limitations, and suggests directions for future research.

Chapter 2

Background

This study is focused on creating a multimodal strategy that will improve the precision and effectiveness of peach tree disease diagnosis by utilising ground-based RGB, UAV, and multispectral imaging. Thus, in this chapter we examine the diseases that might seriously affect peach trees and cause severe losses to the harvest. As well as reviews the existing literature to show the gaps in the methods currently being used.

2.1 Overview of Peach Tree Diseases

2.1.1 Anarsia lineatella

Anarsia lineatella, or the peach twig borer, is a serious insect that affects peach trees. It attacks the vulnerable areas of the tree, like young branches and fruits, seriously harming the tree and its fruit production.[44] Wilting and browning of fresh shoots are among the early indications, which frequently result in dieback, this is shown in Figure 2.1a Peach trees that are severely damaged by Anarsia lineatella may produce fewer and lower-quality fruits. In extreme circumstances, the tree may have reduced development and increased susceptibility to external pressures.

2.1.2 Grapholita molesta

An other significant pest of peach plants is Grapholita molesta, commonly referred to as the oriental fruit moth. This insect mostly attacks the fruits and shoots in orchards, damaging the Stem borer, fruit surface and flesh and resulting in harm that might cause financial losses.[44] Grapholita molesta larvae dig into young shoots, similar to Anarsia lineatella, causing them to wilt and die. This is mainfested mainly in the bracnches

and trunks of the trees ass shown in Figure 2.1b The overall health of the trees is also impacted by the damage to the shoots, leaving them more susceptible to pests and environmental stressors.

2.1.3 Stress Indicators (Due to Various Factors including Disease)

Apart from ailments resulting from Grapholita molesta and Anarsia lineatella, stress symptoms in peach trees can be attributed to several environmental variables such as dryness and nutrient deficits.[25] Accurate diagnosis can be difficult since these stress signals, like wilting, colour changes, and deformation in leaves, might mimic illness symptoms. Stress-affected branches may also exhibit symptoms that are mistaken for illness, such as vitality loss and dieback. In addition to disease symptoms, environmental variables need to be taken into account because stressed trees are more susceptible to pests and diseases, which can worsen their state. Figure 2.1c shows an example of water stress.

(a) Anarsia lineatella[57] (b) Grapholita molesta (c) Water Stress

Figure 2.1: Peach Tree Diseases

2.2 Overview of the AI

2.2.1 Image-based Approaches in Agriculture

Traditionally, agricultural studies have made substantial use of image-based techniques to monitor crop health, identify diseases, and evaluate environmental conditions. Because Convolutional Neural Networks (CNNs) are able to record spatial hierarchies in images, they have become the de facto architecture for processing image data. In this field, models like VGG, ResNet, and Inception are commonly used. These models, however, may miss other important aspects that could be recorded using non-visual data because they mostly concentrate on visual information.[58][20][17]

2.2.2 Tabular Data and Its Importance

Agricultural disease detectionheavily relies on imagery as well as tabular data from several sensors, environmental measures, and handwritten information. A thorough understanding of crop health and productivity requires knowledge of factors including plant physiology, weather patterns, and soil qualities, all of which can be found in this data. It has been demonstrated that models that efficiently handle and integrate this tabular data, including transformer-based architectures and Multilayer Perceptrons (MLPs), enhance image-based techniques by offering extra context that images by themselves cannot.[48]

2.2.3 Multimodal Learning

The concept of multimodal learning aims to enhance predictive performance by merging data from several modalities, including images and tabular data. Multimodal models can be used in the agricultural setting to combine numerical and visual data to enable better disease detection. Effectively combining these disparate data kinds presents a difficulty because each modality may provide unique, sometimes non-overlapping information.[6]

2.2.4 Fusion Techniques in Multimodal Learning

Lahat et a.'s literature[33] has suggested a number of fusion strategies, each having pros and cons. Early fusion allows the model to learn joint representations from the beginning by combining modalities at the input level. Contrarily, late fusion handles each modality independently before combining its outputs, which is advantageous in situations where the modalities have unique properties. By combining modalities at an intermediate layer, intermediate fusion provides a medium ground that balances the advantages of early and late fusion. Figure 2.2 shows the different fusion techniques. These fusion techniques are essential for creating multimodal models that work well, particularly for challenging tasks like agricultural classification where different data sources offer complementary but differing insights.

2.2.5 Attention Mechanisms

The capabilities of multimodal models have been significantly improved by recent developments in attention mechanisms. Attention mechanisms can help the model handle

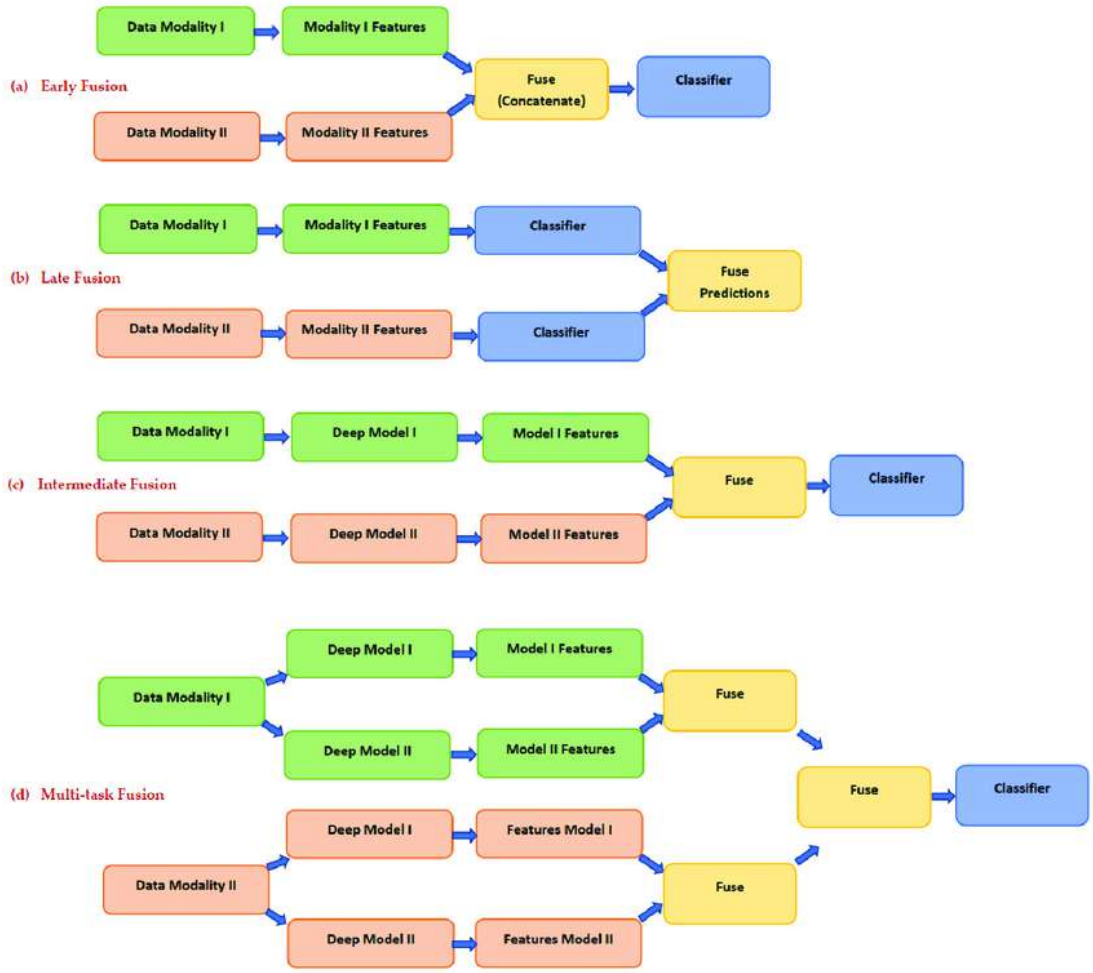


Figure 2.2: Fusion Techniques in Multimodal Learning[4]

complicated, multimodal inputs by enabling it to concentrate on the most pertinent portions of the input data. Attention-augmented models can be used in agricultural applications to simultaneously weight significant elements in the tabular data (e.g., soil pH levels) and prioritise essential locations in pictures (e.g., sick leaves).[9]

2.3 Overview of Existing Research

In this section we have review the existing literature and divided it according to single modal vs multimodal. For the single modal we have focused on image based modal and tabular based modal.

2.3.1 Single Modal in Crop Disease Detection: Image-Based Data

2.3.1.1 Ground RGB Images:

The capacity of ground-based RGB imaging to capture obvious symptoms like discolouration and lesions on plant leaves has made it a popular tool for crop disease identification. Deep learning methods for analysing these photos have been investigated and refined in a number of research, each making a distinct contribution to the field.[59][28]

Mohanty et al. (2016) used a deep convolutional neural network (CNN) to categorise 26 diseases across 14 crop species, laying the groundwork for the use of deep learning in plant disease identification. With an astounding accuracy of over 99%, their model demonstrated the promise of CNNs for high-precision RGB image classification tasks. However, the predetermined selection of crops and diseases limited the study's scope and made it less generalisable to a wider range of agricultural scenarios.[42]

Sladojevic et al. (2016), on the other hand, expanded on this framework by creating a more comprehensive model that could identify several illnesses in various plant species using just one CNN. The goal of this method was to increase the model's adaptability and suitability for a range of agricultural environments.[54] Sladojevic et al. focused on model generalisation, which enabled it to handle a wider range of scenarios with less species-specific adjustments, even though they achieved accuracy comparable to Mohanty et al.

By training a model on a dataset containing over 87,000 photos and 25 distinct plant species and 58 illnesses, Ferentinos (2018) increased the scalability of CNNs. With a 99.53% accuracy rate, the model proved that deep learning models could be scaled up to handle a wide range of crops and illnesses with good accuracy. In contrast to previous research, Ferentinos's work demonstrated CNNs' resilience when used on bigger and more varied datasets.[21]

By contrasting the effectiveness of many CNN architectures, including as AlexNet, GoogLeNet, and LeNet, for the job of plant disease identification in tomato leaves, Brahimi et al. (2017) offered an alternative viewpoint. With a 99.3% accuracy rate, they discovered that GoogLeNet performed better than the other models. The comparative analysis played a crucial role in helping to determine which network topologies were best for particular tasks. This aspect was not thoroughly examined in the earlier studies, which were mainly concerned with the performance of individual models.[11]

Comparing both research, it can be seen that Sladojevic et al. (2016) and Mohanty

et al. (2016) showed how effective CNNs can be in detecting diseases, but with different emphasis—Sladojevic on generalisation across various species, while Mohanty focused on accuracy with a particular dataset. By demonstrating that CNNs could handle bigger and more varied datasets without sacrificing accuracy, Ferentinos (2018) built on earlier efforts. In the meanwhile, Brahimi et al. (2017) offered helpful insights on the efficacy of various CNN architectures as well as suggestions for model selection. Lastly, Too et al. (2019) addressed one of the primary issues in the field by demonstrating that transfer learning may maintain high accuracy with smaller datasets, adding a practical perspective to the conversation.

2.3.1.2 Ground Multispectral Images:

Multispectral imaging captures data across multiple wavelengths, including near-infrared (NIR) bands, making it superior to RGB in detecting physiological changes before visible symptoms appear. Zhang et al. (2020) utilized multispectral imaging to monitor wheat yellow rust, showcasing the early detection capabilities of NIR bands.[61] Mishra et al. (2020) applied multispectral imaging to detect maize gray leaf spot, finding that combining spectral bands improved early detection accuracy.[41] Hunt et al. (2018) used multispectral imagery to detect Phytophthora infestans in potatoes, emphasizing the utility of red-edge and NIR bands in soil-borne disease detection.[27] Sankaran et al. (2017) demonstrated the use of multispectral data in detecting soybean rust, further validating the superiority of this modality for early stress detection.[50] Zhou et al. (2021) successfully used multispectral data to monitor bacterial leaf blight in rice, highlighting its effectiveness across diverse crops.[65]

2.3.1.3 UAV/Aerial Images (RGB and NDVI):

Over the past decade, there has been a considerable advancement in the use of Unmanned Aerial Vehicles (UAVs) for crop disease diagnosis. These UAVs use a variety of imaging modalities, such as RGB, multispectral, and NDVI (Normalised Difference Vegetation Index) sensors. These technologies, which provide both geographical and temporal resolution that exceeds conventional ground-based methods, have made it possible to monitor crop health in great detail and on a vast scale.[52]

RGB Imaging

RGB sensors are a foundational tool in UAV-based crop disease detection, capturing high-resolution visible light imagery crucial for identifying symptoms such as leaf

discoloration and lesions. Calderón et al. (2015) utilized UAV RGB imagery to detect Verticillium wilt in olive orchards, highlighting its effectiveness in mapping disease spread over large areas.[13] Similarly, Bhandari et al. (2020) combined RGB imagery with machine learning to classify leaf spot diseases in wheat, emphasizing the integration of advanced analysis techniques for enhanced accuracy.[10] Lowe et al. (2017)extended the use of RGB imagery to monitor grapevine leafroll-associated virus-3, demonstrating the capability of RGB sensors in tracking disease progression over time.[35] De Souza et al. (2017) focused on early detection of coffee leaf rust using RGB imagery, stressing the importance of high-resolution images in identifying small lesions.[16] Finally, Nigon et al. (2015) used RGB imagery to detect Cercospora leaf spot in sugar beet, proving its utility in large-scale monitoring.[43]

NDVI Imaging

NDVI, which measures vegetation vigor using the red and NIR bands, is widely used in UAV-based crop monitoring. Zhao et al. (2018) demonstrated NDVI's effectiveness in distinguishing healthy and infected rice plants across large fields.[62] Jiang et al. (2019) combined NDVI with other spectral indices to enhance the detection of Northern Leaf Blight in maize, showing that integrating NDVI with other indices improves detection accuracy.[29] Thorp et al. (2017)used NDVI to map rust in wheat, confirming its utility in large-scale disease monitoring.[56] Liu et al. (2018) applied NDVI in monitoring late blight in potato fields, effectively highlighting areas of reduced vigor due to disease.[34] Zheng et al. (2021) combined NDVI with multispectral imagery to improve wheat disease detection and mapping, demonstrating the advantages of multimodal data integration.[63]

2.3.2 Single Modal in Crop Disease Detection: Tabular Data

2.3.2.1 Using Machine Learning to Detect Agricultural Diseases

The agriculture industry has made substantial use of machine learning, especially gradient boosting algorithms like XGBoost and tree-based methods like Random Forest (RF), for disease diagnosis. A comprehensive assessment of machine learning applications in agriculture was presented by Kamilaris and Prenafeta-Boldú (2018), who highlighted the models' interpretability and resilience when processing tabular data. Their research showed that by utilising a range of agronomic factors, such as weather, soil characteristics, and crop management techniques, these models perform exceptionally well in tasks like yield prediction and disease outbreak forecasting.[30]

In order to estimate wheat and barley yields, Richetti et al. (2023) used RF and XGBoost, demonstrating a practical use of these models. Their dataset comprised vegetation indicators and environmental parameters. The results demonstrated the effectiveness of these models, particularly in situations where computational resources are scarce and model interpretability is essential. Their findings demonstrate the usefulness of these models in agricultural applications, since they outperformed more intricate deep learning architectures.[47]

Similar to this, Mahlein et al. (2018) used spectrum data to identify and categorise agricultural illnesses, demonstrating the value of RF in plant pathology. According to their research, RF is a useful tool in precision agriculture since it can handle high-dimensional data efficiently and provide accurate disease categorisation.[37]

2.3.2.2 Deep Learning Approaches to Tabular Data

The potential of deep learning architectures to capture intricate, non-linear correlations in massive datasets is attracting more and more attention, even if standard machine learning models have demonstrated considerable promise. This trend is best illustrated by TabNet, which Arsenio et al. (2020) introduced. Their research on the disease prediction of tomato plants showed that TabNet might perform better than traditional machine learning models such as Support Vector Machines (SVM) and Gradient Boosting Machines. Because TabNet's architecture incorporates feature selection, it can concentrate on the most important parts of the data, which improves its efficiency in handling large, complicated agricultural datasets.[3]

Multi-Layer Perceptrons (MLPs), a type of deep learning model, are becoming more and more popular because of their capacity to represent complex patterns seen in tabular data. But as Richetti et al. pointed out, these networks' architecture and depth have a big impact on how well they work. They discovered that models with larger layers frequently perform better by capturing broader associations across features, even while deeper networks do not always beat shallower ones. This implies that even while deep learning models have a lot of promise, applying them will need careful consideration of the unique properties of the data as well as computing limitations.[47]

Kruse et al. (2022) investigated the use of convolutional neural networks (CNNs) to the processing of tabular data in agriculture in a different study. They discovered that CNNs, which are often used for image data, could be modified to handle tabular data, especially when it came to spotting patterns associated with the onset of sickness. Their research demonstrated how deep learning models are adaptable to many data modalities,

which makes them useful instruments for agricultural research.[32]

2.3.3 Multimodal in Crop Disease Detection

To improve the precision and promptness of crop disease identification, multimodal techniques integrate multiple sensor modalities such as RGB, multispectral, thermal, and hyperspectral imaging.

According to Maimaitijiang et al. (2020), merging RGB and hyperspectral imagery from UAVs enhances yield prediction models and the early diagnosis of soybean illnesses.[38] Using NDVI and multispectral data, Zhao et al. (2018) show that their multimodal technique enables for reliable identification of disease transmission across wide fields in their monitoring of rice blast disease.[62] To identify rice sheath blight, Duan et al. (2019) use deep learning models in conjunction with RGB and multispectral data obtained by UAVs, highlighting the enhanced precision and resilience of disease classification using multimodal data fusion. [18]

In order to monitor grapevine leafroll-associated virus-3 in vineyards, Lowe et al. (2017) combine RGB imaging with thermal data, offering a thorough understanding of the disease's influence on crop physiology.[35] In order to monitor bacterial leaf blight in rice, Zhou et al. (2021) combine multispectral data with NDVI. This allows for accurate mapping of disease severity and focused treatments in broad fields.[64]

Subsequent research highlights the significance of multimodal data in improving the precision of disease identification. Citrus greening disease detection with the combination of hyperspectral and thermal photography is explored by Sankaran et al. (2017), highlighting the advantages of collecting spectral and physiological changes in infected trees.[51] Rumpf et al. (2010) demonstrate the possibility of multimodal imaging in capturing minor physiological changes by combining RGB and fluorescence imaging to detect early symptoms of sugar beetroot illnesses.[49] In order to monitor wheat rust, Guan et al. (2021) combine RGB, multispectral, and LiDAR data. They show that multimodal data improves spatial and temporal resolution, which is important for early intervention.[24] Polder et al. (2017) demonstrate that integrating visible and NIR data enhances disease detection accuracy under different light circumstances.[45] They do this by using RGB and near-infrared (NIR) imaging to identify late blight in potato crops. In order to detect Fusarium head blight in wheat, Meyer et al. (2019) investigate the integration of UAV-based hyperspectral and thermal data, emphasising the advantages of multimodal approaches in capturing both temperature and reflectance

variations associated with the disease.[40]

2.4 Gaps in Existing Research

Although multimodal techniques for crop monitoring and disease detection have made great strides, there are still a number of gaps that need to be filled in order to properly utilise these technologies in precision agriculture.

The inadequate standardisation and integration of multimodal data is one significant gap. Despite a wealth of research highlighting the advantages of merging disparate data sources, standardised techniques to efficiently integrate and handle multimodal data remain lacking. The majority of previous studies concentrate on particular combinations of modalities, frequently customised for certain crops or environmental circumstances, which restricts the applicability of their conclusions. More research and development are still needed to create data fusion methods that are broadly applicable and adaptable to different agricultural situations [19][31].

Real-time processing and scalability present another major challenge. It has been challenging to apply multimodal techniques at bigger scales, like entire farms or agricultural areas. Many research are carried out in controlled situations or on comparatively small plots, which may not transfer well to bigger, more varied agricultural environments. Furthermore, the high computing needs of processing big multimodal datasets frequently limit real-time data processing and analysis—which is essential for timely disease detection in precision agriculture. This restriction prevents these technologies from being used more widely in practice [31][38].

In order to advance precision agriculture and guarantee that multimodal techniques can be successfully applied at scale, which will result in more sustainable and effective farming practices, these gaps must be filled.

Chapter 3

Methodology

In order to develop a multimodal disease detection system for peach trees, this study combines data from ground multispectral imagery, ground RGB imagery, and UAV imagery. Tabular data was also extracted from the ground multispectral imagery and UAV imagery. Advanced ML algorithms, such as a multimodal deep learning model that incorporates convolutional neural networks (CNNs) for image data and MLP for tabular data, were used to handle the data. In this chapter we provide the full methodology that was done throughout this research.

3.1 Data Collection

Peach Tree Disease Detection Dataset from the University of Western Macedonia is a comprehensive[14], multimodal resource for precision agriculture research, with an emphasis on the identification and analysis of stress brought on by diseases like *Grapholita molesta* and *Anarsia lineatella* in peach trees. The dataset includes 889 peach trees that were carefully documented over the course of a full production season, which runs from July 2021 to September 2022. Numerous imaging modalities, including ground-based RGB images, aerial UAV images, and ground multispectral images, are included in this recording. The trees are classified into four distinct groups according to their state of health: *healthy* (the trees do not contain any of the other two diseases but contain water stress), *Grapholita molesta* (affecting tree trunks and branches), *Anarsia lineatella* (presence of withered tops on branches), and *Dead Trees* (trees that died of disease). Accompanying accurate tree mapping inside the orchard, the dataset's structure allows for a thorough investigation of the correlation between environmental conditions and disease occurrence.[14][36][46]

3.2 Manual Annotation of Bounding Boxes

At first, the image as a whole was used, however this method produced subpar results. For the majority of the epochs, the training and validation accuracy were constant, suggesting that the model was only guessing and not learning efficiently. This was probably caused by a class imbalance, which by coincidence made the model forecast accurately the majority of the time. We reasoned that employing bounding boxes might enhance classification performance as a solution.

Bounding boxes for ground RGB, multispectral, and UAV images were made manually. By precisely defining the classes of the data within the images, these bounding boxes ensured proper mapping and analysis. Following their integration into the dataset, these bounding boxes provided each tree with comprehensive spatial information across many image modalities.

The bounding boxes for the ground RGB imagery and the ground multispectral imagery was done to identify each class according to the description provide in the Read Me file for the dataset. Bounding box annotations were stored in JSON files and contained the class name and each bounding box's coordinates.

For the UAV images bounding boxes where created around each tree in the orchard so as to link the tree with the rest of its images in the ground RGB imagery and grounf multispectral imagery. Figure 3.2 shows a sample of the final result. The coordinates of each bounding box and the associated tree ID were included in the bounding box annotations, which were kept in JSON files.

3.3 Data Preprocessing

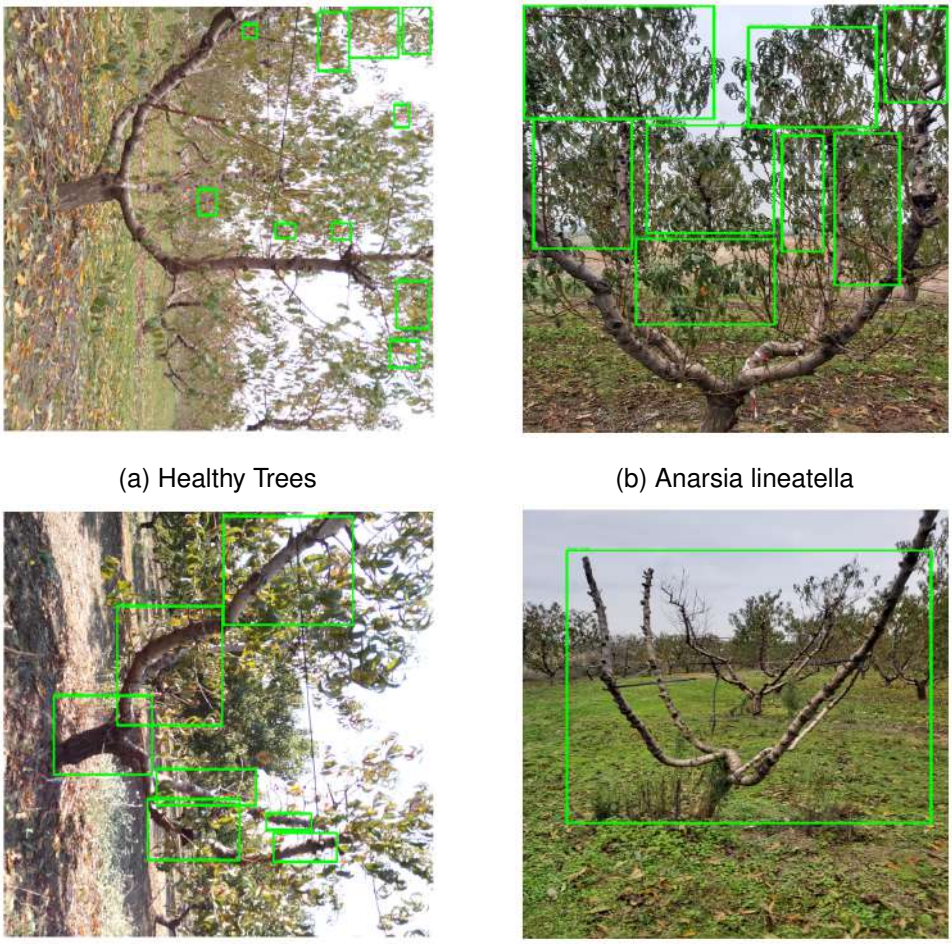
The preprocessing steps done to the data are as follows:

1. **Data Preparation**

   In order to prepare the data, a number of datasets about the health of trees in orchards had to be loaded and organised. Tree mapping data, image metadata, and multispectral images are among the data gathered across a number of dates.

   - Tree Mapping Data: The data for the tree mapping was loaded into different DataFrames. Understanding the arrangement and structure of the orchard, and the condition of each individual tree, required knowledge of these data.

   - Label Data: A CSV file with labels for every tree was loaded. The trees' health state was indicated by these labels, and this information was subsequently correlated with the image data.



(a) Healthy Trees

(b) Anarsia lineatella

(c) Grapholita molesta

(d) Dead Trees

Figure 3.1: Data Classes with Bounding Boxes

2. **Image Metadata and Loading**

   Images of various kinds were taken from several directories, such as:

   - UAV Images: Several spectral bands, including red, green, and near-infrared (NIR), were among them.

   - Ground RGB Images: Standard RGB images with their bounding box annotations captured from the ground.

   - Multispectral Images: Red, green, NIR, and red-edge bands were present in these images.

3. **Image-Tree Mapping**

   A function was created to extract tree IDs from the filenames in order to associate the images with certain trees. By mapping images to particular trees using their IDs, dictionaries connecting each tree to its matching image type—RGB,

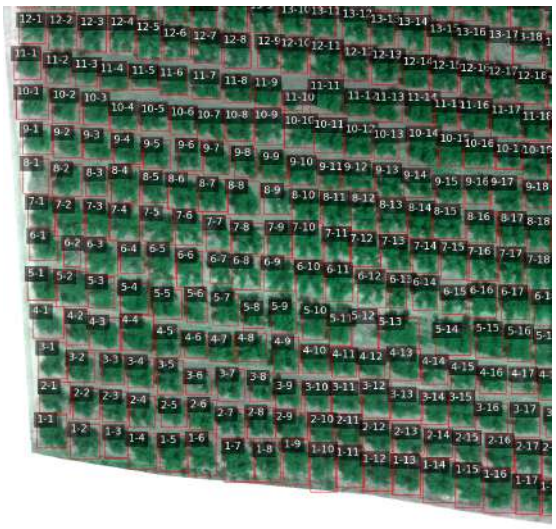


Figure 3.2: Part of RGB UAV with Bounding Boxes

multispectral, etc.—were created. With this phase, the accuracy of each tree's data representation across all modalities was guaranteed.

4. **Data Transformation**

   - Image Resizing: To guarantee consistent input dimensions for models, the image size was standardised using the transform feature in the torch library. The images were resizes to 224x224 pixels.

   - Normalization: The pixel values were normalized for images to fall within a specific range (e.g., 0-1 for RGB values). This was also done using the transform feature.

5. **Data Cleaning and Filtering**

   The following actions were done to make sure the dataset was clean and prepared for additional processing:

   - Data Filtering: To concentrate exclusively on trees with recognised health statuses, rows with labels equal to 0 were eliminated.

   - Validation of Image Paths: To ensure that there were no erroneous or missing picture data, the dataset was filtered so that only rows containing valid strings for each image path were included.

   - Removal of Non-Numeric Columns: Non-numeric columns were found and removed from the feature columns that were used in the analysis that followed.

6. **Splitting Data**
   Three sets were created from the cleaned dataset: test, validation, and training. The information was split as follows: 60% of the data in the training set, 20% of the data is the validation set and 20% of the data is the test set. In order to prepare the data for model training and evaluation, this split was done straight from the CSV file.

3.4 Model Definition

3.4.1 Feature Extraction

There are various tabular and image feature extractors used, including as convolutional networks and MLP. The following steps were performed for feature extraction:

1. **Creation of Multimodal Dataset**
   By merging tree IDs, matching images, bounding box annotations, and health status labels, a comprehensive multimodal dataset was produced. This dataset contained:
   - Date of Data Collection
   - Tree ID
   - Paths to Various Image Types: Orchard mapping image, UAV image (Red, Green, NIR, NDVI), ground RGB image, and multispectral images(Red, Green, NIR, NDVI, RGB).
   - Manually Created Bounding Boxes: For ground RGB and multispectral images.
   -Tree Health Status Labels

2. **Image Annotation and Visualization**
   The multispectral images and ground-based trees were precisely surrounded by rectangles drawn using the bounding boxes that were manually generated. Labels identifying the health condition or other pertinent information were attached to these boundary boxes. After that, the annotated images were shown for close examination. Furthermore, a technique was created to visualise the tree health status based on the CSV data by plotting the orchard mapping grid using the UAV images and the tree mapping csv file which contained the tree mapping in terms of row-column (Figure 3.2 shows a zoomed in version and Figure A.1 shows the full final result where the bounding boxes are named with the same names of the

ground RGB and multispectral images). The trees which were represented as circles were colour-coded based on their class, making it easier to examine the orchard visually.

3. **Vegetation Index Calculation**
   The health and vigour of plants are measured numerically using reflectance measurements in various spectral bands to create vegetation indices. They are extensively employed in precision agriculture and remote sensing to track crop status, identify stress, and direct management choices. The vegetation indicators calculated are NDVI, GNDVI, EVI, NDRE, RVI. The benefits, computations, and explanations for a number of important vegetation indicators are provided in appendix A.2
   Vegetation indices were computed for every tree using the UAV photos and bounding boxes that were created in the preceding stage. However, since the multispectral images already had this data for every image, bounding boxes were not necessary for direct computation.

4. **Data Integration and Output**
   The multimodal dataset contained the calculated vegetation indices in addition to the coordinates of the trees in the images. Colour mapping was used to display the NDVI images, and the final photos were saved. The resultant multimodal dataset was improved with bounding box annotations, vegetation indices, and image routes. It was then saved as a CSV file for further analysis. This dataset serves as the basis for the examination of the orchard's tree health that follows.

5. **Image Feature Extraction:**
   As the image feature extractor, we investigate various cutting-edge convolutional neural network (CNN) designs. These models are made up of 4 benchmark models and 2 suggested models:

   - **InceptionV3 (Benchmark):** To address significant image changes, a deep CNN with auxiliary classifiers and factorised convolutions is used.[55]
   - **ResNet152 (Benchmark):** A deep residual network that can be trained to very deep levels by using skip connections to lessen the vanishing gradient issue.[26]
   - **VGG19 (Benchmark):** A more straightforward CNN architecture consisting of 19 layers, distinguished by its uniform arrangement of tiny convolu-

tion filters.[53]

- **Vision Transformer (ViT) (Benchmark):** This transformer-based architecture uses self-attention methods to capture long-range dependencies in pictures by treating image patches as tokens.[15]
- **Attention-Augmented Convolutional Networks (Suggested):** To enable the network to concentrate on the most pertinent areas of the image, we additionally include attention methods into the classic CNN architectures (ResNet and Inception).[8] The base for the models was taken from a github code[23] and are as follows (The architectures are shown in Figure 3.3):

  (a) **Attention-Augmented Convolutional Layer (AACN Layer)**
  The Attention-Augmented Convolutional Layer (AACN Layer), which augments or substitutes conventional convolutional layers in conventional CNN architectures, is the central novelty in our models. By introducing an attention mechanism, the AACN Layer enhances feature representation and overall model performance by enabling the network to dynamically focus on significant input regions.

  i. **Structure and Parameters:**
  - **Input Channels**: The AACN Layer can receive an input tensor with a predefined number of channels.
  - **Attention Parameters:** The key (k) and value (v) dimensions, which parameterise the layer, regulate the size of the key and value tensors of the attention mechanism in respect to the number of input channels.
  - **Multi-Head Focus:** We employ multi-head attention with a given number of heads (num_heads) to assist the model in capturing a range of input feature attributes.
  - **Size of Kernel:** The AACN Layer maintains a mutable kernel size to guarantee adherence to the spatial dimensions of the present architecture.

  ii. Attention Mechanism:
  The attention mechanism in the AACN Layer computes attention scores by taking the dot product between the query and key tensors. These scores are then scaled and normalised to yield the weighted sum of the value tensors. Through this process, the model is able

to identify and prioritise significant geographical locations in the feature maps. The attention-augmented output and the traditional convolutional output are then combined to generate the layer's final output.

(b) **Attention-Augmented ResNet18**

The first model we propose is the Attention-Augmented ResNet18, which is derived from the widely used ResNet18 architecture.[8]

i. **Integration of AACN Layers:** Certain convolutional layers of the original ResNet18 are swapped out for AACN Layers in this architecture. The replacement process is executed with precision to guarantee the preservation of the residual connections, which are essential to the ResNet architecture.

ii. **Model Training:** Standard backpropagation is used to train the Attention-Augmented ResNet18, and the optimisation methods and loss function are the same as those used to train conventional ResNet models. The addition of attention mechanisms enhances the model's ability to learn discriminative features without changing the training process as a whole.

(c) **Attention-Augmented InceptionV3** The second model developed in this study is the Attention-Augmented InceptionV3, based on the InceptionV3 architecture.

i. **Modification of Inception Modules:** AACN Layers are added to the Inception modules in this design, which are distinguished by parallel convolutions with different kernel sizes. These modules incorporate the attention mechanism, which enables the network to concurrently respond to pertinent information at several scales.

ii. **Model Architecture:** The output of each Inception module combines attention-augmented and conventional convolutional outputs. The network can incorporate spatial attention and retain the multi-scale feature extraction capabilities of InceptionV3 thanks to this dual-pathway architecture.

iii. **Training and Optimization:** The InceptionV3 variation is trained using standard deep learning frameworks and methodologies, just like the Attention-Augmented ResNet18. The model performs

better on tasks involving complicated, multi-scale patterns because of the superior concentration that attention provides.

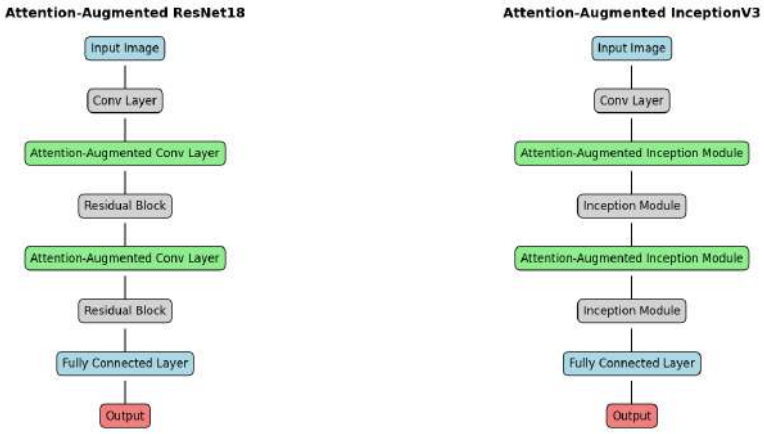


Figure 3.3: Attention-Augmented ResNet18 and Attention-Augmented InceptionV3 Architectures

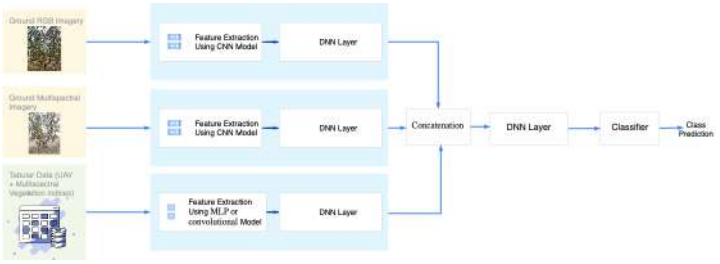
To enable the network to output feature embeddings instead of classification probabilities, the last classification layer is removed from each picture feature extractor. The next fusion layer receives these embeddings after that.

6. **Tabular Data Feature Extraction:**
   We examine several designs designed to extract significant features from numerical data, taking into account the tabular format of the CSV data:

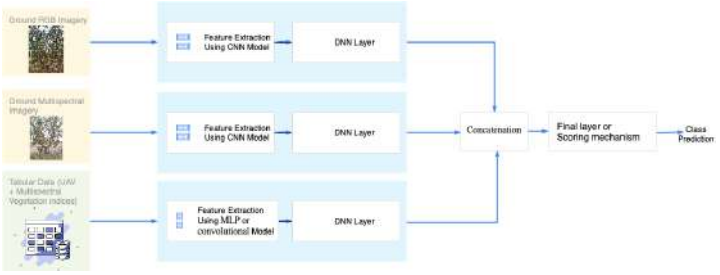
   - Multilayer Perceptron (MLP): This deeply neural network model uses a fully linked deep neural network at several depths: small (50 layers) and deep (100 layers). The purpose of the MLP networks is to identify non-linear correlations in the tabular data.[2]
   - Convolutional Neural Network for Tabular Data: We use one-dimensional convolutions over the tabular data. To determine how well different configurations—from small (50 layers) to deep (100 layers) networks—capture spatially-like dependencies between features are examined.[12]

3.4.2 Multimodal Fusion

We apply and assess two different fusion algorithms to combine the features extracted from both modalities:



(a) Intermediate Fusion



(b) Late Fusion

Figure 3.4: Fusion Methods

- Late Fusion: This technique joins features from the image and CSV data right before the last layer of classification. When working with several types of information, it is helpful that each modality can separately acquire its representation (Figure 3.4b).

- intermediary Fusion: Here, before being fused with the CSV data, the image characteristics are analysed via an intermediate fully connected layer, allowing for some interaction between modalities prior to the final conclusion (Figure 3.4a).

3.4.3 Classifier

Our architecture's last classifier is a fully linked layer that generates probabilities for every class. The number of classes in the dataset determines the output dimension of this layer. To normalise the output into probability distributions over the potential classes, we employ the softmax function.

3.5 Model Training and Validation

The model was trained using a cross-entropy loss function and the Adam optimizer. Early stopping was employed to prevent overfitting, monitoring the validation loss to determine when to stop training. The model's parameters are modified during training in order to reduce loss, and each epoch's training and validation accuracy is monitored.

Chapter 4

Results and Discussion

The findings in this chapter offer a thorough examination of the efficacy of different models used to identify diseases in peach trees utilising both tabular (CSV) and image-based (CNN) data. In order to increase classification performance, the study investigates two types of approaches: single-modal, in which models process image or tabular data individually, and multimodal, in which data from both modalities is combined.

The test set was used to evaluate the trained models in terms of how well they predicted the health status of the trees. A number of metrics, including F1-score, recall, accuracy, and precision, were calculated to assess how effective the multimodal strategy was and confusion matrices are generated to analyze the model's predictions in detail.

4.1 Single-Modal Analysis

In the methodology section, we independently trained each modality using the outlined models. The primary variable across all studies was batch size, while the model settings were largely consistent. We experimented with batch sizes of 8, 16, 32, 64, and 128, finding that a batch size of 32 yielded the best results.

4.1.1 Ground RGB Imagery Imagery

This section provides a thorough examination of the findings from the training and assessment of the models for the Ground RGB Imagery imagery's diseased crop and tree health classification.

4.1.1.1 Model Performance Analysis

Each model's forecasting ability across the dataset is summarized by overall accuracy and loss measures. The models' performance, including train, validation, and test accuracies and losses, is detailed in the tables. 4.1.

Model	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss	Test Accuracy	Test Loss
InceptionV3	100.00%	0.2438	99.22%	0.4343	95.35%	0.1395
ResNet152	100.00%	0.0023	99.22%	0.5954	97.67%	0.1408
VGG19	100.00%	0.0005	98.45%	1.5904	97.67%	0.1220
ViT	99.12%	0.1480	89.12%	2.5412	88.37%	0.5242
AttentionAugmentedInceptionV3	100.00%	0.1034	95.12%	1.5402	93.02%	0.6157
AttentionAugmentedResNet18	98.45%	0.4324	86.05%	1.9324	81.40%	0.6790

Table 4.1: Ground RGB Imagery Model Performance Summary

As shown in Table 4.1, ResNet152 and VGG19 achieved the highest accuracy, with InceptionV3 close behind, making these models the top performers for this classification task. The deep architectures of ResNet152 and VGG19 contributed to their strong performance, with ResNet152 excelling in capturing complex patterns. InceptionV3 also performed well, demonstrating its robustness with balanced validation and test accuracies. On the other hand, ViT showed lower accuracy, likely due to its architecture being more suited for larger datasets. The attention-augmented models (AttentionAugmentedResNet18 and AttentionAugmentedInceptionV3) performed reasonably well, but did not surpass the top models. Recurring high validation accuracy with decreasing training loss raises the possibility of overfitting in some models, particularly the deeper models such as ResNet152 and attention-augmented variations.

4.1.1.2 Precision, Recall, and F1-Score Analysis

The precision, recall, and F1-score measures are used in this section to provide a deeper understanding of how each model deals with the different classes. These measurements are summarised in Table 4.2.

Model-wise Analysis: InceptionV3 exhibits a well-balanced performance, resulting in a robust F1-score and steady accuracy over various datasets. With its well-represented classes, in particular, its architecture is well-suited for this categorisation task. ResNet152 does well because of its residual connections and deep design, which

Model	Precision	Recall	F1-score
InceptionV3	0.714286	0.687500	0.695055
ResNet152	0.730769	0.750000	0.740000
VGG19	0.730769	0.750000	0.740000
ViT	0.608974	0.689103	0.640000
AttentionAugmentedInceptionV3	0.709416	0.708333	0.708132
AttentionAugmentedResNet18	0.389881	0.448718	0.416667

Table 4.2: Ground RGB Imagery Precision, Recall and F1-Scores

enable it to efficiently capture complex patterns. Its depth does, however, raise the possibility of overfitting, which could result in somewhat less accuracy in more difficult tasks. Although efficient, VGG19 is less able to handle unbalanced datasets than ResNet152 due to its more basic methodology and absence of sophisticated characteristics like residual connections. Because ViT relies on large-scale data for training, it performs poorly on smaller datasets despite its capacity to capture long-range connections, leading to lower accuracy and F1-scores. By integrating attention processes, AttentionAugmentedInceptionV3 model manages to beat the base InceptionV3 on more difficult tasks, while it still has issues with overfitting and class imbalance. By incorporating attention mechanisms, AttentionAugmentedResNet18 model builds upon ResNet152, although it has the same difficulties with under-represented classes and possible overfitting.

Class-wise Performance: Regarding specific classes, the models generally performed better on the more prevalent classes but struggled with those that were under-represented. The classification report and confusion matrix for each model are available in Appendix B.1.1. Additionally, a sample comparing true labels with predicted labels, along with their confidence levels, is shown in Figure 4.1. The following observations were made:

- **Anarsia lineatella:** Most models demonstrated high recall and precision for this class, indicating that it is well-represented in the dataset and relatively easy to classify.
- **Grapholita molesta:** Similar to Anarsia lineatella, this class was accurately classified by all models, with InceptionV3 and ResNet152 performing particularly well.
- **Dead Trees:** All models showed zero precision and recall for this class, as none could correctly predict any samples. This suggests that the models may

have become biased towards the more common classes, neglecting this under-represented class. Such class imbalance often leads to poor performance on minority classes in machine learning models.

- **Healthy:** The Healthy class was frequently misclassified across all models, likely because it shares similar leaf patterns with other classes in the dataset, making it difficult to distinguish.

Figure 4.1: Incorrect Predictions of Ground RGB Imagery for Attention Augmented Resnet18 with its confidence level

The majority of models, especially ResNet152, VGG19, and InceptionV3, have high accuracy, which suggests good generalisation to the test set for the main classes. Nonetheless, the persistent challenge of class imbalance is highlighted by the persistent difficulty in correctly anticipating the Dead Trees class.

4.1.2 Ground Multispectral Imagery

The results of the training and evaluation of the models for the diseased crop and tree health categorisation using Ground Multispectral Imagery are thoroughly examined in this part.

4.1.2.1 Model Performance Analysis

The accuracy and loss values of the models during training, validation, and testing provide an overview of their performance on the multispectral imagery dataset. An overview of these metrics for each model is given in Table 4.3.

The best results are obtained by ResNet152 and InceptionV3, which show that they are resilient and flexible enough to handle the multispectral images dataset. ViT and VGG19 perform less well than expected; in particular, VGG19 has serious problems with generalisation, as seen by its low test accuracy and validation. While the

Model	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss	Test Accuracy	Test Loss
InceptionV3	100.00%	0.0074	94.74%	0.0873	89.47%	0.2173
ResNet152	100.00%	0.0076	97.37%	0.0544	97.37%	0.0785
VGG19	100.00%	0.0097	84.21%	0.4482	68.42%	1.5892
ViT	92.04%	0.2184	100.00%	0.0794	81.58%	0.2944
AttentionAugmentedInceptionV3	100.00%	0.0175	92.11%	0.2665	92.11%	0.2907
AttentionAugmentedResNet18	99.12%	0.0655	84.21%	0.6973	92.11%	0.2066

Table 4.3: Ground Multispectral Imagery Model Performance Summary

AttentionAugmented models perform passably, they share several flaws with the non-augmented models, especially when it comes to managing class disparities. Similar to the ground RGB imagery, recurring high validation accuracy with decreasing training loss raises the possibility of overfitting in some models.

4.1.2.2 Precision, Recall, and F1-Score Analysis

The precision, recall, and F1-score metrics provide deeper insights into how well each model handles different classes. Table 4.4 summarizes these metrics.

Model	Precision	Recall	F1-score
InceptionV3	0.641009	0.669643	0.654881
ResNet152	0.983333	0.875000	0.908046
VGG19	0.788889	0.698413	0.655520
ViT	0.654386	0.676587	0.659091
AttentionAugmentedInceptionV3	0.705592	0.687500	0.690862
AttentionAugmentedResNet18	0.666667	0.732143	0.696360

Table 4.4: Ground Multispectral Imagery Precision, Recall and F1-Score

Model-wise Analysis: ResNet152 excels in both precision and recall, making it the top performer for this dataset. Its ability to accurately classify data across all classes, including the less common ones, is reflected in its strong F1-scores, establishing it as a reliable model. InceptionV3 follows closely, offering a well-balanced performance, though slightly less accurate than ResNet152. On the other hand, ViT and VGG19 struggle with consistency due to their sensitivity to class imbalances and architectural design. While the AttentionAugmented models show some improvement in recall, they still face challenges similar to their base models, particularly in precision.

Class-wise Performance: The following findings can be drawn from the models' performance in each class. A sample comparing true labels with predicted labels, along with their confidence levels, is shown in Figure 4.2):

- **Anarsia lineatella:** This class was well-represented in the dataset, as evidenced by the strong performance of all models. Models like AttentionAugmentedInceptionV3 and ResNet152 demonstrated near-perfect precision and recall.
- **Grapholita molesta:** Most models performed well on this class, with particularly good recall and precision values from ResNet152 and InceptionV3.
- **Dead Trees:** Several models encountered difficulties with this class, with varying degrees of success. ResNet152 and InceptionV3 achieved strong recall, but ViT and VGG19 struggled.
- **Healthy:** This class proved to be the most challenging, with some models, including the AttentionAugmented and InceptionV3 models, failing to accurately predict any samples. This suggests a significant bias towards the more common classes, contributing to the low performance in this under-represented class.

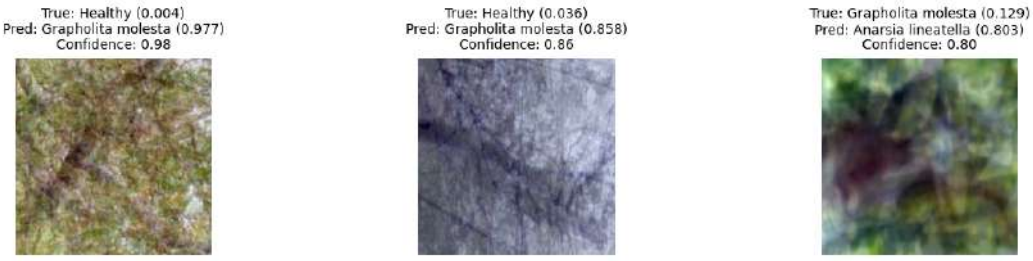


Figure 4.2: Incorrect Predictions of Ground Multispectral Imagery for Attention Augmented Resnet18 with its confidence level

Confusion Matrix Analysis: The confusion matrices offer a graphic depiction of the model's predictions, showing the areas in which the models were successful and unsuccessful. The following are the main findings from the confusion matrices the full results of the confusion matrix and the classification report can be found in Appendix B.1.2:

- **ResNet152 and InceptionV3:** These models show a high degree of confidence in their predictions, especially for classes that are well-represented, such as

Dead Trees and Anarsia lineatella. The less represented classes do show some perplexity, while ResNet152 does better overall in handling these.

- **VGG19 and ViT:** These models exhibit a great deal of ambiguity, especially when it comes to the categorisation of healthy samples. VGG19 performs worse overall since it tends to misclassify several samples in various classes.
- **AttentionAugmented Models:** Although these models handle some classes better than before, they still have perplexity, especially when it comes to the Healthy class. While they can be helpful, the attention mechanisms are not a perfect solution to the problems caused by class disparities.

For this multispectral photography dataset, ResNet152 is by far the most dependable model overall, with InceptionV3 coming in second. Even if they are capable, the other models suffer from class imbalances and need to be fine-tuned or have their architecture changed in order to be more generalisable.

4.1.3 Tabular Data

4.1.3.1 Model Performance Analysis

The models' accuracy and loss numbers during training, validation, and testing provide an overview of how well they performed on the classification job. For every model, these metrics are summarised in Table 4.5.

Model	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss	Test Accuracy	Test Loss
mlp_small	96.25%	0.1955	96.15%	0.1917	95.03%	0.2405
mlp_medium	96.25%	5.5452	96.15%	5.5452	95.03%	5.5452
conv_small	96.25%	5.5452	96.15%	5.5452	95.03%	5.5452
conv_medium	96.25%	0.3185	96.15%	0.3222	95.03%	0.3537

Table 4.5: Model Performance Summary

In training, validation, and testing, the models show a high degree of accuracy, which is mostly due to the dominating class in the dataset. Both the mlp_medium and conv_small models show problems, especially during testing and validation, when the loss stays around 5.5452, suggesting insufficient meaningful learning. Comparatively speaking, the conv_medium model performs better in loss reduction; but, as test accuracies reveal, all models eventually have difficulty with generalisation.

4.1.3.2 Precision, Recall, and F1-Score Analysis

More detailed information about the models' performance, particularly their effectiveness in handling various classes, can be found in the precision, recall, and F1-score measures. These metrics are summarized in Table 4.6. **Model-wise Analysis:** Each

Model	Precision	Recall	F1-score
mlp_small	0.23757	0.25	0.243627
mlp_medium	0.23757	0.25	0.243627
conv_small	0.23757	0.25	0.243627
conv_medium	0.23757	0.25	0.243627

Table 4.6: Tabular Data Precision, Recall and F1-Score

model shows consistent precision, recall, and F1-scores across all classes, revealing a common pattern in their performance. These metrics underscore a significant issue with predicting minority classes, where the models consistently fall short of delivering useful results. Although the models perform well in predicting the dominant class (Anarsia lineatella), they struggle to generalize to less frequent classes. The imbalance between recall and precision is further highlighted by the poor F1-scores, which emphasize the models' difficulties in addressing class imbalances.

Class-wise Performance: The following findings can be drawn from looking at how well the models perform in each class:

- **Anarsia lineatella:** Given its dominance in the dataset, all models do remarkably well on this class. Strong confidence and accuracy are indicated by the models' constant near-perfect precision and recall for this class.
- **Grapholita molesta, Dead Trees, and Healthy:** With precision, recall, and F1-scores of 0.00, these classes present formidable obstacles to all models. The models' incapacity to generalise to the under-represented classes and their high bias towards the Anarsia lineatella class are indicated by their failure to properly predict any instances of these classes.

Confusion Matrix Analysis: The models' predictions are illustrated through confusion matrices, which highlight areas of success and failure. The following are key observations from the confusion matrices:

- **Dominant Class (Anarsia lineatella):** The confusion matrices show high values along the diagonal corresponding to this class, indicating that the models accurately predict the dominant class.

- **Minority Classes:** The models exhibit significant uncertainty when predicting the other classes. The low precision and recall metrics are reflected in the off-diagonal elements, which reveal that predictions for these classes are either completely absent or incorrectly classified as the dominant class.

Overall, the dataset's class imbalance has a significant impact on the models. Although they excel in the dominating class, they have very little ability to generalise to other classes.

4.2 Multimodal Analysis

This section examines how well the models perform with tabular data, multispectral images, and ground RGB images. We assess how well various fusion strategies—specifically, late and intermediate fusion—work when merging these modalities. For tabular data, the fusion studies used CNN models such as InceptionV3, ResNet152, VGG19, and ViT, and MLP models and cons models with 25 and 50 layers. To avoid overfitting, the models were trained with the Adam optimiser using a learning rate of 0.001, different batch size (8, 16, 32, 64, 128) were used but the best was found to be 32, and a dropout rate of 0.5. The CNN models that were used for the image based data were pretrained.

4.2.1 Late Fusion

4.2.1.1 Model Performance Analysis

The findings imply that attention-mechanism-based models, such as AttentionAugmentedResNet18, have a tendency to exhibit better performance consistency across various configurations. This is probably because these models can concentrate on pertinent features while reducing the architectural complexity of the model. Table 4.7 summarises the performance of several model designs employing a late fusion approach.

Overfitting is a problem for classical architectures like VGG19 and ResNet152, especially when combined with more intricate models like convolutional networks or the 50-layer MLP. ViT exhibits a modest level of performance, indicating the promise of transformer-based models. However, the exact combination of data and model architecture can affect the model's performance.

The performance of AttentionAugmentedInceptionV3 is not constant, and in more complicated settings, it tends to overfit. While InceptionV3 operates rather well on its own, when combined with larger MLP architectures, it is prone to overfitting.

All things considered, attention-based models—especially AttentionAugmentedResNet18—show the greatest potential for addressing the difficulties this dataset presents, offering a decent trade-off between accuracy and generalisation throughout the training, validation, and testing stages.

CSV Model	CNN Model	Train Loss	Train Accuracy	Val Loss	Val Accuracy	Test Loss	Test Accuracy
MLP with 25 layers	InceptionV3	0.9160	93.75%	1.0823	71.88%	1.3222	53.12%
	ResNet152	1.0806	82.29%	14.5552	90.62%	1.5224	84.38%
	VGG19	7.7314	38.54%	1.2821	34.38%	1.2461	46.88%
	ViT	0.4386	90.62%	0.3790	90.62%	0.4108	87.50%
	AttentionAugmentedInceptionV3	0.9516	89.58%	0.5179	87.50%	0.7783	75.00%
	AttentionAugmentedResNet18	0.2940	91.67%	0.0660	96.88%	0.5385	93.75%
MLP with 50 layers	InceptionV3	0.1848	98.96%	0.5889	90.62%	0.5413	84.38%
	ResNet152	0.1165	97.92%	0.3228	93.75%	0.3574	90.62%
	VGG19	0.5619	90.62%	0.2401	93.75%	0.2435	90.62%
	ViT	0.3157	90.62%	0.3575	90.62%	0.3214	87.50%
	AttentionAugmentedInceptionV3	0.2605	97.92%	0.2158	84.38%	1.4023	65.62%
	AttentionAugmentedResNet18	0.3451	97.92%	0.0942	96.88%	0.4610	90.62%
Conv Model with 25 layers	InceptionV3	0.0948	100.00%	0.1114	93.75%	0.6372	93.75%
	ResNet152	0.1183	92.71%	0.1992	90.62%	1.3503	90.62%
	VGG19	1.3009	54.17%	0.9050	43.75%	1.0738	50.00%
	ViT	0.2947	92.71%	0.3091	93.75%	0.3609	90.62%
	AttentionAugmentedInceptionV3	0.8995	64.58%	0.5462	84.38%	0.6827	87.50%
	AttentionAugmentedResNet18	0.1347	91.67%	0.1034	96.88%	0.3645	93.75%
Conv Model with 50 layers	InceptionV3	1.0322	53.12%	0.3301	96.88%	0.5347	93.75%
	ResNet152	0.9481	55.21%	2.2583	34.38%	2.2145	46.88%
	VGG19	0.1433	92.71%	0.1152	90.62%	0.2265	93.75%
	ViT	0.2899	90.62%	0.3556	93.75%	0.3325	90.62%
	AttentionAugmentedInceptionV3	0.0329	98.96%	0.3517	96.88%	0.7321	93.75%
	AttentionAugmentedResNet18	0.1045	94.79%	0.0794	96.88%	0.3675	93.75%

Table 4.7: Performance metrics of late fusion method

4.2.1.2 Precision, Recall, and F1-Score Analysis

Table 4.8 shows the Precision, Recall and F1-Score for late fusion.

CSV Model	CNN Model	Precision	Recall	F1-score
MLP with 25 layers	InceptionV3	0.156250	0.333333	0.212766
	ResNet152	0.627451	0.666667	0.645833
	VGG19	0.156250	0.333333	0.212766
	ViT	0.587302	0.622222	0.601881
	AttentionAugmentedInceptionV3	0.470085	0.444444	0.402633
	AttentionAugmentedResNet18	0.625000	0.666667	0.645161
MLP with 50 layers	InceptionV3	0.518519	0.444444	0.404762
	ResNet152	0.587302	0.622222	0.601881
	VGG19	0.604167	0.644444	0.623656
	ViT	0.596491	0.622222	0.603641
	AttentionAugmentedInceptionV3	0.441799	0.466667	0.452107
	AttentionAugmentedResNet18	0.625000	0.644444	0.634409
Conv Model with 25 layers	InceptionV3	0.627451	0.666667	0.645833
	ResNet152	0.602778	0.622222	0.612186
	VGG19	0.383142	0.355556	0.286195
	ViT	0.605229	0.644444	0.623611
	AttentionAugmentedInceptionV3	0.596491	0.622222	0.603641
	AttentionAugmentedResNet18	0.627451	0.666667	0.645833
Conv Model with 50 layers	InceptionV3	0.627451	0.666667	0.645833
	ResNet152	0.156250	0.333333	0.212766
	VGG19	0.627451	0.666667	0.645833
	ViT	0.611111	0.644444	0.624869
	AttentionAugmentedInceptionV3	0.627451	0.666667	0.645833
	AttentionAugmentedResNet18	0.627451	0.666667	0.645833

Table 4.8: Multimodal Late Fusion Precision, Recall and F1-Score

Model-wise Analysis: When combined with deeper Conv models, AttentionAugmentedResNet18 performs exceptionally well in the multimodal late fusion scenario, routinely attaining high precision, recall, and F1-scores across a range of configurations. This model shows how resilient it is and how well it can combine features from many modalities.

Nonetheless, certain models exhibit inconsistent results, such as ResNet152 and VGG19. ResNet152 does well in some combinations but performs poorly in others, especially when combined with the 50-layer Conv Model, where it shows poor precision and recall. Additionally variable, VGG19 performs well in some configurations but poorly in others, which might point to model pairing sensitivity.

InceptionV3 and ViT continue to work admirably in the majority of settings, with a usually balanced precision and recall. Nevertheless, AttentionAugmentedInceptionV3 exhibits erratic behaviour, functioning well in certain situations but less successfully in others, which may indicate that the setting affects how efficient the attention mechanism is.

Class-wise Performance: Several patterns become apparent when examining

the Intermediate Fusion models' performance across classes. Anarsia lineatella and Grapholita molesta provide F1-scores of 0.97 and 0.90, respectively, for the InceptionV3 with MLP (25 Layers) model, which performs well for these two species. However, the Dead Tree class yields an F1-score of 0.00 for this model. Across most classes, the ResNet152 with MLP (25 Layers) model exhibits notable difficulties, with an overall accuracy of only 22%. Similarly, the VGG19 with MLP (25 Layers) has very low F1-scores and underperforms, especially for Anarsia lineatella and Healthy.

Although it still fails on Dead Tree, the ViT with MLP (25 Layers) model performs better, particularly for Grapholita molesta and Anarsia lineatella. On the other hand, the AttentionAugmentedInceptionV3 with MLP (25 Layers) model performs exceptionally well in every class, obtaining nearly flawless F1-scores and a 97% overall accuracy.

The InceptionV3 with MLP (50 Layers) and ViT with MLP (50 Layers) are the two 50-layer models that exhibit balanced performance, however they are still not very successful with Dead Tree. Out of all of them, AttentionAugmentedResNet18 with MLP (50 Layers) performs the best; yet, it still has issues with several classes.

Overall, the data shows that performance varies significantly throughout models and classes. Certain models, such as AttentionAugmentedInceptionV3, routinely outperform others; this is especially true of models built on ResNet152. The outcomes show that more work has to be done to enhance performance, particularly for under-represented classes like Dead Tree. Appendix B.2.2 shows the classification reports for each model configuration.

Confusion Matrix Analysis: The performance disparities between the VGG19 and Vision Transformer (ViT) models over a range of dataset sizes and model settings are illustrated by the confusion matrices that you supplied. Because of its strong diagonal lines in its confusion matrices, VGG19 generally shows high performance with small and medium datasets. This shows that even with insufficient data, the neural design of VGG19 is well suited for efficient classification. The medium dataset improves accuracy even more by lowering the number of misclassifications.

On the other hand, the ViT models perform more inconsistently, especially on smaller datasets where the confusion matrices indicate a larger rate of misclassifications. But the ViT models get much better when the dataset size goes up to medium, with more prominent diagonal lines signifying higher accuracy. This pattern illustrates how more data is required by the ViT architecture in order to properly capture nuances and patterns in the input.

Furthermore, it seems that both model types benefit from the addition of late

convolutional layers, particularly when working with medium-sized datasets. It appears that these layers improve feature extraction, which raises classification accuracy and decreases misclassifications. Overall, deeper feature extraction procedures improve both models, but VGG19 is more dependable across a range of dataset sizes. ViT models show promise when given enough data. Appendix B.2.2 shows the confusion matrix for each model configuration.

4.2.2 Intermediate Fusion

4.2.2.1 Model Performance Analysis

A summary of the various model architectures' performance with an intermediate fusion method is provided in Table 4.9.

CSV Model	CNN Model	Train Loss	Train Accuracy	Val Loss	Val Accuracy	Test Loss	Test Accuracy
MLP with 25 layers	InceptionV3	0.0101	100.00%	0.1651	93.75%	0.4863	87.50%
	ResNet152	1.1298	59.38%	205.9879	53.12%	12.7994	53.12%
	VGG19	1.1181	52.08%	0.7673	62.50%	0.9466	46.88%
	ViT	0.2827	92.71%	0.3182	87.50%	0.4219	90.62%
	AttentionAugmentedInceptionV3	0.4482	93.75%	0.0022	100.00%	2.0357	90.62%
	AttentionAugmentedResNet18	0.1915	96.88%	0.0782	93.75%	0.3288	93.75%
MLP with 50 layers	InceptionV3	0.1874	97.92%	0.1177	93.75%	0.8883	90.62%
	ResNet152	1.1046	59.38%	1.5578	65.62%	1.2319	68.75%
	VGG19	1.2532	44.79%	0.8310	62.50%	0.9480	46.88%
	ViT	1.8505	76.04%	0.7159	75.00%	0.7215	65.62%
	AttentionAugmentedInceptionV3	0.4005	95.83%	1.4249	71.88%	1.6646	56.25%
	AttentionAugmentedResNet18	0.3446	92.71%	0.2018	96.88%	1.1611	90.62%
Conv Model with 25 layers	InceptionV3	0.0225	100.00%	0.2370	93.75%	0.4450	84.38%
	ResNet152	1.5261	79.17%	0.4069	96.88%	0.4381	93.75%
	VGG19	0.4614	81.25%	0.2914	93.75%	0.3976	93.75%
	ViT	0.4977	90.62%	0.4374	90.62%	0.4555	87.50%
	AttentionAugmentedInceptionV3	0.0880	95.83%	0.8870	96.88%	0.2192	90.62%
	AttentionAugmentedResNet18	0.3146	94.79%	0.4130	96.88%	0.7914	93.75%
Conv Model with 50 layers	InceptionV3	0.9122	84.38%	3.7757	68.75%	2.4215	50.00%
	ResNet152	1.4661	91.67%	0.8564	96.88%	1.7343	87.50%
	VGG19	0.2279	93.75%	0.0999	96.88%	0.2180	93.75%
	ViT	0.1610	93.75%	0.3746	87.50%	0.3118	87.50%
	AttentionAugmentedInceptionV3	0.1338	93.75%	0.4651	96.88%	0.4272	90.62%
	AttentionAugmentedResNet18	0.2175	92.71%	0.0721	96.88%	0.3034	93.75%

Table 4.9: Performance metrics of intermediate fusion method

According to the findings, AttentionAugmentedResNet18 exhibits the best consistency across configurations, demonstrating high accuracy and robust generalisation in both the validation and test stages. Both InceptionV3 and AttentionAugmentedIn-

ceptionV3 exhibit strong performance, however occasionally they exhibit overfitting tendencies. ResNet152 and VGG19 perform inconsistently, especially when combined with MLP models, but perform better when combined with convolutional models, suggesting that architectural selection is important. ViT demonstrates promise but has room for improvement as it maintains a decent level of performance across various setups. In general, attention-based models such as AttentionAugmentedResNet18 are the most dependable options to deal with the dataset's intricacies, offering a robust trade-off between generalisation and accuracy.

4.2.2.2 Precision, Recall, and F1-Score Analysis

Table 4.10 shows the Precision, Recall and F1-Score for intermediate fusion.

CSV Model	CNN Model	Precision	Recall	F1-score
MLP with 25 layers	InceptionV3	0.625000	0.577778	0.597372
	ResNet152	0.156250	0.333333	0.212766
	VGG19	0.156250	0.333333	0.212766
	ViT	0.611111	0.644444	0.624869
	AttentionAugmentedInceptionV3	0.627451	0.666667	0.645833
	AttentionAugmentedResNet18	0.625000	0.666667	0.645161
MLP with 50 layers	InceptionV3	0.583007	0.622222	0.601389
	ResNet152	0.484480	0.511111	0.492997
	VGG19	0.156250	0.333333	0.212766
	ViT	0.437908	0.466667	0.451389
	AttentionAugmentedInceptionV3	0.208333	0.333333	0.256410
	AttentionAugmentedResNet18	0.587302	0.622222	0.601881
Conv Model with 25 layers	InceptionV3	0.611111	0.555556	0.580247
	ResNet152	0.605229	0.644444	0.623611
	VGG19	0.627451	0.666667	0.645833
	ViT	0.587302	0.622222	0.601881
	AttentionAugmentedInceptionV3	0.625000	0.644444	0.634409
	AttentionAugmentedResNet18	0.605229	0.644444	0.623611
Conv Model with 50 layers	InceptionV3	0.156250	0.333333	0.212766
	ResNet152	0.563492	0.600000	0.578892
	VGG19	0.627451	0.666667	0.645833
	ViT	0.726190	0.766667	0.731801
	AttentionAugmentedInceptionV3	0.627451	0.666667	0.645833
	AttentionAugmentedResNet18	0.811111	0.811111	0.811111

Table 4.10: Multimodal Intermediate Fusion Precision, Recall and F1-Score

Model-wise Analysis: There is a great deal of variation in the way various model combinations perform in the multimodal intermediate fusion setting. Once again, AttentionAugmentedResNet18 performs exceptionally well, particularly when combined with the 50-layer Conv model, attaining the maximum F1-score of 0.811. This demonstrates

how attention mechanisms are useful for identifying intricate patterns and enhancing classification precision. On the other hand, models such as VGG19 and ResNet152 show uneven performance, especially when paired with more complex MLP or Conv models. Though they are sensitive to the model's depth, InceptionV3 and ViT generally perform well, suggesting that the architectural complexity may have an impact on their efficacy. In multimodal classification tasks, attention-based models generally perform better than their non-augmented counterparts, especially in configurations with deeper networks, indicating the benefit of incorporating attention mechanisms.

Class-wise Performance: The classification results for each model configuration are displayed in Appendix B.2.1.

Confusion Matrix Analysis: The most reliable models in this intermediate fusion configuration, AttentionAugmentedInceptionV3 and AttentionAugmentedResNet18, achieve good classification accuracy across most classes, as seen by the confusion matrices. While InceptionV3 likewise exhibits strong performance, it has trouble reliably recognising every instance of some classes, such Dead Tree and Healthy. However, ResNet152 and VGG19 continuously perform worse, demonstrating severe difficulties in accurately identifying several groups.

According to the analysis, attention-mechanism-augmented models (such as AttentionAugmentedInceptionV3 and AttentionAugmentedResNet18) significantly increase classification performance, especially when it comes to differentiating between difficult classes. These models produce more accurate predictions by efficiently capturing and utilising important characteristics across several classes. This suggests that attention processes work well in multimodal learning tasks where a variety of complicated patterns and input sources are involved. The confusion matrix for each model configuration are displayed in Appendix B.2.1.

4.2.3 Comparison of Fusion Techniques

Both approaches typically achieve high training accuracies and loss, especially when using models like as InceptionV3, AttentionAugmentedResNet18, and ViT. Nevertheless, late fusion has difficulties when using models like ResNet152 and VGG19, particularly when the model is configured as a 25-layer MLP. In this configuration, the models have trouble training well, as shown by their lower training accuracies and higher losses. Conversely, intermediate fusion demonstrates a markedly high training loss, indicating instability during training with specific designs, and similarly difficulties

with ResNet152 in the 25-layer MLP configuration.

In terms of test performance and validation, the late fusion approach yields a mixed bag of results. While models such as InceptionV3 and ResNet152 exhibit good performance in the 50-layer MLP designs, they show significant declines in the 25-layer setups, suggesting that there may be an overfitting problem. Conversely, AttentionAugmentedResNet18 is a dependable model option in the late fusion approach since it consistently yields high validation and test performance across many configurations. However, as difficulties with generalisation are evident, VGG19 continuously performs poorly in the majority of setups, especially in the 25-layer models.

In contrast, intermediate fusion yields more consistent outcomes in a variety of configurations. Particularly when paired with Conv models, InceptionV3 and ViT provide robust validation and test accuracy, underscoring their potent generalisation capabilities. Attention-based models, particularly AttentionAugmentedResNet18, maintain high accuracy and low loss across all configurations, demonstrating robust generalisation capabilities, while ResNet152's performance varies significantly depending on the configuration, performing poorly in the 25-layer MLP setup but improving in the 50-layer Conv setup.

In summary, when integrating various modalities, the intermediate fusion method turns out to be the more dependable strategy. In example, with designs like ResNet152 and VGG19, it mitigates some of the instability and overfitting difficulties encountered in late fusion by providing consistent performance over a range of model topologies and depths. While there are advantages to both fusion techniques, intermediate fusion offers a more robust and balanced framework that is especially useful for difficult classification tasks where generalisation is crucial throughout the training, validation, and testing phases.

Chapter 5

Conclusion

5.1 Summary of Findings

5.1.1 Single Modal

Ground RGB Results: For minority classes, the models' recall and precision are low, pointing to a serious class imbalance and poor generalisation. They show considerable loss during testing, especially with minority classes, despite having strong training and validation accuracy, which suggests overfitting to the majority class. Confusion matrices also show that minority classes are frequently misclassified as the majority class, underscoring the drawbacks of depending only on ground RGB imagery.

Multispectral Results: Beyond what RGB imaging provides, the additional spectral data can improve feature detection and possibly improve class distinction. Like RGB models, they may experience overfitting and poor generalisation, particularly with minority classes, and are yet constrained by class imbalances. Multispectral data is valuable, but it also adds complexity to the model, increasing the likelihood of overfitting and creating convergence issues.

Tabular Data Results: Models find it challenging to represent class complexity in tabular data because to the lack of spectral and geographical context, particularly when attempting to detect visually tiny distinctions that are challenging to quantify. Tabular models frequently perform poorly on minority classes due to their over-reliance on dominant features, as evidenced by the persistently low F1-scores, recall, and precision for all classes except the dominant one. Furthermore, tabular data's restricted flexibility makes it difficult for the model to generalise effectively across various classes.

5.1.2 Multimodal

The integration of features from several modalities produced greater precision, recall, and F1-scores across a range of models, as shown by both late and intermediate fusion methods. In particular, models with constant high performance were ViT and AttentionAugmentedResNet18, demonstrating that multimodal fusion contributes to the capture of more extensive patterns and hence improves overall accuracy. Nonetheless, several models encountered difficulties with particular combinations, suggesting that the fusion technique and model choice require meticulous deliberation. All things considered, multimodal fusion turned out to be a strong method that successfully combined the advantages of several data kinds to enhance classification results.

5.2 Limitations

Due to a lack of data in developing countries with the modalities required (multispectral images, ground imagery, UAV imagery, and environmental parameters), for the research we used a Peach Tree Disease Detection dataset located in Greece.[14] The main obstacle this study encountered was the imbalance in the dataset, especially across the different classes. The under-representation of several classes made it challenging to train models efficiently. Lower precision, recall, and F1-scores for these under-represented categories show that models biassed towards more frequent classes as a result of this imbalance performed poorly on minority classes.

5.3 Future Work

This project will need to be improved in a number of important areas in the future. Initially, it is imperative to tackle the imbalance in the dataset. This can be achieved by utilising methods such as data augmentation, synthetic data generation, or advanced loss functions like focal loss, which can improve the performance of the model on under-represented classes. Understanding and believing in the behaviour of the model will depend on the development of techniques to better analyse and explain decisions made by the model, such as integrated gradients or SHAP values.

Another crucial factor is scalability, which guarantees that the models can manage bigger datasets and a variety of data distributions. To achieve this, it could be necessary to optimise computing resources and apply domain adaptation strategies. Ultimately, the practical usability and effectiveness of these models will need to be validated by integration into real-world applications with continuous monitoring and fine-tuning.

Bibliography

[1] Gashaw T. Abate et al. Digital tools and agricultural market transformation in africa: Why are they not at scale yet, and what will it take to get there? *Food Policy*, 116:102439.

[2] Luis B Almeida. Multilayer perceptrons. In *Handbook of Neural Computation*, pages C1–2. CRC Press, 2020.

[3] Habiba Arsenio, Devansh Mahajan, Zhaoxia Di, and David Langer. Tabnet: Attentive interpretable tabular learning. *Proceedings of the 2020 ACM Conference on Data Science and Learning*, pages 18–26, 2020.

[4] Omneya Attallah. Multitask deep learning-based pipeline for gas leakage detection via e-nose and thermal imaging multimodal fusion. *Chemosensors*, 11:364, 06 2023.

[5] Ishana Attri, Lalit Kumar Awasthi, and Teek Parval Sharma. Machine learning in agriculture: a review of crop management applications. *Multimedia Tools and Applications*, 83(5):12875–12915.

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[7] Yadeta Bedasa and Kumala Deksisa. Journal of agriculture and food research. *Journal of Agriculture and Food Research*, 15:100978.

[8] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019.

[9] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks, 2020.

[10] M. Bhandari, A. Ramcharan, J. Bartholomé, S. Tripathy, A. Sarkar, and S. Sankaran. Detection of wheat leaf diseases using rgb uav imaging and machine learning techniques. *Remote Sensing*, 12(11):1832, 2020.

[11] Mounir Brahimi, Kamel Boukhalfa, and Abdelmalek Moussaoui. Deep learning for plant disease detection and classification using image processing. *Computers and Electronics in Agriculture*, 142:361–369, 2017.

[12] Ljubomir Buturović and Dejan Miljković. A novel method for classification of tabular data using convolutional neural networks. *BioRxiv*, pages 2020–05, 2020.

[13] R. Calderón, J. A. Navas-Cortes, C. Lucena, and P. J. Zarco-Tejada. Detection of verticillium wilt in olive orchards using uav-based multispectral and thermal imagery. *Remote Sensing*, 7(5):5584–5610, 2015.

[14] C. Chaschatzis, C. Karaiskou, E. Mouratidis, E. Karagiannis, and P. Sarigiannidis. Detection and characterization of stressed sweet cherry tissues using machine learning. *Drones*, 6(1), 2022.

[15] Wuyang Chen, Xianzhi Du, Fan Yang, Lucas Beyer, Xiaohua Zhai, Tsung-Yi Lin, Huizhong Chen, Jing Li, Xiaodan Song, Zhangyang Wang, et al. A simple single-scale vision transformer for object localization and instance segmentation. *arXiv preprint arXiv:2112.09747*, 2021.

[16] C. H. W. de Souza, R. A. Krohling, and M. Kampel. Detection of coffee leaf rust using uav-based rgb imaging and machine learning algorithms. *International Journal of Remote Sensing*, 38(9):2313–2325, 2017.

[17] VG Dhanya, A Subeesh, NL Kushwaha, Dinesh Kumar Vishwakarma, T Nagesh Kumar, G Ritika, and AN Singh. Deep learning based computer vision approaches for smart agricultural applications. *Artificial Intelligence in Agriculture*, 6:211–229, 2022.

[18] Yanan Duan, Guangjian Yan, Xuegong Zhang, Yuhua Ma, Jianmin Zhang, and Yiming Ren. Multimodal uav-based crop disease detection and monitoring using deep learning and data fusion. *Remote Sensing*, 11(5):524, 2019.

[19] O. B. Falana and O. I. Durodola. Multimodal remote sensing and machine learning for precision agriculture: A review. *Journal of Engineering Research and Reports*, 23(8):30–34, 2022.

[20] Helia Farhood, Ivan Bakhshayeshi, Matineh Pooshideh, Nabi Rezvani, and Amin Beheshti. Recent advances of image processing techniques in agriculture. *Artificial Intelligence and Data Science in Environmental Sensing*, pages 129–153, 2022.

[21] Konstantinos P Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018.

[22] Food and Agriculture Organization of the United Nations. The state of food security and nutrition in the world, 2020.

[23] Martin Ger. Attention-augmented convolutional networks. `https://github.com/MartinGer/Attention-Augmented-Convolutional-Networks`, 2024. GitHub repository.

[24] Hongbo Guan, Guoyu Jia, Yabo Zhao, Yiming Tian, Jianxin Cui, and Qiang Feng. Multimodal data fusion for monitoring wheat rust using rgb, multispectral, and lidar data. *Computers and Electronics in Agriculture*, 184:106092, 2021.

[25] Monia Guizani, Samira Maatallah, and Aida Ltifi. Influence of water stress on the nutritional quality of peach fruits. *JOURNAL OF OASIS AGRICULTURE AND SUSTAINABLE DEVELOPMENT*, 4(2):140–147, 2022.

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

[27] E. Hunt, W. Akemi, and R. Ward. Multispectral uav imagery for detecting phytophthora infestans in potato crops. *Remote Sensing*, 10(6):923, 2018.

[28] C Jackulin and S Murugavalli. A comprehensive review on detection of plant disease using machine learning and deep learning approaches. *Measurement: Sensors*, 24:100441, 2022.

[29] Y. Jiang, C. Li, R. W. Ward, and A. Ebadi. Enhanced detection of northern leaf blight in maize using uav-based multispectral and hyperspectral imagery. *Remote Sensing*, 11(1):103, 2019.

[30] Andreas Kamilaris and Francesc X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.

[31] Priyabrata Karmakar, Shyh Wei Teng, Manzur Murshed, Shaoning Pang, Yanyu Li, and Hao Lin. Crop monitoring by multimodal remote sensing: A review. *Remote Sensing Applications: Society and Environment*, 33:101093, 2024.

[32] Christopher Kruse, Larissa Roese-Koerner, and Sebastian Bretthauer. Convolutional neural networks for tabular data: A novel approach in agriculture. *Journal of Precision Agriculture*, 23:245–256, 2022.

[33] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: An overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[34] H. Liu, R. Jin, H. Hu, Y. Song, and J. Liu. Ndvi-based uav monitoring of late blight in potato fields. *Remote Sensing*, 10(4):570, 2018.

[35] A. Lowe, N. Harrison, and A. P. French. Monitoring grapevine leafroll-associated virus-3 in vineyards using uav-based rgb imaging. *Precision Agriculture*, 18(4):481–494, 2017.

[36] A. Lytos, T. Lagkas, P. Sarigiannidis, M. Zervakis, and G. Livanos. Towards smart farming: Systems, frameworks and exploitation of multiple sources. *Computer Networks*, 172(107147), 2020.

[37] Anne-Katrin Mahlein, Michael T Kuska, Janina Behmann, Gerie Polder, and Achim Walter. Application of random forest for the classification of crop diseases based on hyperspectral and thermal imaging data. *Computers and Electronics in Agriculture*, 145:183–195, 2018.

[38] Maitiniyazi Maimaitijiang, Valery Sagan, Paheding Sidike, Sam Hartling, Fabio Esposito, and Frank Fritschi. Soybean yield prediction from uav using multimodal data fusion and deep learning. *Remote Sensing of Environment*, 237:111599, 2020.

[39] Robert Mendelsohn. The impact of climate change on agriculture in developing countries. *Journal of Natural Resources Policy Research*, 1(1):5–19.

[40] Günter E. Meyer, Timothy W. Hindman, Rajasekhar Khosla, Luke A. Dasi, and Barbara M. Niendorf. Hyperspectral and thermal uav imaging for fusarium head blight detection and disease severity assessment in wheat. *Agriculture*, 9(11):227, 2019.

[41] P. Mishra, V. Singh, P. Singh, and A. Kumar. Multispectral uav imaging for early detection of maize gray leaf spot disease. *Precision Agriculture*, 21(2):379–391, 2020.

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

[43] T. J. Nigon, D. R. Turnbull, and L. R. Khot. Detection of cercospora leaf spot in sugar beet using uav rgb imagery. *Remote Sensing Applications: Society and Environment*, 2:96–103, 2015.

[44] Nikos T. Papadopoulos, Brett R. Blaauw, Panagiotis Milonas, and Anne L. Nielsen. Biology and management of insect pests. *CABI*, page 366–420, 2023.

[45] Gerrit Polder, Ferdinand van der Heijden, Jeroen van Doorn, and Roeland van der Schoor. Detection of potato late blight using rgb and nir imaging under varying light conditions. *Computers and Electronics in Agriculture*, 133:25–31, 2017.

[46] P. Radoglou-Grammatikis, P. Sarigiannidis, T. Lagkas, and I. Moscholios. A compilation of uav applications for precision agriculture. *Computer Networks*, 172(107148), 2020.

[47] Diogo Richetti, Bruno Francois, and Nathalie Moulin. Machine learning methods for predicting wheat and barley yield: A comparison of models using tabular data. *Computers and Electronics in Agriculture*, 193:106579, 2023.

[48] Jonathan Richetti, Foivos I Diakogianis, Asher Bender, André F Colaço, and Roger A Lawes. A methods guideline for deep learning for tabular data in agriculture with a case study to forecast cereal yield. *Computers and Electronics in Agriculture*, 205:107642, 2023.

[49] Torsten Rumpf, Anne-Katrin Mahlein, Uwe Steiner, Ernst-Christian Oerke, Heinz-Wilhelm Dehne, and Lutz Plümer. Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance. *Precision Agriculture*, 11(6):607–623, 2010.

[50] S. Sankaran, A. Mishra, and R. Ehsani. Soybean rust detection using uav-based multispectral imagery and machine learning. *Journal of Precision Agriculture*, 18(4):287–300, 2017.

[51] Sindhuja Sankaran, Joe Mario Maja, Shelby Buchanon, and Reza Ehsani. Hyperspectral and thermal imaging for early detection of citrus greening disease in young trees. *Sensors*, 17(2):211, 2017.

[52] Tej Bahadur Shahi, Cheng-Yuan Xu, Arjun Neupane, and William Guo. Recent advances in crop disease detection using uav and deep learning techniques. *Remote Sensing*, 15(9):2450, 2023.

[53] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[54] Srdjan Sladojevic, Marko Arsenovic, Ana Anderla, Dubravko Culibrk, and Darko Stefanovic. Deep neural networks based recognition of plant diseases by leaf image classification. *Computational intelligence and neuroscience*, 2016:1–11, 2016.

[55] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[56] K. R. Thorp, L. Tian, and S. Jha. Mapping rust disease in wheat using uav-acquired ndvi data. *Journal of Field Robotics*, 34(2):413–423, 2017.

[57] Washington State University. Peach twig borer, 2024. Accessed: 2024-08-17.

[58] Raj Kishor Verma and Kaushal Kishor. Image processing applications in agriculture with the help of ai. In *Infrastructure Possibilities and Human-Centered Approaches With Industry 5.0*, pages 162–181. IGI Global, 2024.

[59] Dashuai Wang, Wujing Cao, Fan Zhang, Zhuolin Li, Sheng Xu, and Xinyu Wu. A review of deep learning in multiscale agricultural sensing. *Remote Sensing*, 14(3):559, 2022.

[60] World Bank. Agriculture overview.

[61] X. Zhang, J. Zhao, H. Yang, X. Yan, Y. Ren, and H. Yuan. Uav-based multispectral imaging for monitoring wheat yellow rust disease and nitrogen status. *Remote Sensing*, 12(9):1478, 2020.

[62] Y. Zhao, Z. Zhang, J. Wang, S. Jiang, H. Sun, and Y. Zhao. Monitoring rice blast disease using uav-based multi-temporal multispectral imagery. *Remote Sensing*, 10(4):545, 2018.

[63] L. Zheng, L. Gao, Y. Zhang, J. Wang, X. Yang, and Z. Li. Fusion of uav multispectral and ndvi data for improved wheat disease detection and mapping. *Remote Sensing*, 13(3):487, 2021.

[64] Liang Zhou, Lu Gao, Yingchun Zhang, Jiantao Wang, Xiaoyun Yang, and Zhiqiang Li. Fusion of uav multispectral and ndvi data for improved wheat disease detection and mapping. *Remote Sensing*, 13(3):487, 2021.

[65] Y. Zhou, Y. Gao, and Z. Jiang. Multispectral uav imaging for monitoring bacterial leaf blight in rice. *Remote Sensing*, 13(3):621, 2021.

Appendix A

Methodology

A.1 UAV with Bounding Boxes

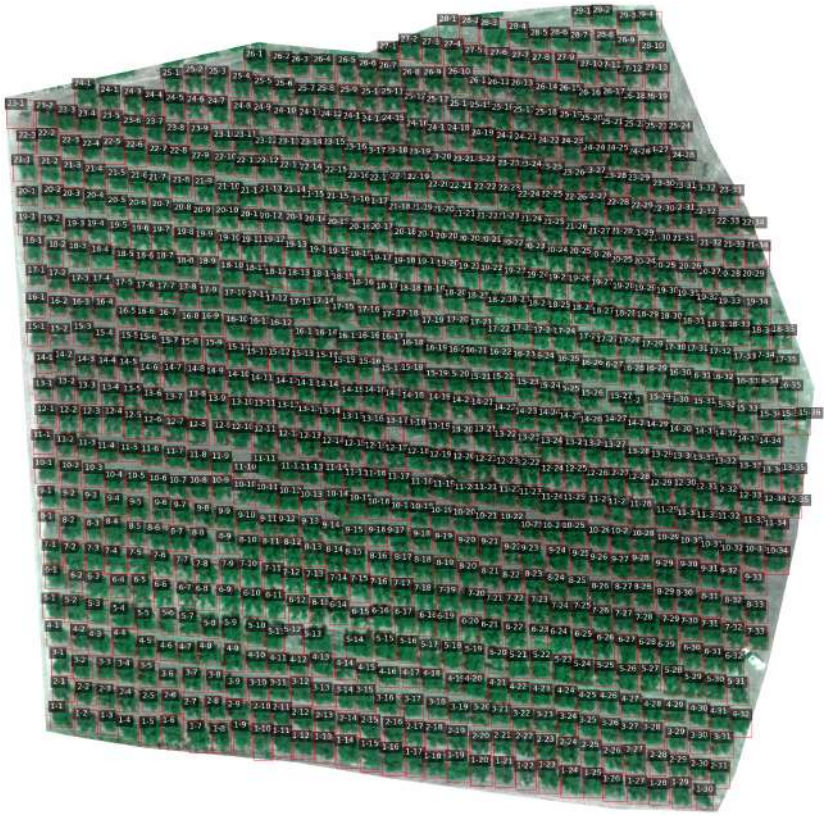


Figure A.1: RGB UAV with Bounding Boxes

A.2 Vegetation Indicators

The benefits, computations, and explanations for a number of important vegetation indicators are provided below: **- NDVI (Normalized Difference Vegetation Index)** The NDVI is a commonly utilised vegetation index due to its ease of calculation and ability to distinguish between healthy and stressed vegetation. It draws attention to the distinctions between surfaces such as soil and water and flora, which reflects strongly in the near-infrared spectrum. The range of NDVI readings is -1 to 1, with higher values (usually between 0.2 and 0.8) denoting healthy vegetation and values around 0 or negative denoting non-vegetated surfaces.

$$\text{NDVI} = \frac{(\text{NIR} - \text{Red})}{(\text{NIR} + \text{Red})} \tag{A.1}$$

Where:

- NIR = Near-Infrared reflectance
- Red = Red reflectance

- EVI (Enhanced Vegetation Index)

EVI enhances NDVI by lessening the impact of soil background and atmospheric circumstances and becomes more dependable in regions with extensive soil exposure or dense vegetation. Because the NDVI tends to saturate (flatten out) in areas with significant biomass, EVI is especially helpful in these areas. In order to account for aerosol scattering, EVI additionally uses the Blue band, resulting in more precise vegetation monitoring.

$$\text{EVI} = G \times \frac{(\text{NIR} - \text{Red})}{(\text{NIR} + C_1 \times \text{Red} - C_2 \times \text{Blue} + L)} \tag{A.2}$$

Where:

- NIR = Near-Infrared reflectance
- Red = Red reflectance
- Blue = Blue reflectance
- G = Gain factor (usually 2.5)
- C_1, C_2 = Coefficients (usually 6 and 7.5, respectively)
- L = Canopy background adjustment (usually 1)

-NDRE (Normalized Difference Red Edge Index)

NDRE is very good at identifying the vegetation stress before the NDVI shows signs of it, this makes it beneficial for the early diagnosis of sickness, nutrient deficiencies, or water stress. Compared to NDVI, it is less susceptible to the effects of soil background, giving a more distinct signal of the state of the vegetation. In precision agriculture, NDRE is frequently used to track crop health, particularly during important growth phases.

$$\text{NDRE} = \frac{(\text{NIR} - \text{Red Edge})}{(\text{NIR} + \text{Red Edge})} \tag{A.3}$$

Where:

- NIR = Near-Infrared reflectance
- Red Edge = Reflectance in the Red Edge spectrum (usually around 705–740 nm)

- GNDVI (Green Normalized Difference Vegetation Index)

GNDVI is comparable to NDVI but uses the green band in place of the red band. Since it is highly sensitive to the amount of chlorophyll, it is useful for tracking photosynthetic activity and the leaf area index (LAI). It is helpful in determining plant vigour and identifying minute variations in the health of the vegetation. Precision agriculture frequently uses GNDVI to track crop growth and apply fertiliser as efficiently as possible.

$$\text{GNDVI} = \frac{(\text{NIR} - \text{Green})}{(\text{NIR} + \text{Green})} \tag{A.4}$$

Where:

- NIR = Near-Infrared reflectance
- Green = Green reflectance

- RVI (Ratio Vegetation Index)

Vegetation health can be quickly evaluated with the help of the straightforward ratio known as RVI. It can be applied to a variety of vegetation monitoring applications and is sensitive to the amount of green biomass. Compared to the NDVI, it is less susceptible to atmospheric conditions and can be helpful in areas with less vegetation cover. To provide a thorough evaluation of the state of the vegetation, RVI is frequently used in conjunction with other indices.

$$\text{RVI} = \frac{\text{NIR}}{\text{Red}} \tag{A.5}$$

Where:

- NIR = Near-Infrared reflectance
- Red = Red reflectance

- TVI (Transformed Vegetation Index)

TVI works especially well for stress detection and plant health monitoring because it increases the sensitivity of vegetation indicators to chlorophyll content. In certain situations, it can distinguish between stressed and healthy vegetation more accurately than the NDVI, particularly in settings with intricate canopy systems. For the purpose of monitoring crop growth phases and directing management procedures, TVI is helpful in precision agriculture.

$$\text{TVI} = 0.5 \times [120 \times (\text{NIR} - \text{Green}) - 200 \times (\text{Red} - \text{Green})] \tag{A.6}$$

Where:

- NIR = Near-Infrared reflectance
- Red = Red reflectance
- Green = Green reflectance

Appendix B

Results

B.1 Single Modal

B.1.1 Ground RGB Imagery

The following tables present the classification reports for each model, highlighting the precision, recall, and F1-score for each class.

B.1.1.1 Classification Report and Confusion Matrix for InceptionV3

Class	Precision	Recall	F1-score
Anarsia lineatella	0.67	1.00	0.80
Grapholita molesta	0.91	1.00	0.95
Healthy	0.93	1.00	0.96
Dead Trees	0.00	0.00	0.00
Accuracy			0.95

Table B.1: Classification Report for InceptionV3

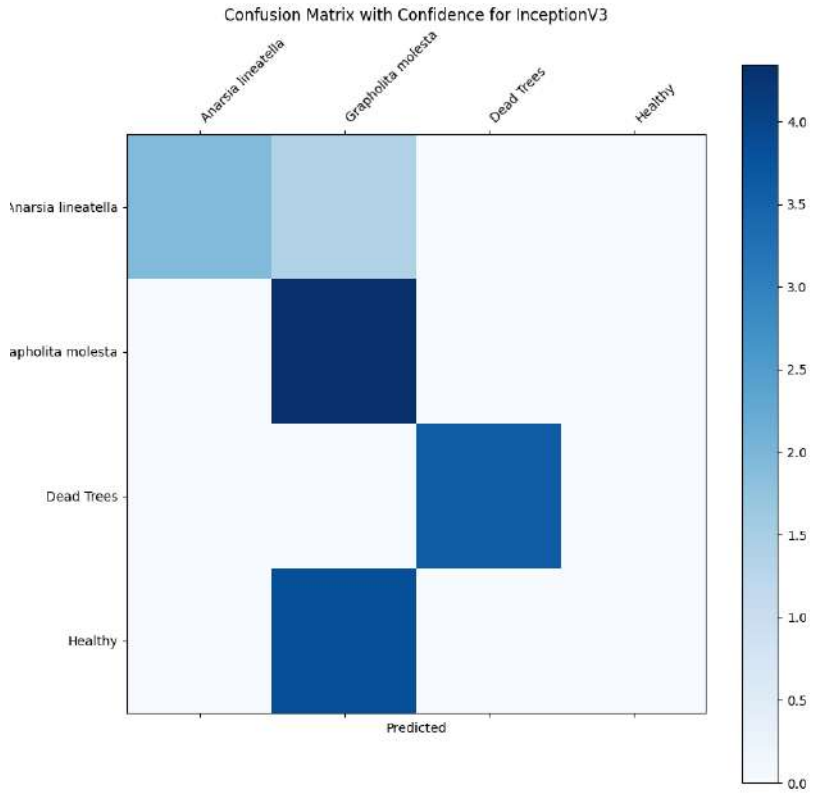


Figure B.1: Confusion Matrix with Confidence for InceptionV3

B.1.1.2 Classification Report and Confusion Matrix for ResNet152

Class	**Precision**	**Recall**	**F1-score**
Anarsia lineatella	1.00	1.00	1.00
Grapholita molesta	0.92	1.00	0.96
Healthy	1.00	1.00	1.00
Dead Trees	0.00	0.00	0.00
Accuracy			0.98

Table B.2: Classification Report for ResNet152

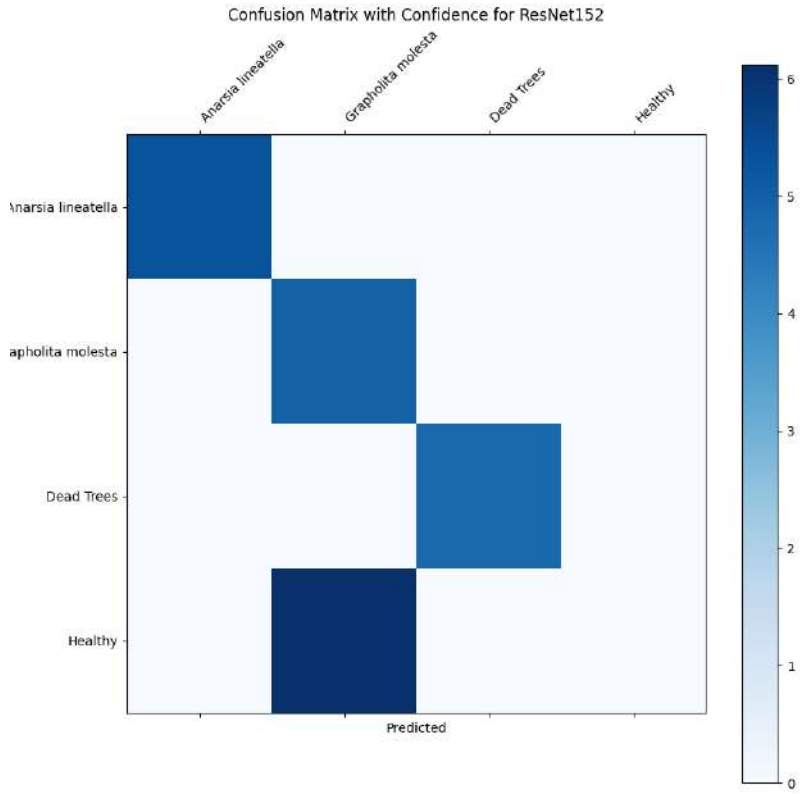


Figure B.2: Confusion Matrix with Confidence for ResNet152

B.1.1.3 Classification Report and Confusion Matrix for VGG19

Class	Precision	Recall	F1-score
Anarsia lineatella	1.00	1.00	1.00
Grapholita molesta	0.92	1.00	0.96
Healthy	1.00	1.00	1.00
Dead Trees	0.00	0.00	0.00
Accuracy			0.98

Table B.3: Classification Report for VGG19

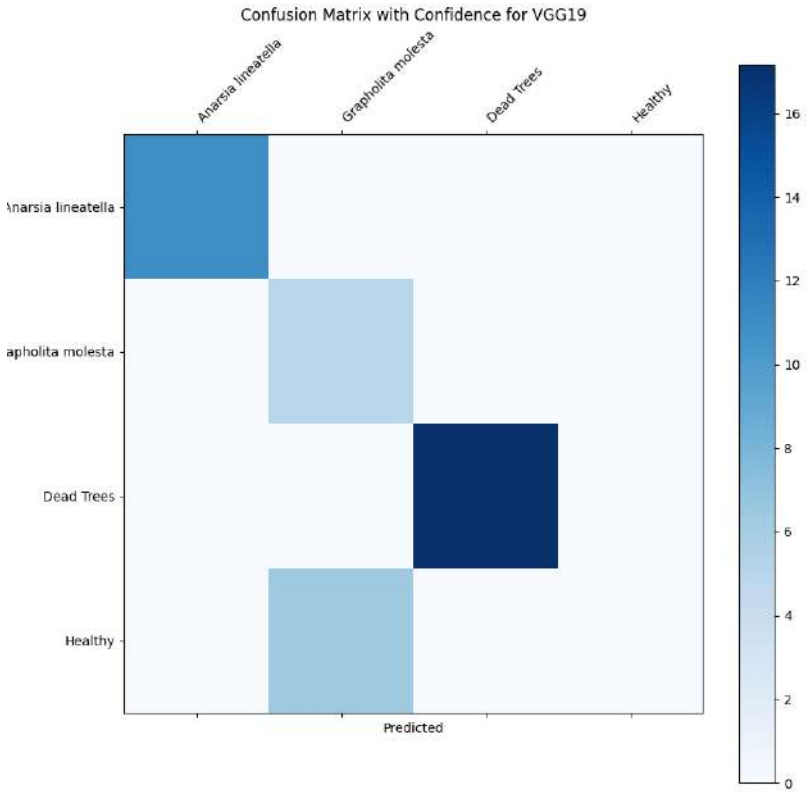


Figure B.3: Confusion Matrix with Confidence for VGG19

B.1.1.4 Classification Report and Confusion Matrix for ViT

Class	Precision	Recall	F1-score
Anarsia lineatella	0.67	1.00	0.80
Grapholita molesta	0.77	0.83	0.80
Healthy	1.00	0.92	0.96
Dead Trees	0.00	0.00	0.00
Accuracy			0.88

Table B.4: Classification Report for ViT

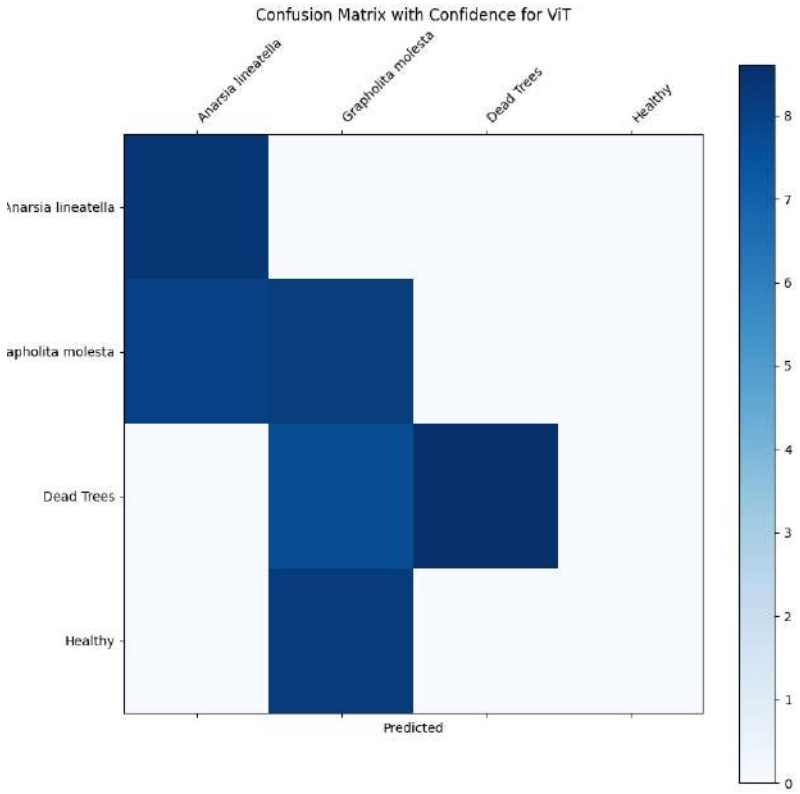


Figure B.4: Confusion Matrix with Confidence for ViT

B.1.1.5 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3

Class	Precision	Recall	F1-score
Anarsia lineatella	1.00	1.00	1.00
Grapholita molesta	0.91	0.83	0.87
Healthy	0.93	1.00	0.96
Dead Trees	0.00	0.00	0.00
Accuracy			0.93

Table B.5: Classification Report for AttentionAugmentedInceptionV3

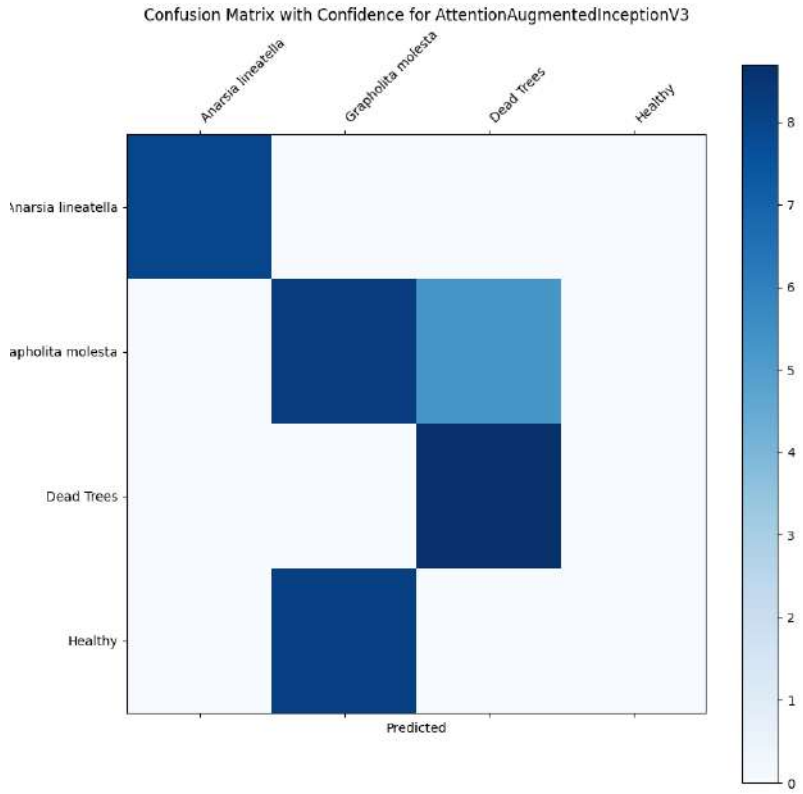


Figure B.5: Confusion Matrix with Confidence for AttentionAugmentedInceptionV3

B.1.1.6 Classification Report and Confusion Matrix for AttentionAugmentedResNet18

Class	Precision	Recall	F1-score
Anarsia lineatella	0.00	0.00	0.00
Grapholita molesta	0.67	0.83	0.74
Healthy	0.89	0.96	0.93
Dead Trees	0.00	0.00	0.00
Accuracy			0.81

Table B.6: Classification Report for AttentionAugmentedResNet18

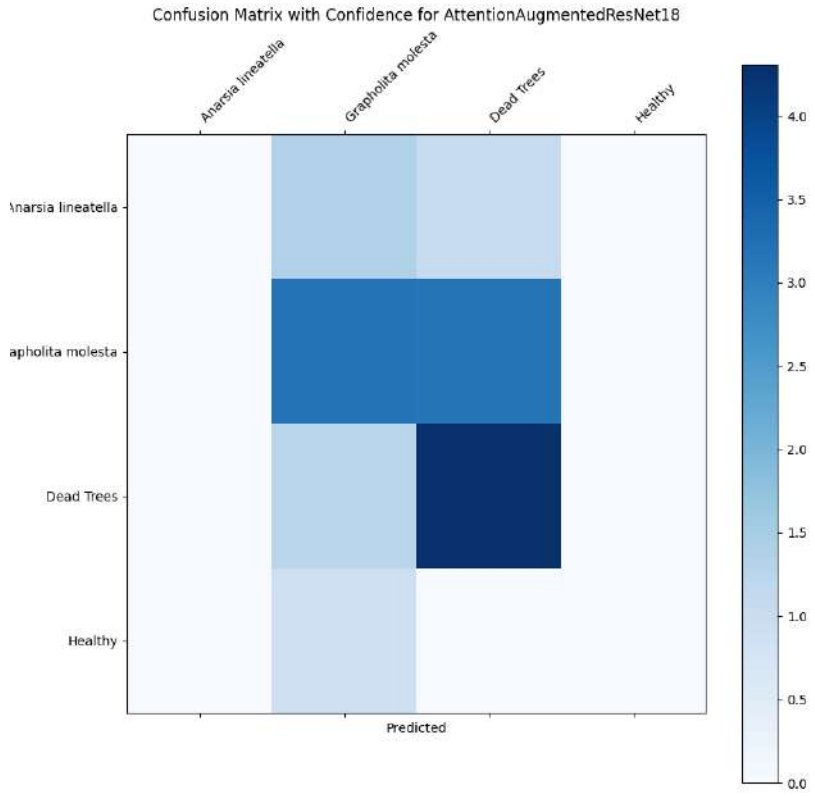


Figure B.6: Confusion Matrix with Confidence for AttentionAugmentedResNet18

B.1.2 Ground Multispectral Imagery

B.1.2.1 Classification Report and Confusion Matrix for InceptionV3

Class	Precision	Recall	F1-score
Anarsia lineatella	0.75	0.75	0.75
Grapholita molesta	0.87	0.93	0.90
Dead Trees	0.95	1.00	0.97
Healthy	0.00	0.00	0.00
Accuracy		0.89	
Macro avg	0.64	0.67	0.65
Weighted avg	0.85	0.89	0.87

Table B.7: Classification Report for InceptionV3

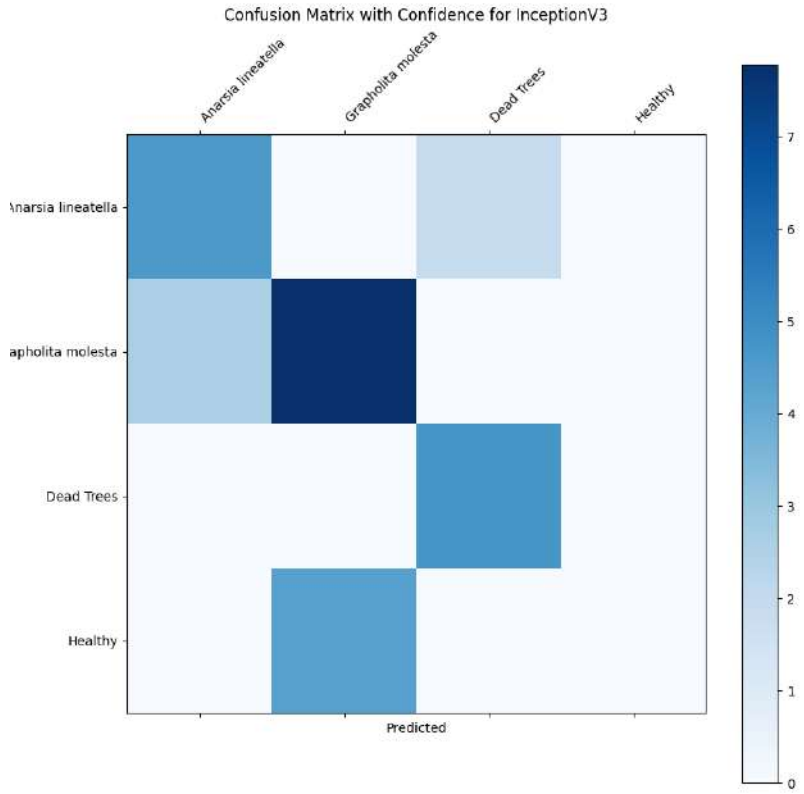


Figure B.7: Confusion Matrix with Confidence for InceptionV3

B.1.2.2 Classification Report and Confusion Matrix for ResNet152

Class	Precision	Recall	F1-score
Anarsia lineatella	1.00	1.00	1.00
Grapholita molesta	0.93	1.00	0.97
Dead Trees	1.00	1.00	1.00
Healthy	1.00	0.50	0.67
Accuracy		0.97	
Macro avg	0.98	0.88	0.91
Weighted avg	0.98	0.97	0.97

Table B.8: Classification Report for ResNet152

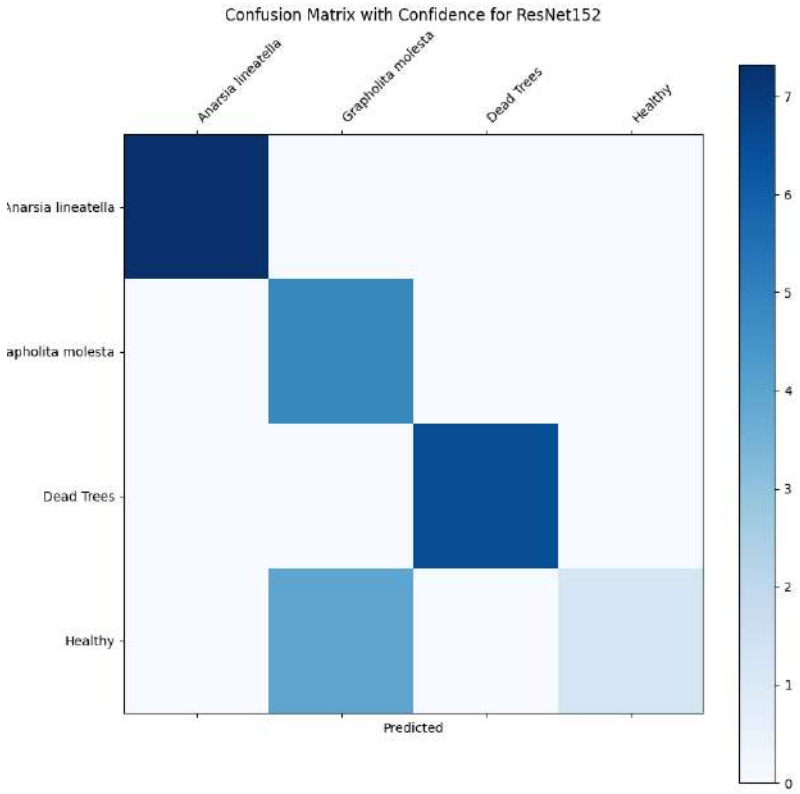


Figure B.8: Confusion Matrix with Confidence for ResNet152

B.1.2.3 Classification Report and Confusion Matrix for VGG19

Class	Precision	Recall	F1-score
Anarsia lineatella	0.27	1.00	0.42
Grapholita molesta	0.89	0.57	0.70
Dead Trees	1.00	0.72	0.84
Healthy	1.00	0.50	0.67
Accuracy		0.68	
Macro avg	0.79	0.70	0.66
Weighted avg	0.88	0.68	0.73

Table B.9: Classification Report for VGG19

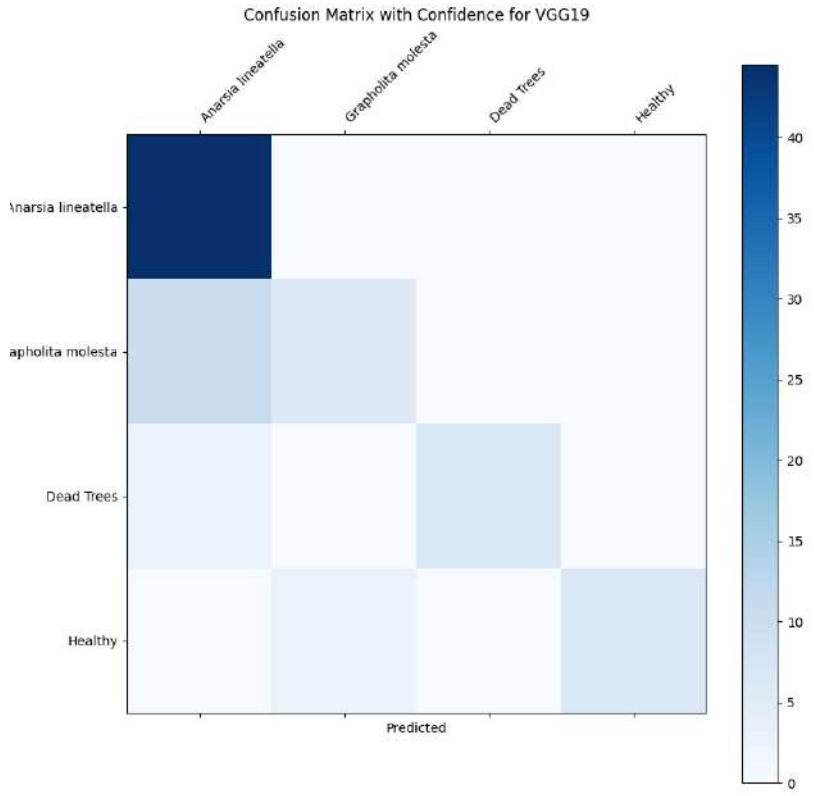


Figure B.9: Confusion Matrix with Confidence for VGG19

B.1.2.4 Classification Report and Confusion Matrix for ViT

Class	Precision	Recall	F1-score
Anarsia lineatella	1.00	1.00	1.00
Grapholita molesta	0.68	0.93	0.79
Dead Trees	0.93	0.78	0.85
Healthy	0.00	0.00	0.00
Accuracy		0.82	
Macro avg	0.65	0.68	0.66
Weighted avg	0.80	0.82	0.80

Table B.10: Classification Report for ViT

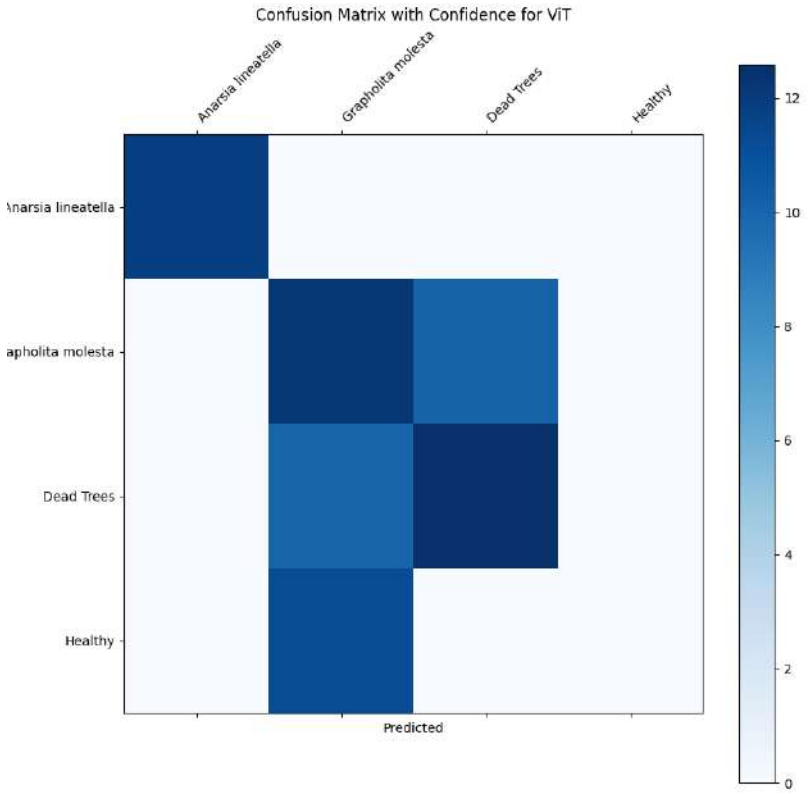


Figure B.10: Confusion Matrix with Confidence for ViT

B.1.2.5 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3

Class	Precision	Recall	F1-score
Anarsia lineatella	1.00	0.75	0.86
Grapholita molesta	0.88	1.00	0.93
Dead Trees	0.95	1.00	0.97
Healthy	0.00	0.00	0.00
Accuracy		0.92	
Macro avg	0.71	0.69	0.69
Weighted avg	0.88	0.92	0.89

Table B.11: Classification Report for AttentionAugmentedInceptionV3

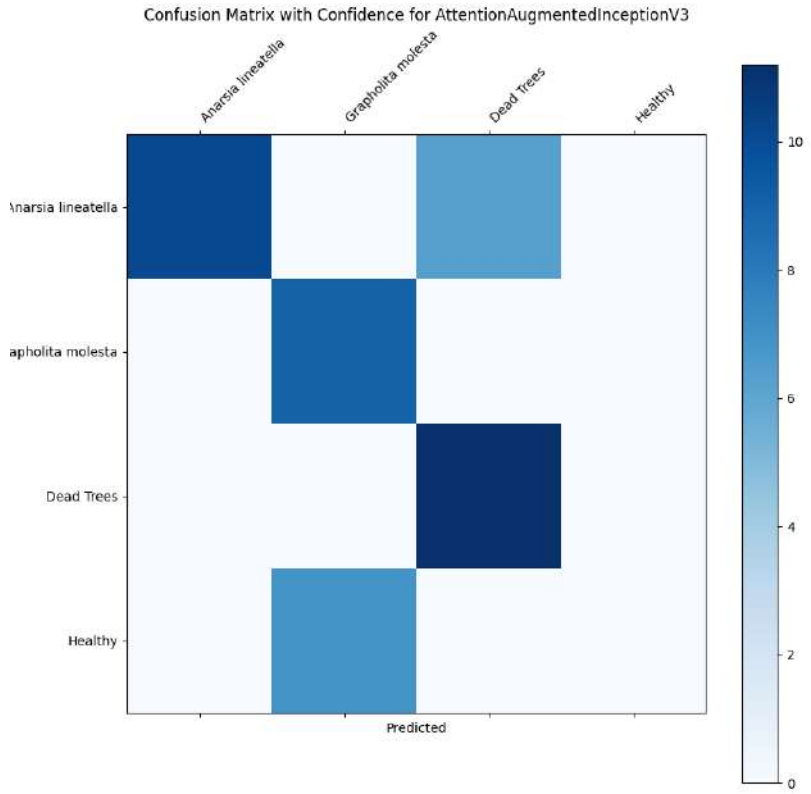


Figure B.11: Confusion Matrix with Confidence for AttentionAugmentedInceptionV3

B.1.2.6 Classification Report and Confusion Matrix for AttentionAugmentedResNet18

Class	Precision	Recall	F1-score
Anarsia lineatella	0.80	1.00	0.89
Grapholita molesta	0.87	0.93	0.90
Dead Trees	1.00	1.00	1.00
Healthy	0.00	0.00	0.00
Accuracy		0.92	
Macro avg	0.67	0.73	0.70
Weighted avg	0.88	0.92	0.90

Table B.12: Classification Report for AttentionAugmentedResNet18

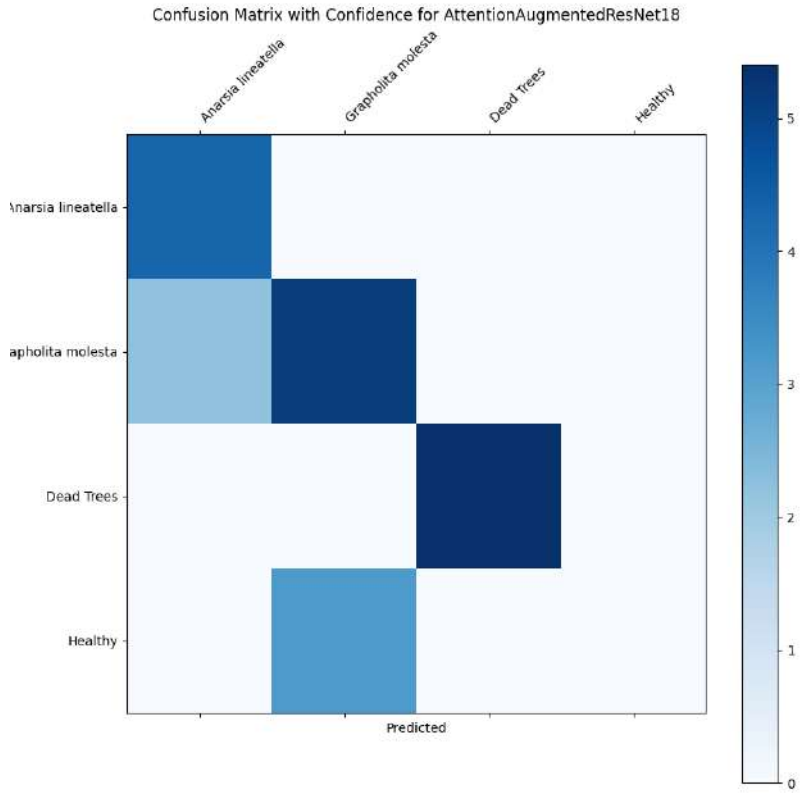


Figure B.12: Confusion Matrix with Confidence for AttentionAugmentedResNet18

B.1.3 Tabular Data

B.1.3.1 Classification Report and Confusion Matrix for MLP with 25 Layers

Class	Precision	Recall	F1-score
Anarsia lineatella	0.95	1.00	0.97
Grapholita molesta	0.00	0.00	0.00
Dead Trees	0.00	0.00	0.00
Healthy	0.00	0.00	0.00
Accuracy			0.95
Macro avg	0.24	0.25	0.24
Weighted avg	0.90	0.95	0.93

Table B.13: Classification Report for MLP with 25 Layers

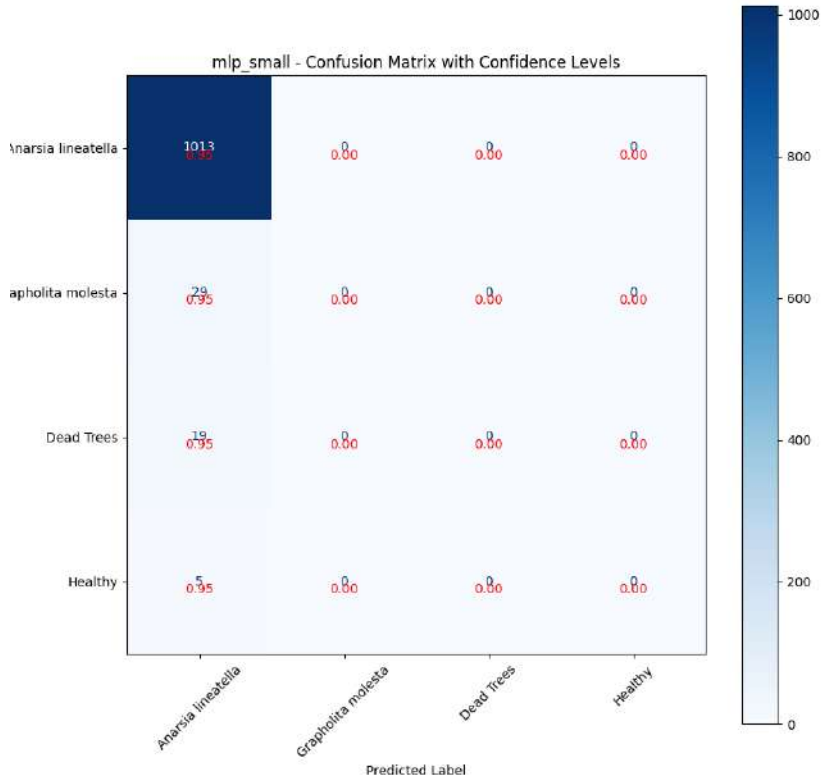


Figure B.13: Confusion Matrix with Confidence for MLP with 25 Layers

B.1.3.2 Classification Report and Confusion Matrix for MLP with 50 Layers

Class	Precision	Recall	F1-score
Anarsia lineatella	0.95	1.00	0.97
Grapholita molesta	0.00	0.00	0.00
Dead Trees	0.00	0.00	0.00
Healthy	0.00	0.00	0.00
Accuracy			0.95
Macro avg	0.24	0.25	0.24
Weighted avg	0.90	0.95	0.93

Table B.14: Classification Report for MLP with 50 Layers

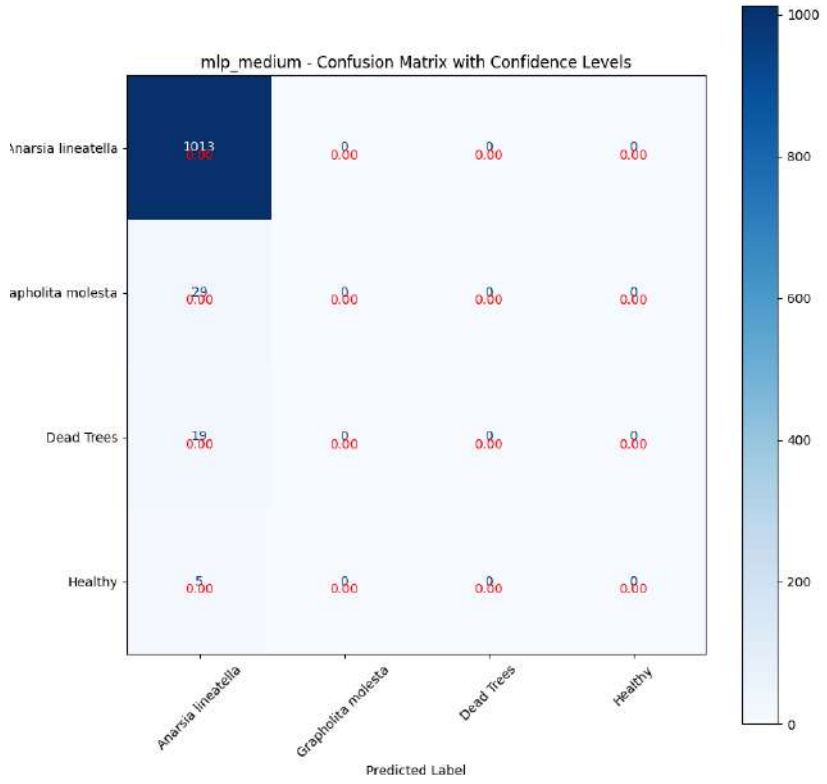


Figure B.14: Confusion Matrix with Confidence for MLP with 50 Layers

B.1.3.3 Classification Report and Confusion Matrix for Conv Model with 25 layes

Class	Precision	Recall	F1-score
Anarsia lineatella	0.95	1.00	0.97
Grapholita molesta	0.00	0.00	0.00
Dead Trees	0.00	0.00	0.00
Healthy	0.00	0.00	0.00
Accuracy			0.95
Macro avg	0.24	0.25	0.24
Weighted avg	0.90	0.95	0.93

Table B.15: Classification Report for Conv Model with 25 layes

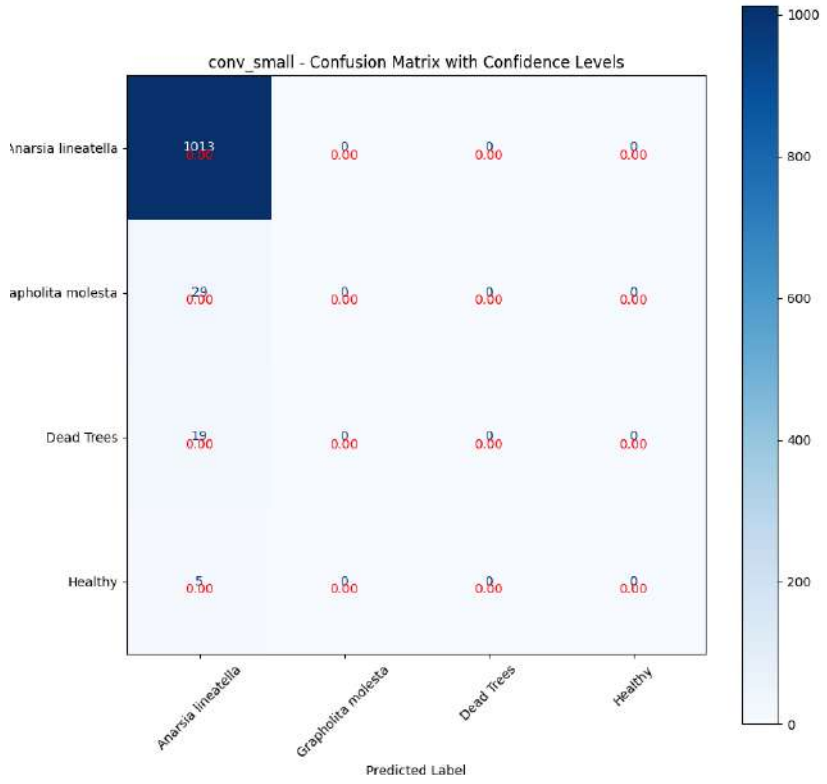


Figure B.15: Confusion Matrix with Confidence for Conv Model with 25 layes

B.1.3.4 Classification Report and Confusion Matrix for Conv Model with 50 layes

Class	Precision	Recall	F1-score
Anarsia lineatella	0.95	1.00	0.97
Grapholita molesta	0.00	0.00	0.00
Dead Trees	0.00	0.00	0.00
Healthy	0.00	0.00	0.00
Accuracy			0.95
Macro avg	0.24	0.25	0.24
Weighted avg	0.90	0.95	0.93

Table B.16: Classification Report for Conv Model with 50 layes

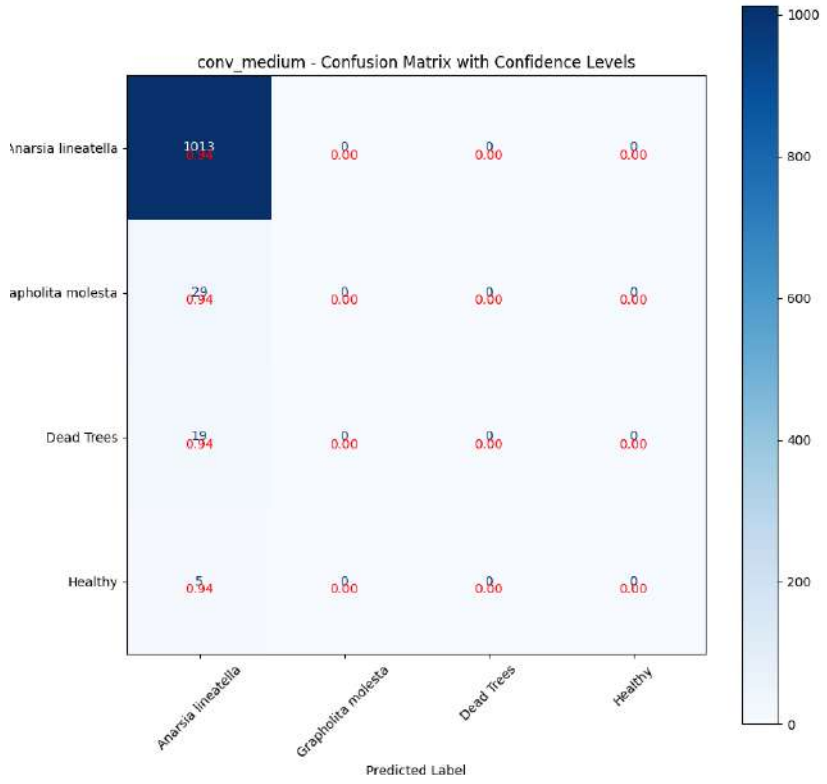


Figure B.16: Confusion Matrix with Confidence for Conv Model with 50 layes

B.2 Multimodal

B.2.1 Intermediate Fusion

B.2.1.1 Classification Report and Confusion Matrix for InceptionV3 and MLP Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.88	1.00	0.94	15
Anarsia lineatella	1.00	1.00	1.00	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.94		
Macro avg	0.47	0.50	0.48	32
Weighted avg	0.88	0.94	0.91	32

Table B.17: Classification Report for InceptionV3 and MLP Model with 25 layers

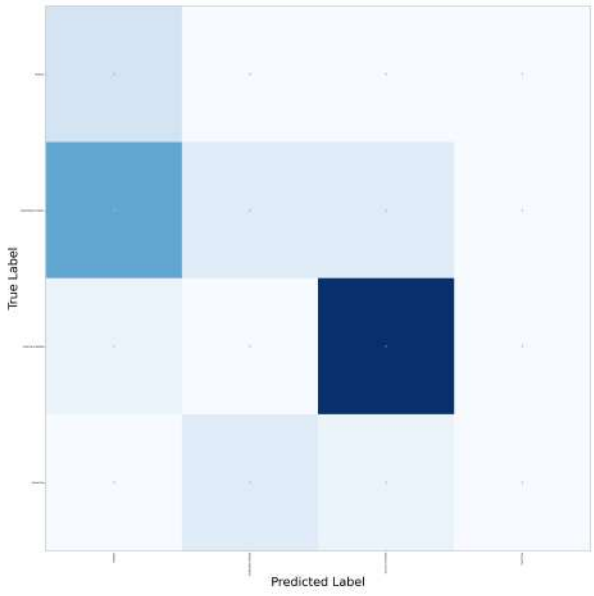


Figure B.17: Confusion Matrix for InceptionV3 and MLP Model with 25 layers

B.2.1.2 Classification Report and Confusion Matrix for ResNet152 and MLP Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	1.00	0.33	0.50	15
Anarsia lineatella	0.56	1.00	0.71	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.62		
Macro avg	0.39	0.33	0.30	32
Weighted avg	0.73	0.62	0.57	32

Table B.18: Classification Report for ResNet152 and MLP Model with 25 layers

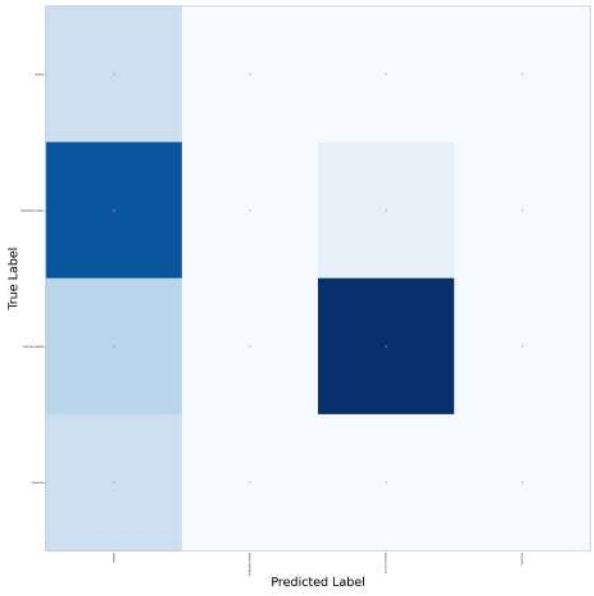


Figure B.18: Confusion Matrix for ResNet152 and MLP Model with 25 layers

B.2.1.3 Classification Report and Confusion Matrix for VGG19 and MLP Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.47	1.00	0.64	15
Anarsia lineatella	0.00	0.00	0.00	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.47		
Macro avg	0.12	0.25	0.16	32
Weighted avg	0.22	0.47	0.30	32

Table B.19: Classification Report for VGG19 and MLP Model with 25 layers

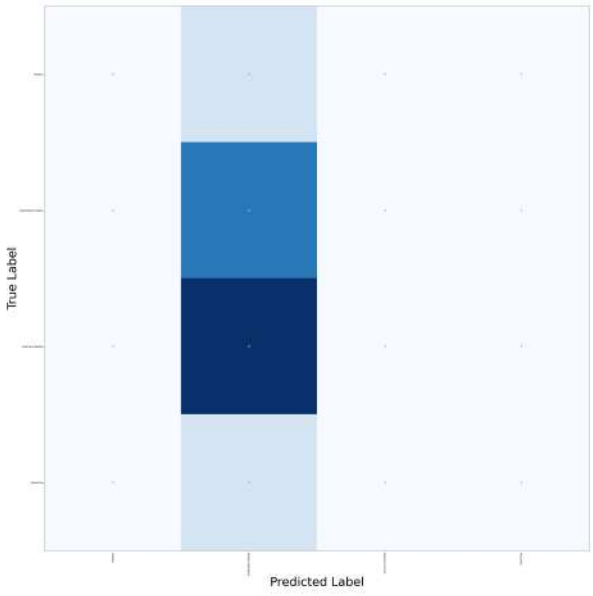


Figure B.19: Confusion Matrix for VGG19 and MLP Model with 25 layers

B.2.1.4 Classification Report and Confusion Matrix for ViT and MLP Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.92	0.73	0.81	15
Anarsia lineatella	0.75	1.00	0.86	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.81		
Macro avg	0.42	0.43	0.42	32
Weighted avg	0.78	0.81	0.78	32

Table B.20: Classification Report for ViT and MLP Model with 25 layers

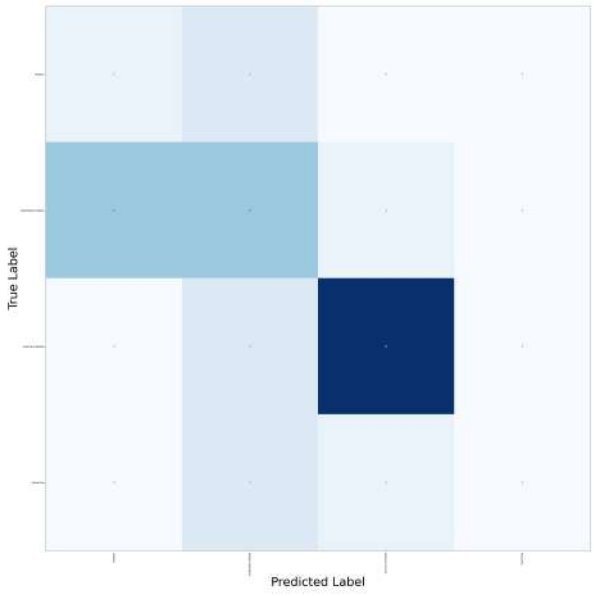


Figure B.20: Confusion Matrix for ViT and MLP Model with 25 layers

B.2.1.5 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and MLP Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.00	0.00	0.00	15
Anarsia lineatella	0.47	1.00	0.64	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.47		
Macro avg	0.12	0.25	0.16	32
Weighted avg	0.22	0.47	0.30	32

Table B.21: Classification Report for AttentionAugmentedInceptionV3 and MLP Model with 25 layers

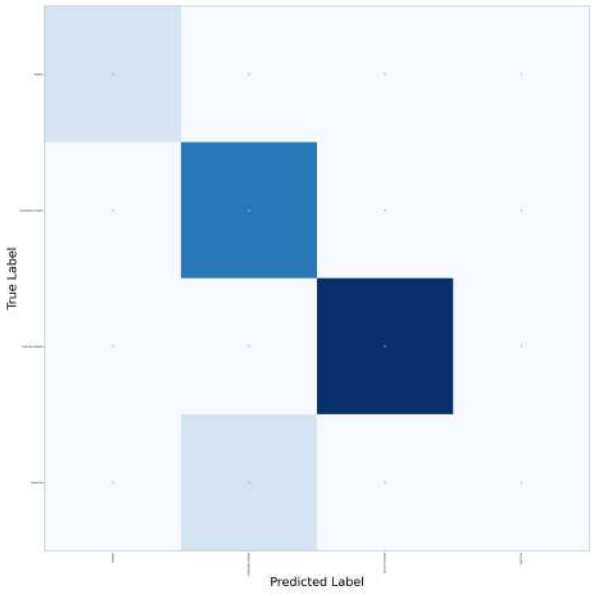


Figure B.21: Confusion Matrix for AttentionAugmentedInceptionV3 and MLP Model with 25 layers

B.2.1.6 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and MLP Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.92	0.73	0.81	15
Anarsia lineatella	0.75	1.00	0.86	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.81		
Macro avg	0.42	0.43	0.42	32
Weighted avg	0.78	0.81	0.78	32

Table B.22: Classification Report for AttentionAugmentedResNet18 and MLP Model with 25 layers

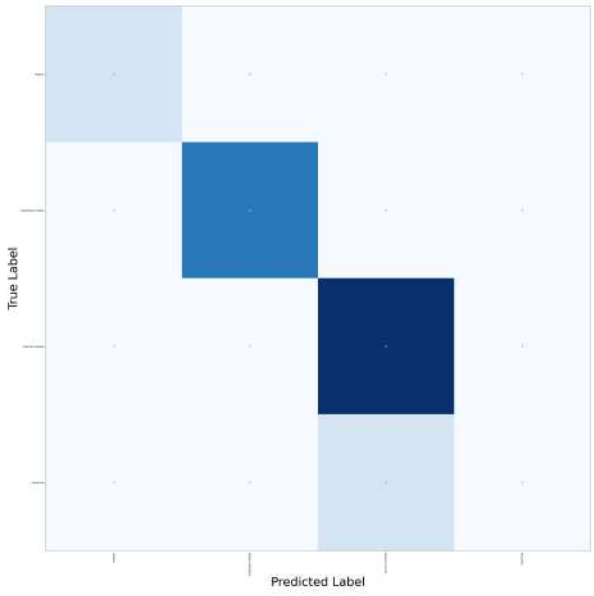


Figure B.22: Confusion Matrix for AttentionAugmentedResNet18 and MLP Model with 25 layers

B.2.1.7 Classification Report and Confusion Matrix for InceptionV3 and MLP Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.83	1.00	0.91	15
Anarsia lineatella	1.00	0.93	0.97	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.91		
Macro avg	0.46	0.48	0.47	32
Weighted avg	0.86	0.91	0.88	32

Table B.23: Classification Report for InceptionV3 and MLP Model with 50 layers

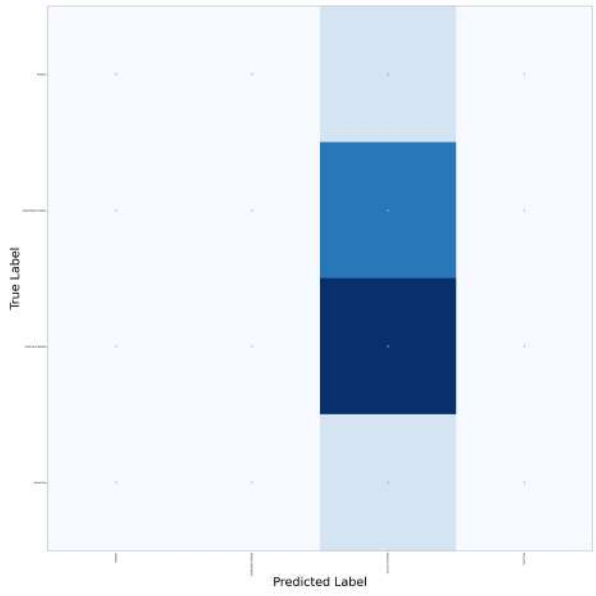


Figure B.23: Confusion Matrix for InceptionV3 and MLP Model with 50 layers

B.2.1.8 Classification Report and Confusion Matrix for ResNet152 and MLP Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	1.00	0.27	0.42	15
Anarsia lineatella	0.54	1.00	0.70	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.59		
Macro avg	0.38	0.32	0.28	32
Weighted avg	0.72	0.59	0.52	32

Table B.24: Classification Report for ResNet152 and MLP Model with 50 layers

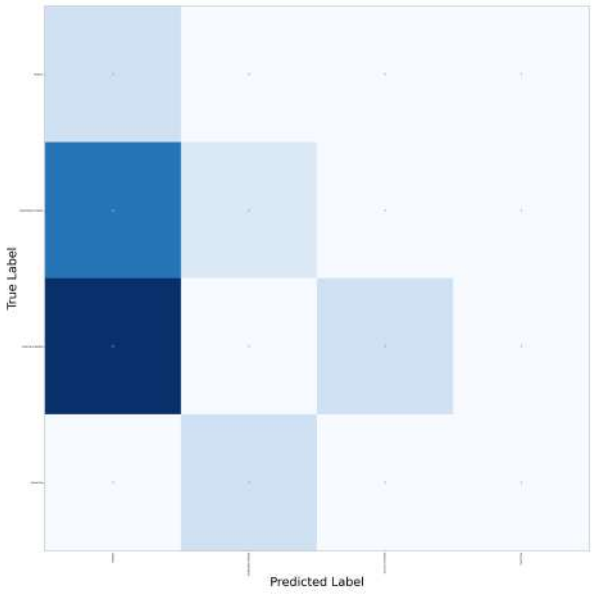


Figure B.24: Confusion Matrix for ResNet152 and MLP Model with 50 layers

B.2.1.9 Classification Report and Confusion Matrix for VGG19 and MLP Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.00	0.00	0.00	15
Anarsia lineatella	0.47	1.00	0.64	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.47		
Macro avg	0.12	0.25	0.16	32
Weighted avg	0.22	0.47	0.30	32

Table B.25: Classification Report for VGG19 and MLP Model with 50 layers

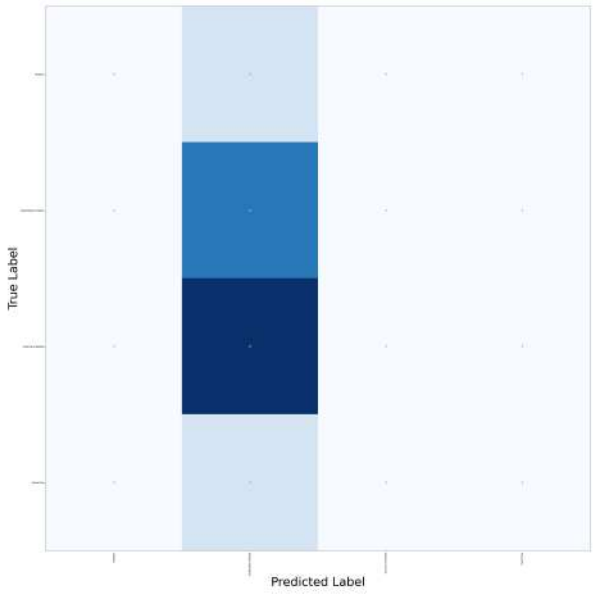


Figure B.25: Confusion Matrix for VGG19 and MLP Model with 50 layers

B.2.1.10 Classification Report and Confusion Matrix for ViT and MLP Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.79	1.00	0.88	15
Anarsia lineatella	1.00	0.87	0.93	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.88		
Macro avg	0.45	0.47	0.45	32
Weighted avg	0.84	0.88	0.85	32

Table B.26: Classification Report for ViT and MLP Model with 50 layers

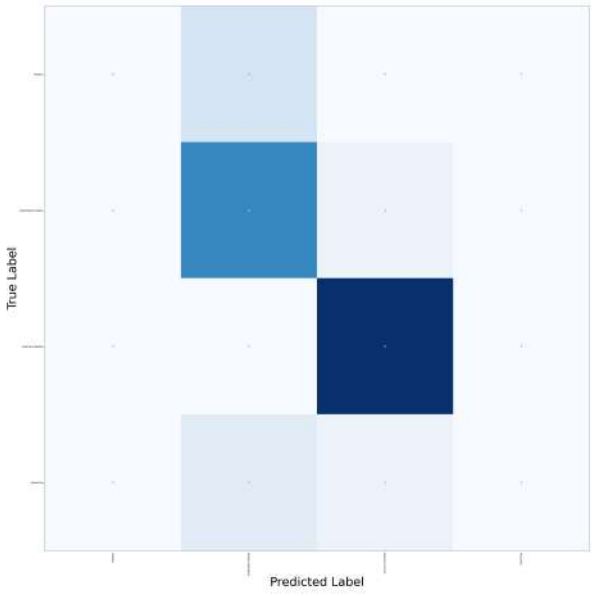


Figure B.26: Confusion Matrix for ViT and MLP Model with 50 layers

B.2.1.11 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and MLP Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.80	0.27	0.40	15
Anarsia lineatella	0.56	1.00	0.71	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.59		
Macro avg	0.34	0.32	0.28	32
Weighted avg	0.64	0.59	0.52	32

Table B.27: Classification Report for AttentionAugmentedInceptionV3 and MLP Model with 50 layers

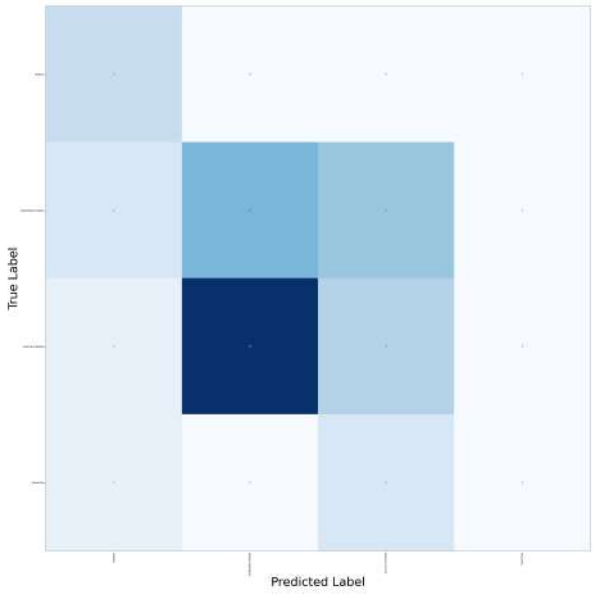


Figure B.27: Confusion Matrix for AttentionAugmentedInceptionV3 and MLP Model with 50 layers

B.2.1.12 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and MLP Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	0
Grapholita molesta	0.92	0.73	0.81	15
Anarsia lineatella	0.75	1.00	0.86	15
Dead Tree	0.00	0.00	0.00	2
Accuracy		0.81		
Macro avg	0.42	0.43	0.42	32
Weighted avg	0.78	0.81	0.78	32

Table B.28: Classification Report for AttentionAugmentedResNet18 and MLP Model with 50 layers

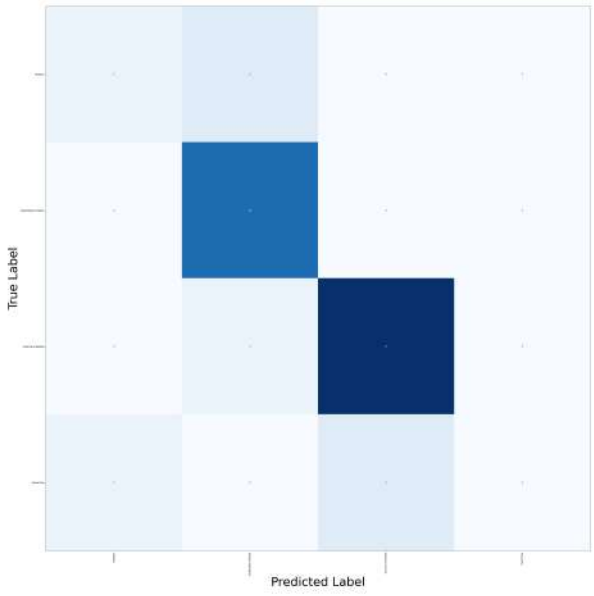


Figure B.28: Confusion Matrix for AttentionAugmentedResNet18 and MLP Model with 50 layers

B.2.1.13 Classification Report and Confusion Matrix for InceptionV3 and Conv Model with 25 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.23	1.00	0.38	3
Grapholita molesta	0.50	0.15	0.24	13
Anarsia lineatella	0.85	0.94	0.89	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.59 (out of 37)		
Macro avg	0.40	0.52	0.38	37
Weighted avg	0.61	0.59	0.55	37

Table B.29: Classification Report for InceptionV3 with Conv Model 25 layers

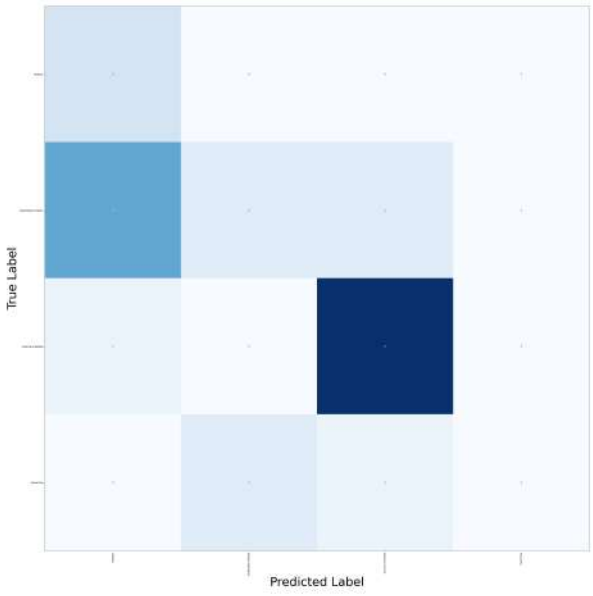


Figure B.29: Confusion Matrix for InceptionV3 with Conv Model 25 layers

B.2.1.14 Classification Report and Confusion Matrix for ResNet152 and Conv Model with 25 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.14	1.00	0.24	3
Grapholita molesta	0.00	0.00	0.00	13
Anarsia lineatella	0.93	0.78	0.85	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.46 (out of 37)		
Macro avg	0.27	0.44	0.27	37
Weighted avg	0.47	0.46	0.43	37

Table B.30: Classification Report for ResNet152 with Conv Model 25 layers

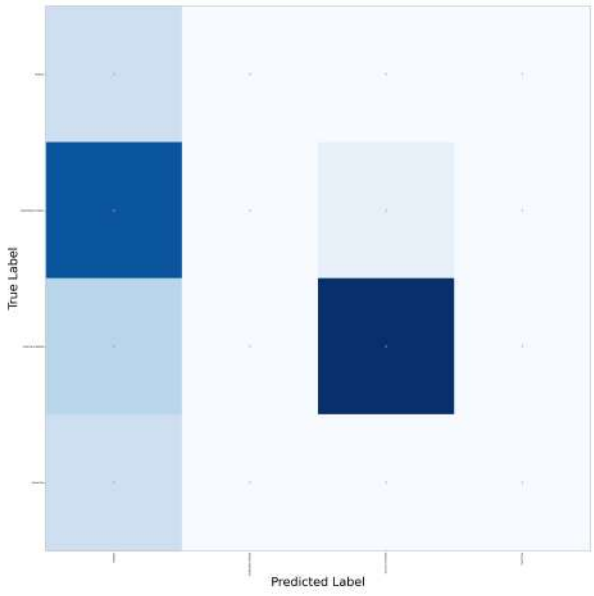


Figure B.30: Confusion Matrix for ResNet152 with Conv Model 25 layers

B.2.1.15 Classification Report and Confusion Matrix for VGG19 and Conv Model with 25 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.35	1.00	0.52	13
Anarsia lineatella	0.00	0.00	0.00	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.35 (out of 37)		
Macro avg	0.09	0.25	0.13	37
Weighted avg	0.12	0.35	0.18	37

Table B.31: Classification Report for VGG19 with Conv Model 25 layers

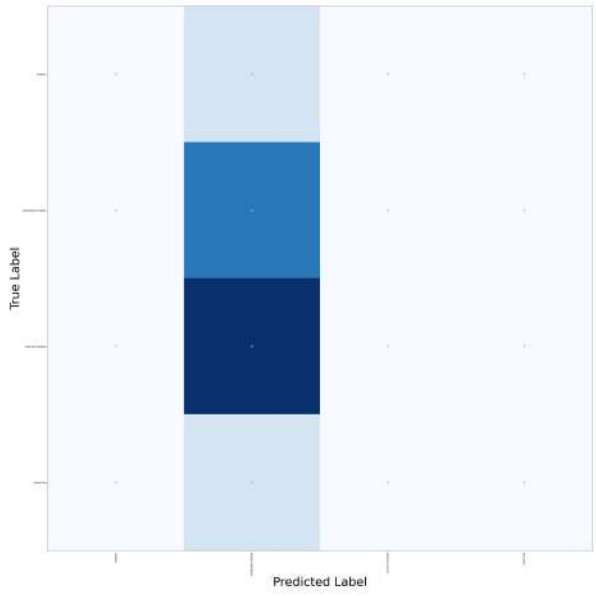


Figure B.31: Confusion Matrix for VGG19 with Conv Model 25 layers

B.2.1.16 Classification Report and Confusion Matrix for ViT and Conv Model with 25 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.14	0.33	0.20	3
Grapholita molesta	0.50	0.46	0.48	13
Anarsia lineatella	0.89	0.89	0.89	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.62 (out of 37)		
Macro avg	0.38	0.42	0.39	37
Weighted avg	0.62	0.62	0.62	37

Table B.32: Classification Report for ViT with Conv Model 25 layers

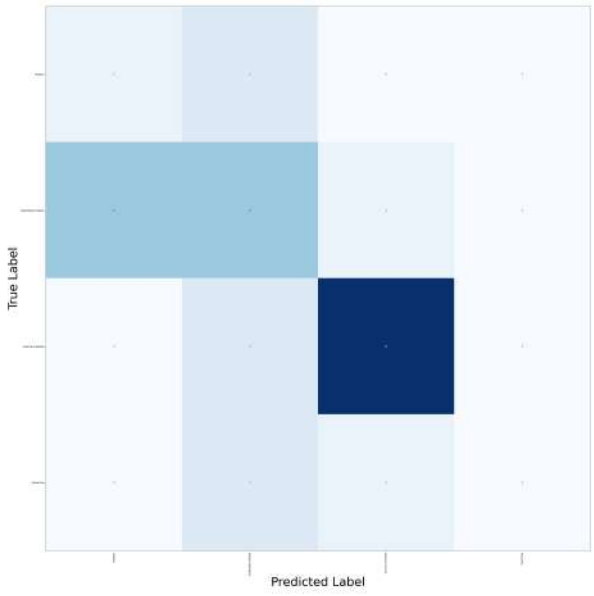


Figure B.32: Confusion Matrix for ViT with Conv Model 25 layers

B.2.1.17 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and Conv Model with 25 layers

Class	Precision	Recall	F1-Score	Support
Healthy	1.00	1.00	1.00	3
Grapholita molesta	0.81	1.00	0.90	13
Anarsia lineatella	1.00	1.00	1.00	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.92 (out of 37)		
Macro avg	0.70	0.75	0.72	37
Weighted avg	0.85	0.92	0.88	37

Table B.33: Classification Report for AttentionAugmentedInceptionV3 with Conv Model 25 layers

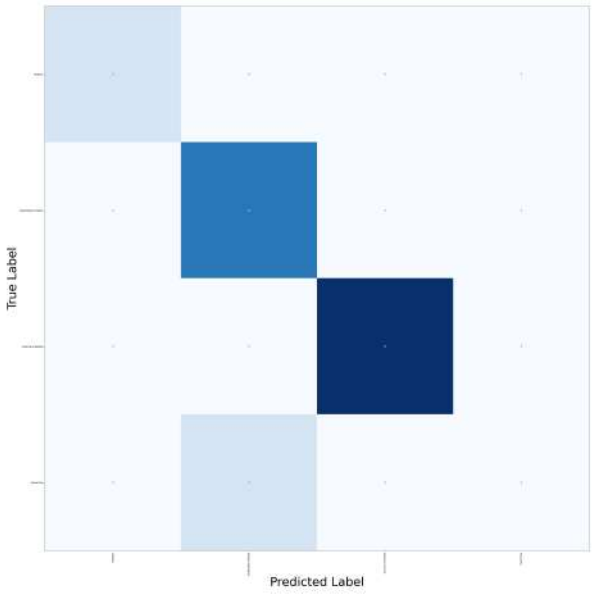


Figure B.33: Confusion Matrix for AttentionAugmentedInceptionV3 with Conv Model 25 layers

B.2.1.18 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and Conv Model with 25 layers

Class	Precision	Recall	F1-Score	Support
Healthy	1.00	1.00	1.00	3
Grapholita molesta	1.00	1.00	1.00	13
Anarsia lineatella	0.86	1.00	0.92	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.92 (out of 37)		
Macro avg	0.71	0.75	0.73	37
Weighted avg	0.85	0.92	0.88	37

Table B.34: Classification Report for AttentionAugmentedResNet18 with Conv Model 25 layers

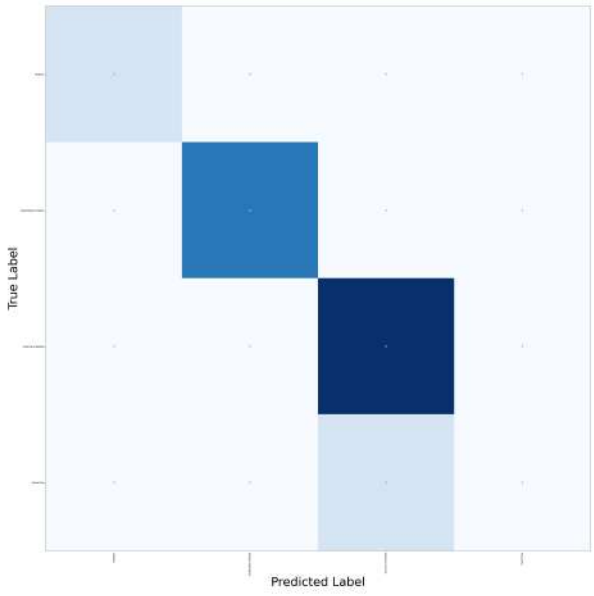


Figure B.34: Confusion Matrix for AttentionAugmentedResNet18 with Conv Model 25 layers

B.2.1.19 Classification Report and Confusion Matrix for InceptionV3 and Conv Model with 50 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.00	0.00	0.00	13
Anarsia lineatella	0.49	1.00	0.65	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.49 (out of 37)		
Macro avg	0.12	0.25	0.16	37
Weighted avg	0.24	0.49	0.32	37

Table B.35: Classification Report for InceptionV3 with Conv Model 50 layers

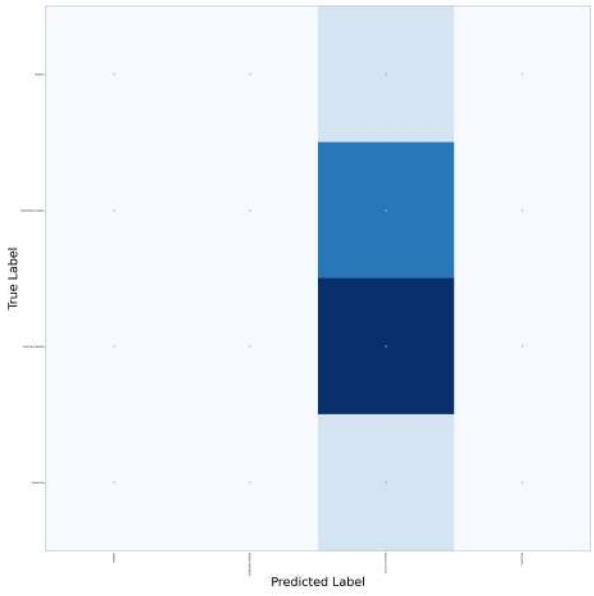


Figure B.35: Confusion Matrix for InceptionV3 with Conv Model 50 layers

B.2.1.20 Classification Report and Confusion Matrix for ResNet152 and Conv Model with 50 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.10	1.00	0.19	3
Grapholita molesta	0.40	0.15	0.22	13
Anarsia lineatella	1.00	0.17	0.29	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.22 (out of 37)		
Macro avg	0.38	0.33	0.17	37
Weighted avg	0.64	0.22	0.23	37

Table B.36: Classification Report for ResNet152 with Conv Model 50 layers

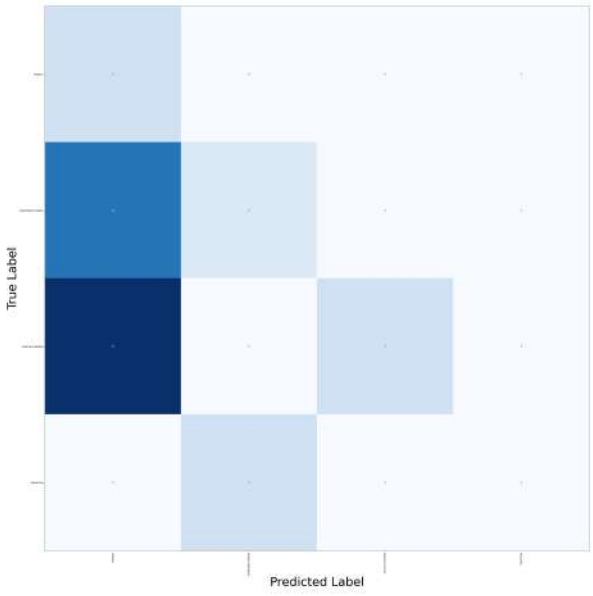


Figure B.36: Confusion Matrix for ResNet152 with Conv Model 50 layers

B.2.1.21 Classification Report and Confusion Matrix for VGG19 and Conv Model with 50 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.35	1.00	0.52	13
Anarsia lineatella	0.00	0.00	0.00	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.35 (out of 37)		
Macro avg	0.09	0.25	0.13	37
Weighted avg	0.12	0.35	0.18	37

Table B.37: Classification Report for VGG19 with Conv Model 50 layers

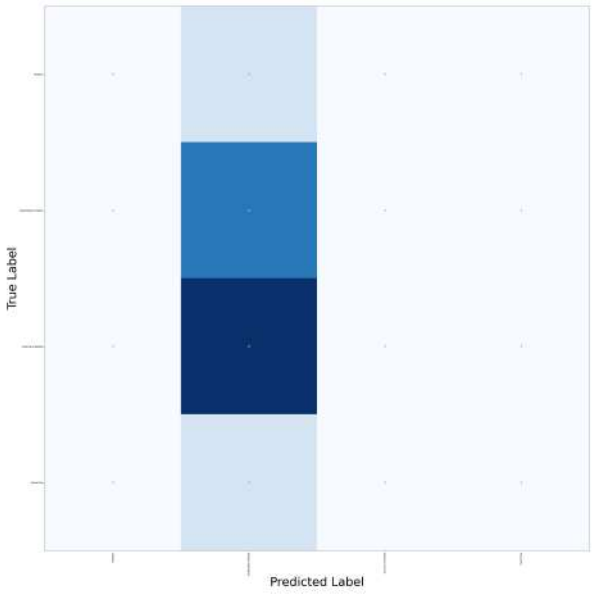


Figure B.37: Confusion Matrix for VGG19 with Conv Model 50 layers

B.2.1.22 Classification Report and Confusion Matrix for ViT and Conv Model with 50 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.71	0.92	0.80	13
Anarsia lineatella	0.90	1.00	0.95	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.81 (out of 37)		
Macro avg	0.40	0.48	0.44	37
Weighted avg	0.69	0.81	0.74	37

Table B.38: Classification Report for ViT with Conv Model 50 layers

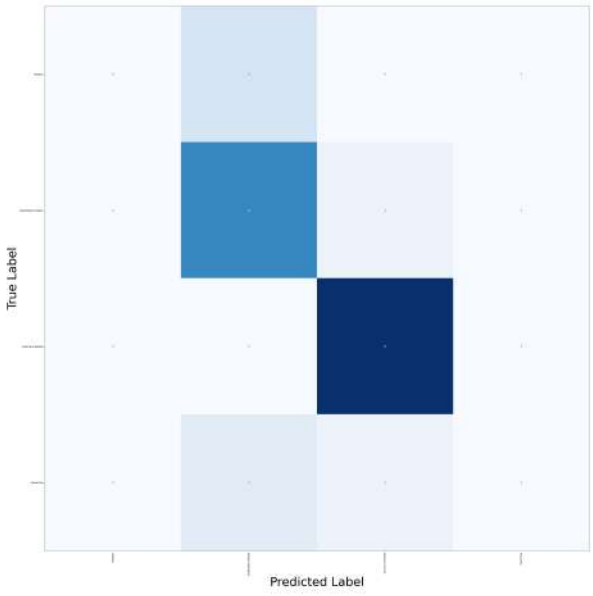


Figure B.38: Confusion Matrix for ViT with Conv Model 50 layers

B.2.1.23 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and Conv Model with 50 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.43	1.00	0.60	3
Grapholita molesta	0.32	0.46	0.38	13
Anarsia lineatella	0.36	0.22	0.28	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.35 (out of 37)		
Macro avg	0.28	0.42	0.31	37
Weighted avg	0.32	0.35	0.31	37

Table B.39: Classification Report for AttentionAugmentedInceptionV3 with Conv Model 50 layers

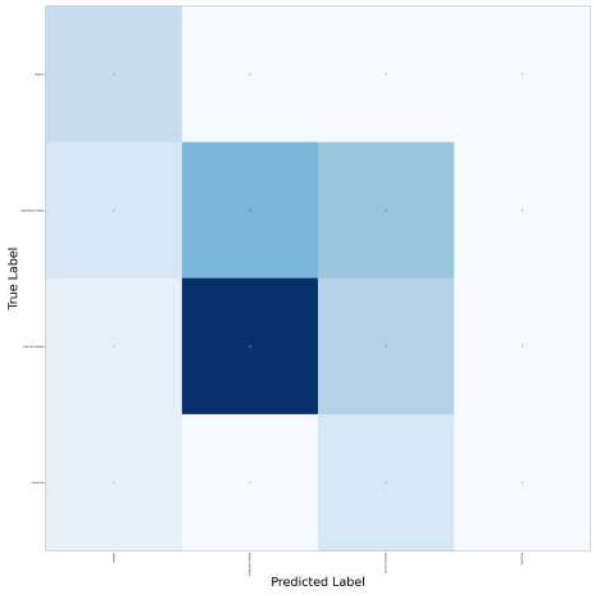


Figure B.39: Confusion Matrix for AttentionAugmentedInceptionV3 with Conv Model 50 layers

B.2.1.24 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and Conv Model with 50 layers

Class	Precision	Recall	F1-Score	Support
Healthy	0.50	0.33	0.40	3
Grapholita molesta	0.81	1.00	0.90	13
Anarsia lineatella	0.89	0.94	0.92	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.84 (out of 37)		
Macro avg	0.55	0.57	0.55	37
Weighted avg	0.76	0.84	0.79	37

Table B.40: Classification Report for AttentionAugmentedResNet18 with Conv Model 50 layers

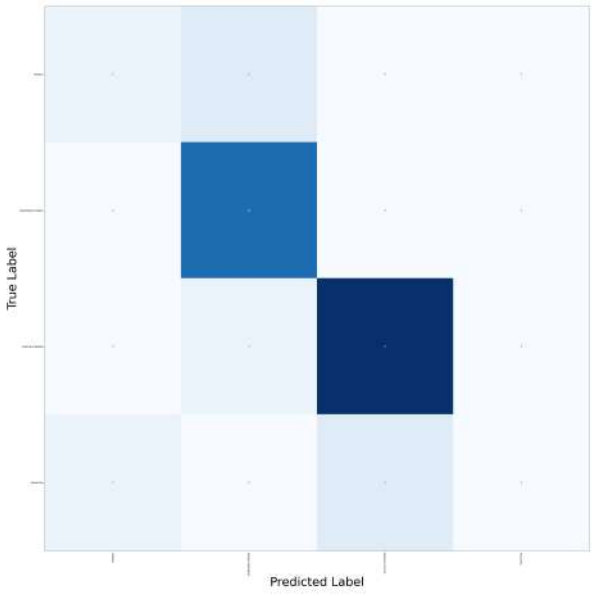


Figure B.40: Confusion Matrix for AttentionAugmentedResNet18 with Conv Model 50 layers

B.2.2 Late Fusion

B.2.2.1 Classification Report and Confusion Matrix for InceptionV3 and MLP with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	1.00	0.67	0.80	3
Grapholita molesta	0.81	1.00	0.90	13
Anarsia lineatella	0.95	1.00	0.97	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.89		
Macro avg	0.69	0.67	0.67	37
Weighted avg	0.83	0.89	0.85	37

Table B.41: Classification Report for InceptionV3 and MLP with 25 layers

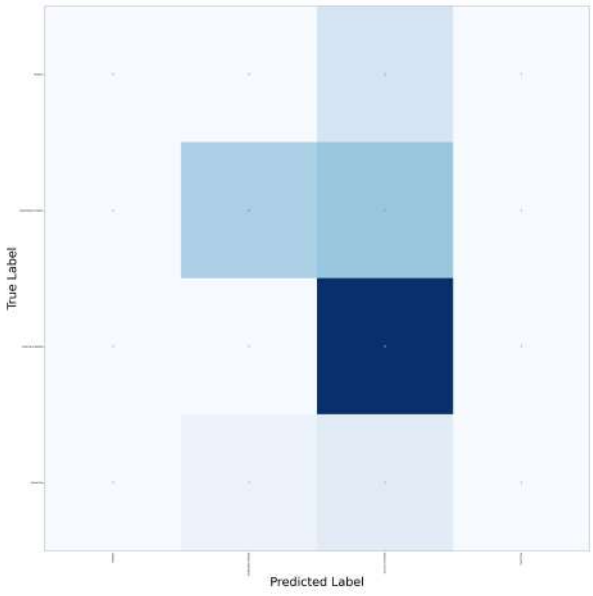


Figure B.41: Confusion Matrix for InceptionV3 and MLP with 25 layers

B.2.2.2 Classification Report and Confusion Matrix for ResNet152 and MLP with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.00	0.00	0.00	13
Anarsia lineatella	0.50	0.28	0.36	18
Dead Tree	0.23	1.00	0.38	3
Accuracy		0.22		
Macro avg	0.18	0.32	0.18	37
Weighted avg	0.26	0.22	0.20	37

Table B.42: Classification Report for ResNet152 and MLP with 25 layers

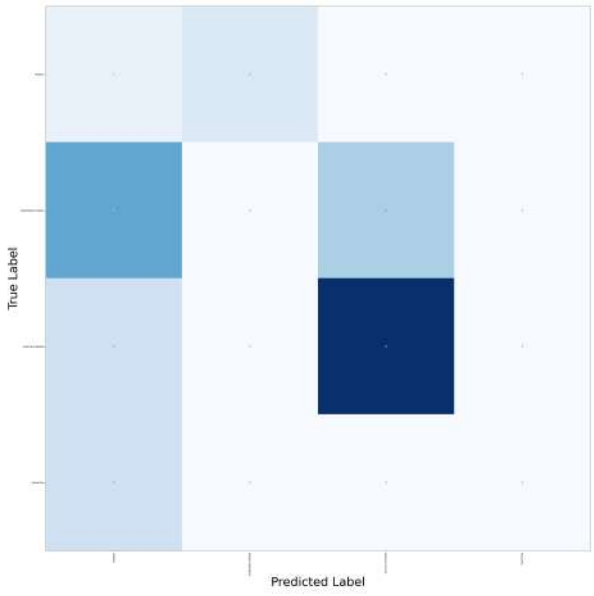


Figure B.42: Confusion Matrix for ResNet152 and MLP with 25 layers

B.2.2.3 Classification Report and Confusion Matrix for VGG19 and MLP with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.35	1.00	0.52	13
Anarsia lineatella	0.00	0.00	0.00	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.35		
Macro avg	0.09	0.25	0.13	37
Weighted avg	0.12	0.35	0.18	37

Table B.43: Classification Report for VGG19 and MLP with 25 layers

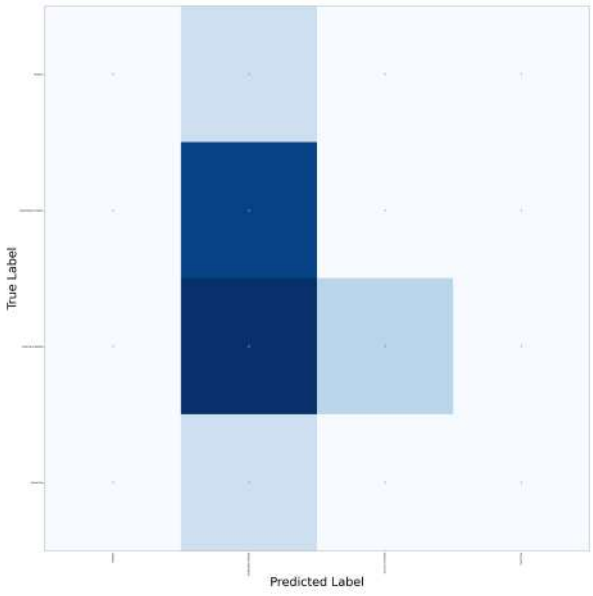


Figure B.43: Confusion Matrix for VGG19 and MLP with 25 layers

B.2.2.4 Classification Report and Confusion Matrix for ViT and MLP with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.62	1.00	0.76	13
Anarsia lineatella	1.00	0.89	0.94	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.78		
Macro avg	0.40	0.47	0.43	37
Weighted avg	0.70	0.78	0.73	37

Table B.44: Classification Report for ViT and MLP with 25 layers

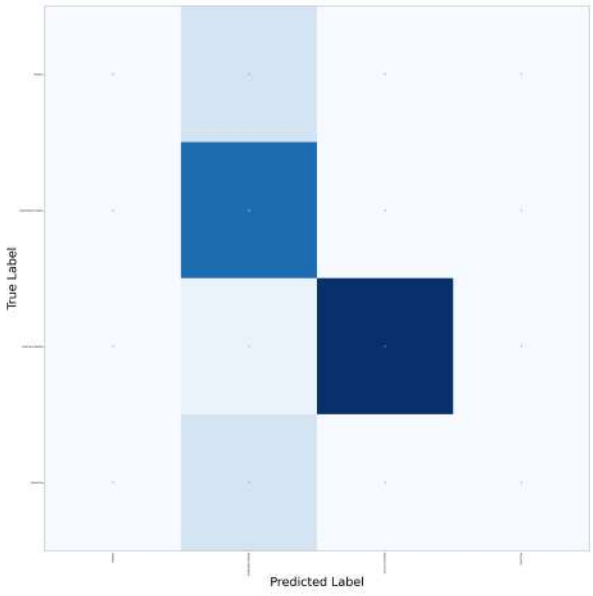


Figure B.44: Confusion Matrix for ViT and MLP with 25 layers

B.2.2.5 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and MLP with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	1.00	1.00	1.00	3
Grapholita molesta	0.93	1.00	0.96	13
Anarsia lineatella	1.00	1.00	1.00	18
Dead Tree	1.00	0.67	0.80	3
Accuracy			0.97	
Macro avg	0.98	0.92	0.94	37
Weighted avg	0.97	0.97	0.97	37

Table B.45: Classification Report for AttentionAugmentedInceptionV3 and MLP with 25 layers

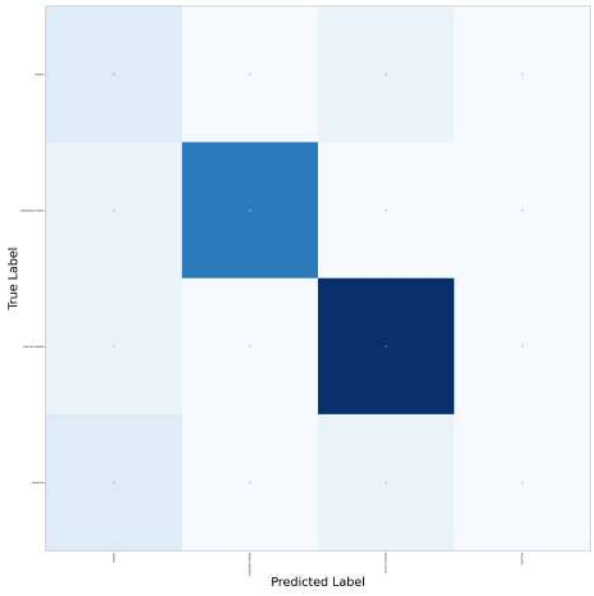


Figure B.45: Confusion Matrix for AttentionAugmentedInceptionV3 and MLP with 25 layers

B.2.2.6 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and MLP with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.33	0.67	0.44	3
Grapholita molesta	1.00	0.92	0.96	13
Anarsia lineatella	0.89	0.94	0.92	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.84		
Macro avg	0.56	0.63	0.58	37
Weighted avg	0.81	0.84	0.82	37

Table B.46: Classification Report for AttentionAugmentedResNet18 and MLP with 25 layers

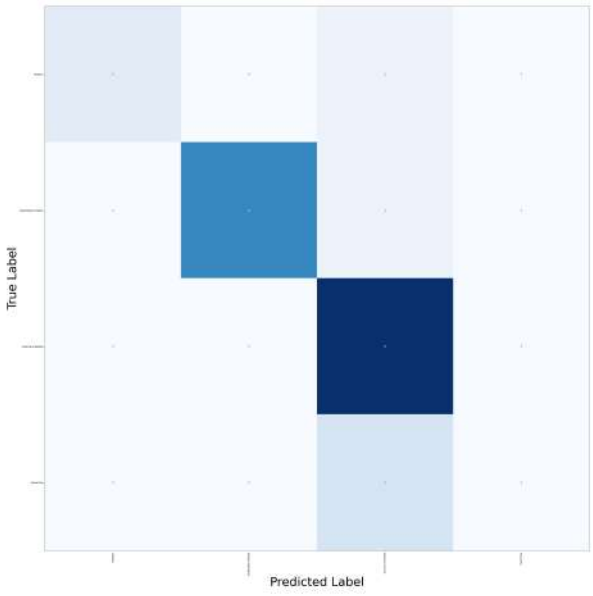


Figure B.46: Confusion Matrix for AttentionAugmentedResNet18 and MLP with 25 layers

B.2.2.7 Classification Report and Confusion Matrix for InceptionV3 and MLP with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.75	1.00	0.86	3
Grapholita molesta	0.62	1.00	0.76	13
Anarsia lineatella	1.00	0.61	0.76	18
Dead Tree	1.00	0.33	0.50	3
Accuracy		0.76		
Macro avg	0.84	0.74	0.72	37
Weighted avg	0.85	0.76	0.75	37

Table B.47: Classification Report for InceptionV3 and MLP with 50 layers

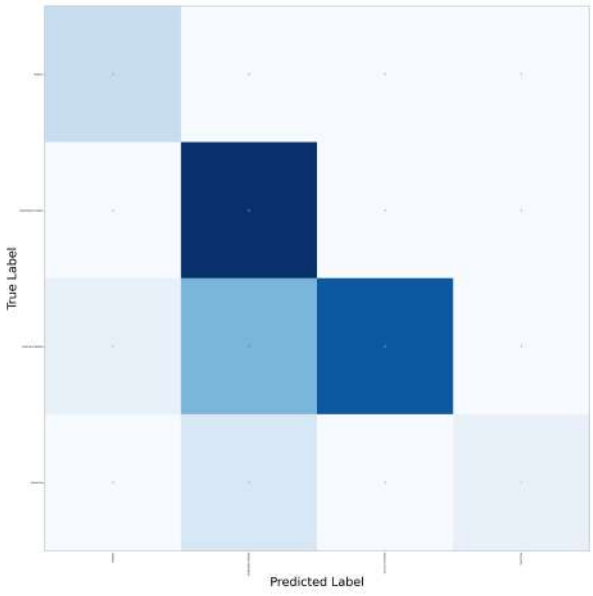


Figure B.47: Confusion Matrix for InceptionV3 and MLP with 50 layers

B.2.2.8 Classification Report and Confusion Matrix for ResNet152 and MLP with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.00	0.00	0.00	13
Anarsia lineatella	0.53	1.00	0.69	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.49		
Macro avg	0.13	0.25	0.17	37
Weighted avg	0.26	0.49	0.34	37

Table B.48: Classification Report for ResNet152 and MLP with 50 layers

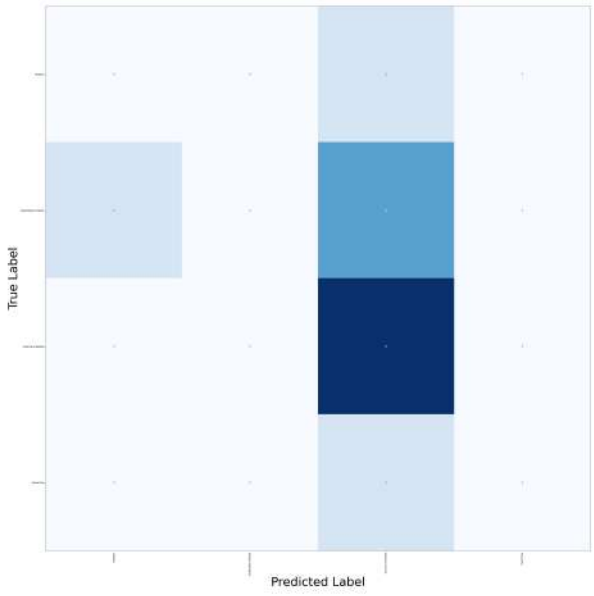


Figure B.48: Confusion Matrix for ResNet152 and MLP with 50 layers

B.2.2.9 Classification Report and Confusion Matrix for VGG19 and MLP with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.30	0.69	0.42	13
Anarsia lineatella	0.00	0.00	0.00	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.24		
Macro avg	0.07	0.17	0.10	37
Weighted avg	0.11	0.24	0.15	37

Table B.49: Classification Report for VGG19 and MLP with 50 layers

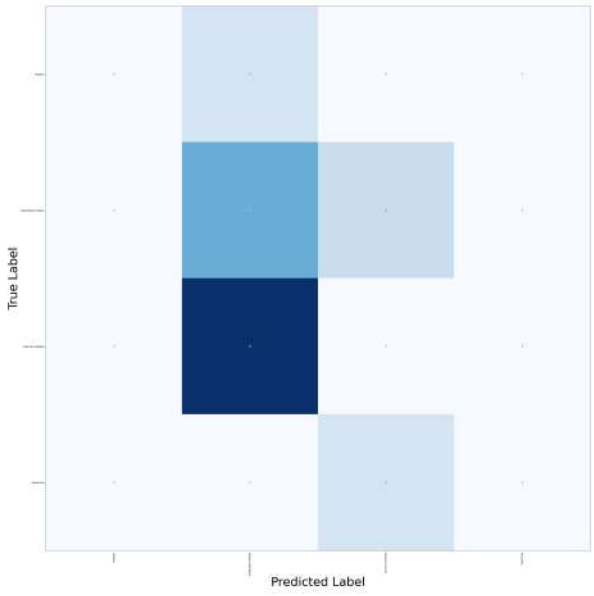


Figure B.49: Confusion Matrix for VGG19 and MLP with 50 layers

B.2.2.10 Classification Report and Confusion Matrix for ViT and MLP with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.59	1.00	0.74	13
Anarsia lineatella	1.00	0.83	0.91	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.76		
Macro avg	0.40	0.46	0.41	37
Weighted avg	0.69	0.76	0.70	37

Table B.50: Classification Report for ViT and MLP with 50 layers

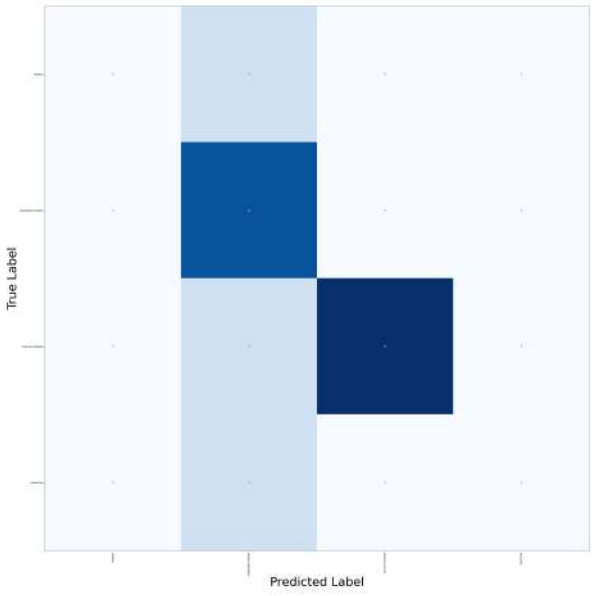


Figure B.50: Confusion Matrix for ViT and MLP with 50 layers

B.2.2.11 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and MLP with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.60	1.00	0.75	3
Grapholita molesta	0.73	0.85	0.79	13
Anarsia lineatella	0.94	0.89	0.91	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.81		
Macro avg	0.57	0.68	0.61	37
Weighted avg	0.76	0.81	0.78	37

Table B.51: Classification Report for AttentionAugmentedInceptionV3 and MLP with 50 layers

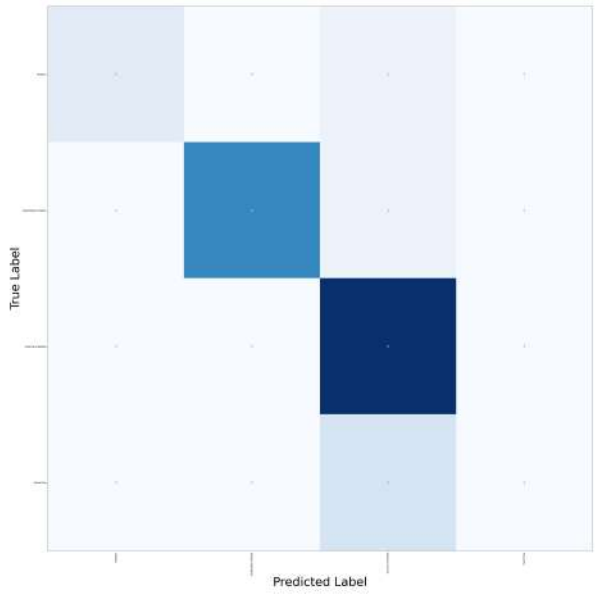


Figure B.51: Confusion Matrix for AttentionAugmentedInceptionV3 and MLP with 50 layers

B.2.2.12 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and MLP with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	1.00	0.33	0.50	3
Grapholita molesta	1.00	0.77	0.87	13
Anarsia lineatella	0.69	1.00	0.82	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.78		
Macro avg	0.67	0.53	0.55	37
Weighted avg	0.77	0.78	0.74	37

Table B.52: Classification Report for AttentionAugmentedResNet18 and MLP with 50 layers

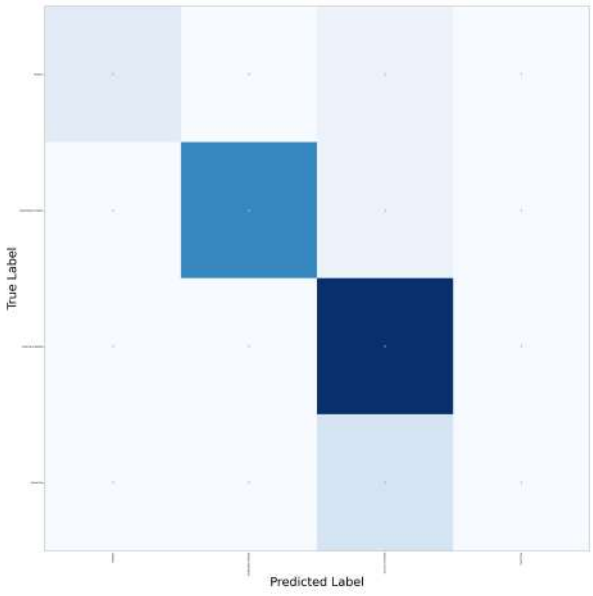


Figure B.52: Confusion Matrix for AttentionAugmentedResNet18 and MLP with 50 layers

B.2.2.13 Classification Report and Confusion Matrix for InceptionV3 and Conv Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.86	0.46	0.60	13
Anarsia lineatella	0.60	1.00	0.75	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.65		
Macro avg	0.36	0.37	0.34	37
Weighted avg	0.59	0.65	0.58	37

Table B.53: Classification Report for InceptionV3 and Conv Model with 25 layers

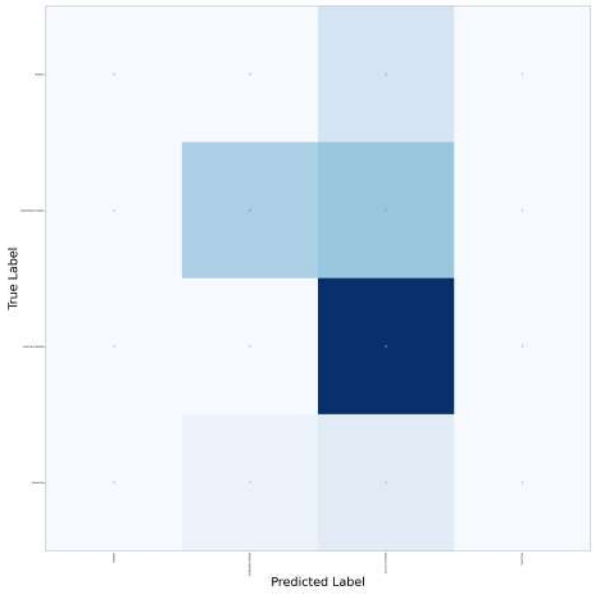


Figure B.53: Confusion Matrix for InceptionV3 and Conv Model with 25 layers

B.2.2.14 Classification Report and Confusion Matrix for ResNet152 and Conv Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.07	0.33	0.11	3
Grapholita molesta	0.00	0.00	0.00	13
Anarsia lineatella	0.75	0.83	0.79	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.43		
Macro avg	0.20	0.29	0.23	37
Weighted avg	0.37	0.43	0.39	37

Table B.54: Classification Report for ResNet152 and Conv Model with 25 layers

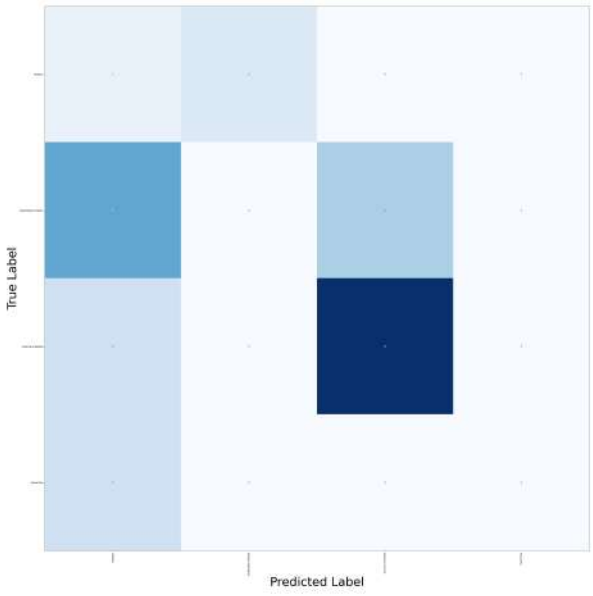


Figure B.54: Confusion Matrix for ResNet152 and Conv Model with 25 layers

B.2.2.15 Classification Report and Confusion Matrix for VGG19 and Conv Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.39	1.00	0.57	13
Anarsia lineatella	1.00	0.22	0.36	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.46		
Macro avg	0.35	0.31	0.23	37
Weighted avg	0.62	0.46	0.38	37

Table B.55: Classification Report for VGG19 and Conv Model with 25 layers

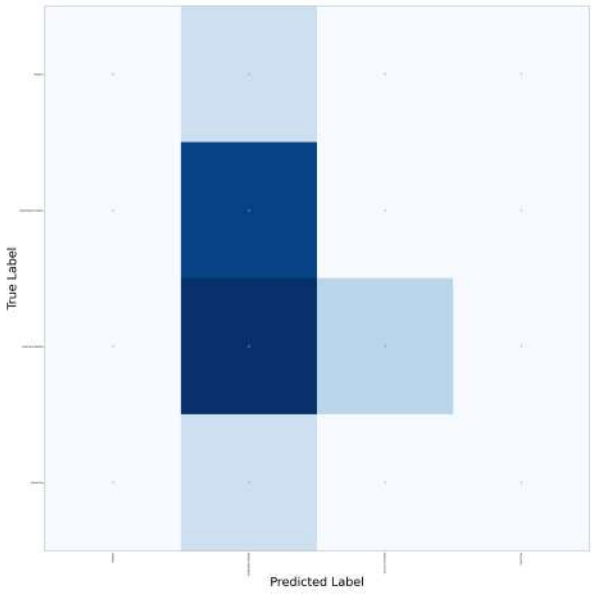


Figure B.55: Confusion Matrix for VGG19 and Conv Model with 25 layers

B.2.2.16 Classification Report and Confusion Matrix for ViT and Conv Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.65	1.00	0.79	13
Anarsia lineatella	1.00	0.94	0.97	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.81		
Macro avg	0.41	0.49	0.44	37
Weighted avg	0.71	0.81	0.75	37

Table B.56: Classification Report for ViT and Conv Model with 25 layers

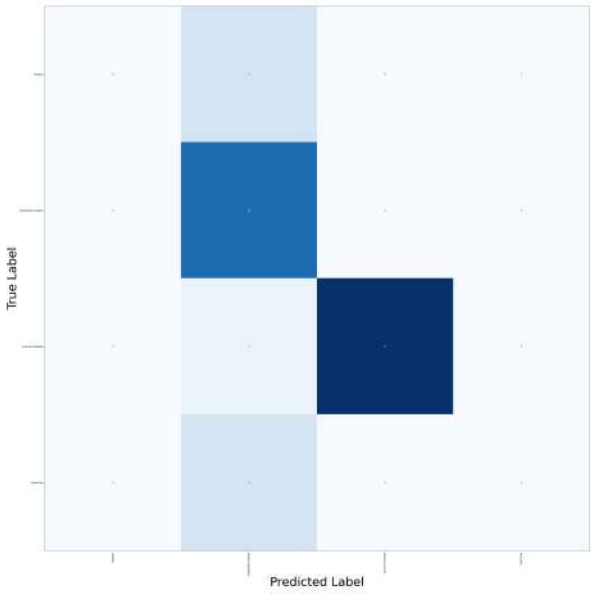


Figure B.56: Confusion Matrix for ViT and Conv Model with 25 layers

B.2.2.17 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and Conv Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	1.00	1.00	1.00	3
Grapholita molesta	0.93	1.00	0.96	13
Anarsia lineatella	1.00	1.00	1.00	18
Dead Tree	1.00	0.67	0.80	3
Accuracy		0.97		
Macro avg	0.98	0.92	0.94	37
Weighted avg	0.97	0.97	0.97	37

Table B.57: Classification Report for AttentionAugmentedInceptionV3 and Conv Model with 25 layers

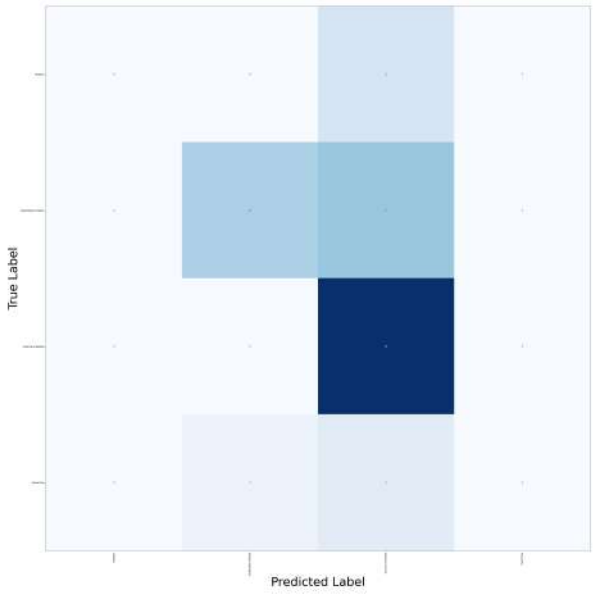


Figure B.57: Confusion Matrix for AttentionAugmentedInceptionV3 and Conv Model with 25 layers

B.2.2.18 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and Conv Model with 25 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.33	0.67	0.44	3
Grapholita molesta	1.00	0.92	0.96	13
Anarsia lineatella	0.89	0.94	0.92	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.84		
Macro avg	0.56	0.63	0.58	37
Weighted avg	0.81	0.84	0.82	37

Table B.58: Classification Report for AttentionAugmentedResNet18 and Conv Model with 25 layers

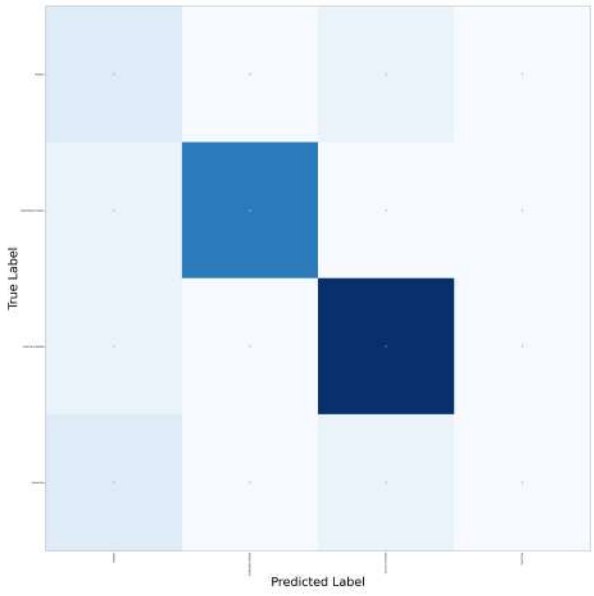


Figure B.58: Confusion Matrix for AttentionAugmentedResNet18 and Conv Model with 25 layers

B.2.2.19 Classification Report and Confusion Matrix for InceptionV3 and Conv Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	1.00	0.33	0.50	3
Grapholita molesta	0.65	1.00	0.79	13
Anarsia lineatella	1.00	0.89	0.94	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.81		
Macro avg	0.66	0.56	0.56	37
Weighted avg	0.80	0.81	0.78	37

Table B.59: Classification Report for InceptionV3 and Conv Model with 50 layers

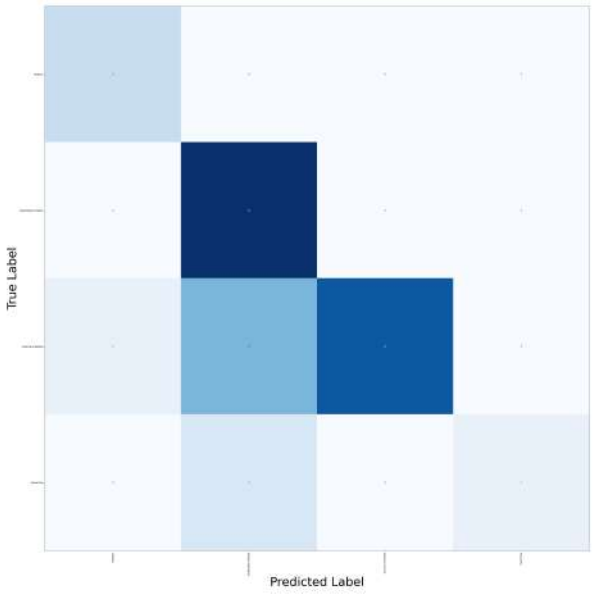


Figure B.59: Confusion Matrix for InceptionV3 and Conv Model with 50 layers

B.2.2.20 Classification Report and Confusion Matrix for ResNet152 and Conv Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.10	1.00	0.18	3
Grapholita molesta	0.86	0.46	0.60	13
Anarsia lineatella	0.00	0.00	0.00	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.24		
Macro avg	0.24	0.37	0.20	37
Weighted avg	0.31	0.24	0.23	37

Table B.60: Classification Report for ResNet152 and Conv Model with 50 layers

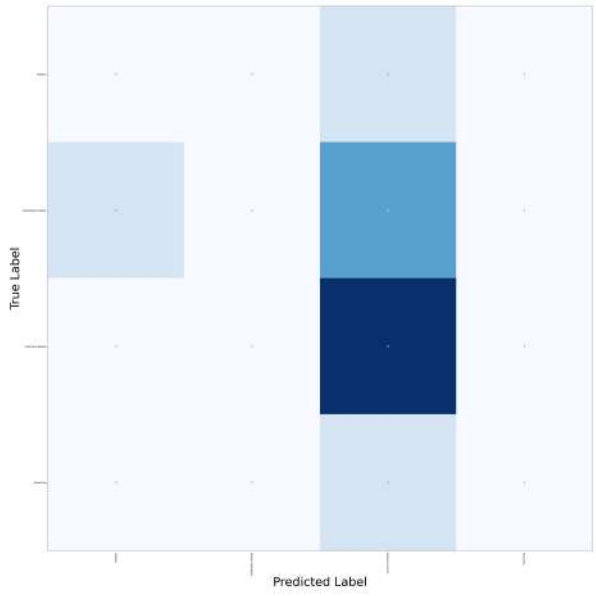


Figure B.60: Confusion Matrix for ResNet152 and Conv Model with 50 layers

B.2.2.21 Classification Report and Confusion Matrix for VGG19 and Conv Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.00	0.00	0.00	13
Anarsia lineatella	0.49	1.00	0.65	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.49		
Macro avg	0.12	0.25	0.16	37
Weighted avg	0.24	0.49	0.32	37

Table B.61: Classification Report for VGG19 and Conv Model with 50 layers

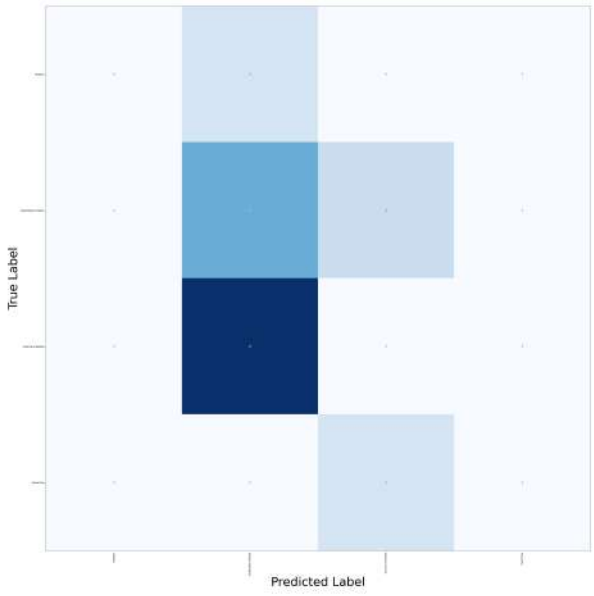


Figure B.61: Confusion Matrix for VGG19 and Conv Model with 50 layers

B.2.2.22 Classification Report and Confusion Matrix for ViT and Conv Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.00	0.00	0.00	3
Grapholita molesta	0.59	1.00	0.74	13
Anarsia lineatella	1.00	0.83	0.91	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.76		
Macro avg	0.40	0.46	0.41	37
Weighted avg	0.69	0.76	0.70	37

Table B.62: Classification Report for ViT and Conv Model with 50 layers

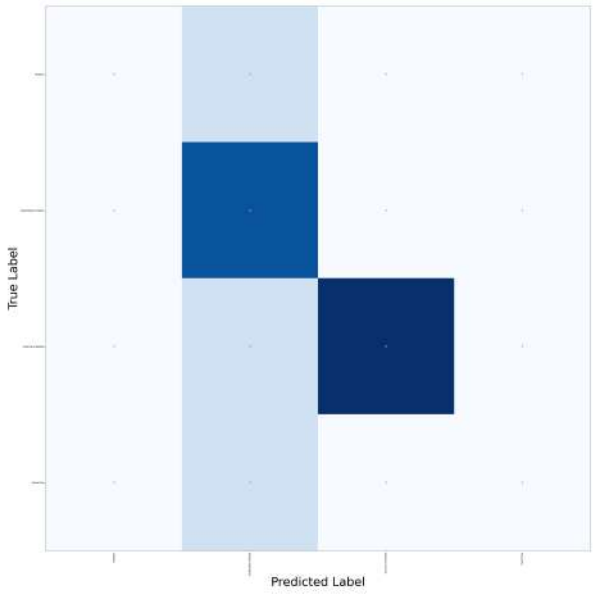


Figure B.62: Confusion Matrix for ViT and Conv Model with 50 layers

B.2.2.23 Classification Report and Confusion Matrix for AttentionAugmentedInceptionV3 and Conv Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	0.60	1.00	0.75	3
Grapholita molesta	0.73	0.85	0.79	13
Anarsia lineatella	0.94	0.89	0.91	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.81		
Macro avg	0.57	0.68	0.61	37
Weighted avg	0.76	0.81	0.78	37

Table B.63: Classification Report for AttentionAugmentedInceptionV3 and Conv Model with 50 layers

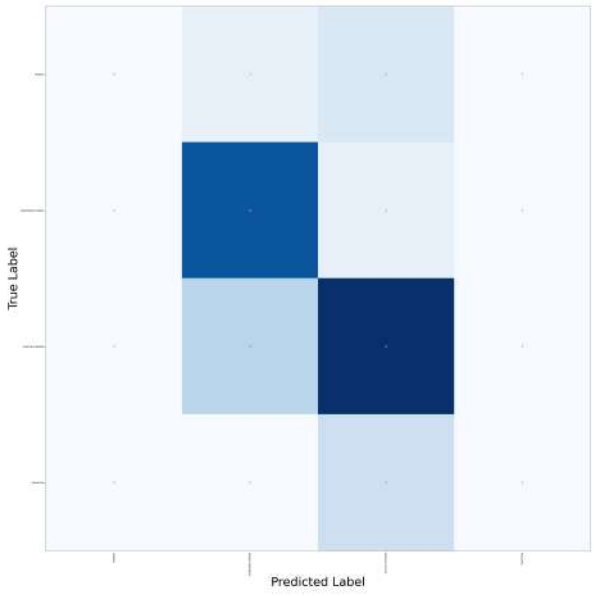


Figure B.63: Confusion Matrix for AttentionAugmentedInceptionV3 and Conv Model with 50 layers

B.2.2.24 Classification Report and Confusion Matrix for AttentionAugmentedResNet18 and Conv Model with 50 layers

Class	Precision	Recall	F1-score	Support
Healthy	1.00	0.33	0.50	3
Grapholita molesta	1.00	0.77	0.87	13
Anarsia lineatella	0.69	1.00	0.82	18
Dead Tree	0.00	0.00	0.00	3
Accuracy		0.78		
Macro avg	0.67	0.53	0.55	37
Weighted avg	0.77	0.78	0.74	37

Table B.64: Classification Report for AttentionAugmentedResNet18 and Conv Model with 50 layers

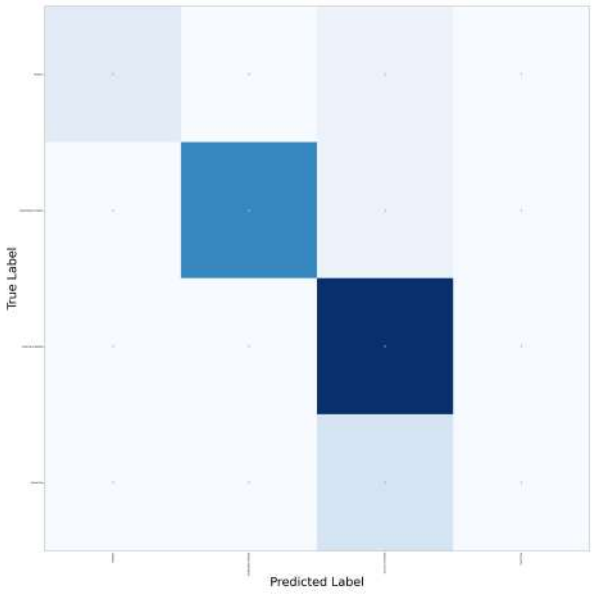


Figure B.64: Confusion Matrix for AttentionAugmentedResNet18 and Conv Model with 50 layers