UNIVERSITI TEKNOLOGI MARA

CONVOLUTIONAL NEURAL NETWORK WITH ENHANCED VISION TRANSFORMER MODEL FOR DEFECT CLASSIFICATION OF ADDITIVE MANUFACTURING PRODUCTS

NUR NAJIHA BINTI KAMARULZAMAN

MSc

June 2026

UNIVERSITI TEKNOLOGI MARA

CONVOLUTIONAL NEURAL NETWORK WITH ENHANCED VISION TRANSFORMER MODEL FOR DEFECT CLASSIFICATION OF ADDITIVE MANUFACTURING PRODUCTS

NUR NAJIHA BINTI KAMARULZAMAN

Thesis submitted in fulfilment
of the requirements for the degree of
Master of Science
(Electrical Engineering)

Faculty of Electrical Engineering

June 2026

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 13 May 2026 to conduct the final examination of Nur Najiha Binti Kamarulzaman on her Masters of Science thesis entitled "Convolutional Neural Network with Enhanced Vision Transformer Model for Defect Classification of Additive Manufacturing Products" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiner recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Muhammad Khusairi Osman, PhD
Associate Professor
Faculty of Electrical Engineering
Universiti Teknologi MARA
(Chairman)

Mohd Firdaus Abdullah, PhD
Senior Lecturer
Faculty of Elctrical Engineering
Universiti Teknologi MARA
(Internal Examiner)

Effariza Hanafi, PhD
Associate Professor
Faculty of Engineering
University Malaya
(External Examiner)

PROFESOR DR HJH ZURAEDA IBRAHIM
Dean
Institute of Postgraduates Studies
Universiti Teknologi MARA

Date: 15 June 2026

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Nur Najiha Binti Kamarulzaman
Student ID. No.	:	2024656996
Programme	:	Master of Science (Electrical Engineering) – CEEE750
Faculty	:	Electrical Engineering
Thesis Title	:	Convolutional Neural Network with Enhanced Vision Transformer Model for Defect Classification of Additive Manufacturing Products
Signature of Student	:	……..
Date	:	June 2026

ABSTRACT

Additive manufacturing (AM), or 3D printing, refers to a process of joining materials layer by layer to create objects. The AM process faces inconsistent outcomes and poses challenges in maintaining quality control. Manual quality control can be time-consuming. Traditional defect inspection method has evolved into an automated strategy by leveraging the benefit of artificial intelligence (AI). Machine learning (ML) is a subset in the AI field, can support decision-making for process optimization, quality control, and system improvement. The existing methods used ML have only focused on the defect classification without considering local and global feature extraction. Local and global feature extraction play an important role, as different types of defects require different feature extraction. The CNN algorithm, as a local features extraction widely used to classify the defects from AM products. However, the CNN unable to identify the relationship between the extracted features, prevents the model from understanding the global context of the images. Vision transformer (ViT), is a global feature extraction model, which enables a more comprehensive understanding of the global context within an image. However, ViT remains limited in capturing local features as transformer rely heavily on global relationships. Hybrid architectures combining CNN and ViT have emerged, leveraging the advantages of both architecture of CNN ViT. However, challenges occur when ViT divides the original input images into a series of non-overlapping patches in fixed size, in which this method destructs the image local-continuity in some degree. In addition, ViT is often determined by some tokens with large amounts of information, whereas most tokens are redundant. To address the limitations, this study proposed a deep learning model introducing CNN with enhanced ViT (CNN-enhanced ViT) to classify types of AM defects including crack, stringing, and no defect arising from the AM process. The main contribution of CNN-enhanced ViT relies on its architecture. The proposed class token extracted from the CNN feature map may serve as a reference for global information. In addition, this study further contributes to enhance the ViT architecture by performing token selection using 1D maxpooling layer. On the other hand, previous study has performed token selection which involves ViT architecture only, but to the best of author's knowledge, no studies have performed CNN-ViT together with token selection, especially in AM scope. In this study, a total of 1500 AM images were collected from the Malaysia Automotive, Robotics and IoT Institute (MARii), which consists of defect classes of crack and stringing and a no defect class. The performance of the proposed CNN-enhanced ViT model was then evaluated and compared with the CNN-ViT model and Swin transformer models using confusion matrix. The confusion matrix results showed that the proposed model CNN-enhanced ViT achieved the best results, demonstrating a value of average accuracy of 96.00% and average precision of 94.28%, sensitivity of 94.00% and F1-score of 94.01% accordingly. The results demonstrate that the proposed CNN-enhanced ViT model is potentially able to produce effective and accurate results in defect classification for AM products, offering a lightweight and scalable solution for intelligent visual inspection system. This work has potential in contributing to the advancement of defect classification systems especially in smart manufacturing industry.

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to Allah SWT for granting me strength, patience, and perseverance for giving me the opportunity to embark on my MsC and for completing this journey successfully. Without His blessings and guidance, this research would not have been possible.

I would like to extend my sincere appreciation and heartfelt thanks to my supervisor, Associate Professor Ir. Dr. Nor Salwa Damanhuri, for the continuous guidance, valuable advice, and unwavering support throughout this research. Her knowledge, encouragement, and constructive feedback have greatly contributed to the successful completion of this study. Her dedication and commitment to academic excellence have inspired me to further enhance my understanding and research capabilities. My gratitude and thanks also go to lecturers who help me along the journey, Associate Professor Dr. Noor Azlina Mohd Salleh, Dr. Belinda Chong Chiew Meng, Associate Professor Ir. Dr. Nor Azlan Othman and Dr. Nur Sa'adah Muhamad Sauki

Special thanks are also extended to the Malaysia Automotive, Robotics and IoT Institute (MARii) for providing the dataset used in this study. The availability of this dataset has played a crucial role in enabling the development, training, and evaluation of the proposed deep learning models for defect classification.

I would also like to acknowledge my friends and colleagues for their encouragement, motivation, and support. Their assistance, whether through academic discussions or moral support, has helped me overcome many challenges during the research process.

Most importantly, I would like to express my deepest gratitude to my beloved parents and family members for their endless love, prayers, patience, and sacrifices. Special thanks to my mother, and my siblings for not even missed to support me every single day to complete this journey. Their emotional and moral support has been my greatest source of strength and motivation throughout this journey. Without their encouragement and belief in me, this achievement would not have been possible. This journey is dedicated to my very dear late father. This piece of victory is dedicated to you.

Finally, I would like to thank everyone who has directly or indirectly contributed to the completion of this thesis. Their support and encouragement are sincerely appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations

AI	Artificial Intelligence
AM	Additive Manufacturing
CAD	Computer-Aided Design
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Tree
FDM	Fused Deposition Modelling
FFF	Fused Filament Fabrication
MARii	Malaysia Automotive, Robotics, And Iot Institute
ML	Machine Learning
NDT	Non-Destructive Test
SDG	Sustainable Development Goals
STL	Standard Tessellation Language
SVM	Support Vector Machine
ViT	Vision Transformer

CHAPTER 1
INTRODUCTION

1.1 Research Background

Additive manufacturing (AM), or 3D printing, refers to a process of joining materials layer by layer to create objects, which is opposed to subtractive manufacturing [1]. This growing technology enables object customization built on demand despite give significant impact to modern industry, as development for mass customization might improve sustainability by reducing waste and energy consumption [2]. However, AM is still susceptible to various types of defects. The defects that occur during the 3D-printing process lead the deterioration in the quality of the printed product [3]. Certain post-processing techniques are used to ensure the printed product has an acceptable surface finish, which require lots of time [4]. In addition, conventional methods of inspecting AM products, such as visual inspection and non-destructive test are tedious, ineffective use of manpower, time-consuming and not suitable for large-scale production [5].

The conventional defect inspection method has evolved into an automated strategy by leveraging the benefit of artificial intelligence (AI), which is capable of mimicking the human brain and performing tasks like what humans can do [6]. Machine learning (ML) is a subset in the AI field, which allows the machine to learn from the example such as images as input and thus improve the performance over time by using the model to make predictions [7]. Furthermore, ML can support decision-making for process optimization, quality control, and system improvement [8]. Zhang et al. stated an automated systems can lower operational costs despite enhancing the speed, precision, and reliability through model optimization [9].

Previously, researchers have started to integrate the convolutional neural network (CNN) algorithm, as a local features extraction method, to classify the defects from AM products. For example, Chen et al. has designed a CNN based model to recognize normal filamentary materials and materials with different degrees of defects [10]. On the other hand, Kozhay et al. utilized CNN-based methods for detecting anomalies such as layer shift and over-extrusion that arise from fused deposition

modelling (FDM) 3D printers, ensuring thorough quality assessment while lowering calibration error [11]. CNN-based models perform well in identifying defects since CNN can extract from low-level features to high-level features. Furthermore, CNN is capable of solving a wide variety of problems, as it was able to generalize learned features during training [12]. However, the CNN unable to identify the relationship between the extracted features, prevents the model from understanding the global context of the images [13]. Understanding the whole context of the image is important for the defect classification task since surface defects span across regions, necessitate the recognition of whole defect patterns.

The key limitation of CNN architecture lies in its inability to effectively capture global contextual information [14]. Vision transformer (ViT), on the other hand, is a global feature extraction model, which provides an alternative approach that enables a more comprehensive understanding of the global context within an image [15]. For example, a study by Wang et al. employed ViT, a transformer-based approach, to classify the printing quality regarding different levels of extrusion from the fused filament fabrication (FFF) printing method [16]. Similarly, Vasan et al. implements ViT in classifying six types of steel surface defects while configuring optimal parameters with 96.39% accuracy [17]. These studies indicate that ViT exhibits excellent performance in image classification tasks by effectively capturing the global context of input images. Self-attention mechanism in ViT also contributes to comprehension of the relationship between different regions of an image [13].

Despite the powerful global context capability in ViT, ViT remains limited in capturing local features as transformer blocks rely heavily on global relationships and model depth, while potentially overlooking local feature details [18]. In addition, improvement need to be performed as challenges occur when ViT divides the original input images into a series of non-overlapping patches in fixed size, in which this method destructs the image local- continuity in some degree, and also limits the ViT performance in visual tasks [19], despite leading to marginalization of information and loss of local tokens features [18]. Furthermore, the calculated amount of ViT increases exponentially as the number of tokens increases, and the final prediction result of ViT is often determined by some tokens with large amounts of information, whereas most of tokens are redundant [19].

Recent models have been developed as the architectures combining local and global extraction offer a robust method for tackling complex visual tasks [14]. A study

by Dai et al. proposed CoAtNet which combines local feature extraction using convolution and global feature extraction using self-attention in a single model [20]. It stacks convolutional blocks (MBConv) before the transformer blocks and the self-attention in transformer being replaced with relative attention. The proposed model generalizes well in large dataset like Imagenet-1k. Another study by Liu et al. develops swin transformer that use window-based self-attention to extract local features and shifted window being used for interaction between the extracted features [21]. Swin transformers construct hierarchical representations by starting from small-sized patches and gradually merging the patches in deeper layers. The proposed model is efficient for segmentation and detection tasks because it produces multi-scale feature maps similar to CNN algorithms. However, the reliance on window-based self-attention causes global information to propagate gradually across layers, which may limit the performance in image classification.

In contrast, hybrid architectures that combine CNN and ViT can overcome some limitations of the swin transformer by leveraging the local feature extraction capabilities of CNN and the global attention mechanism of ViT. However, there are at least two considerations to combine both architectures. First consideration is the complexity in designing the architecture, in which combining CNN and ViT is difficult due to differences in feature hierarchies and scale representation [22]. Second consideration is the training efficiency, as complex feature integration can challenge deployment on edge devices with limited computational resources.

Hence, in this study, a deep learning model introducing CNN with enhanced ViT (CNN-enhanced ViT) is developed to classify types of AM defects, including crack, stringing, and no defect product within small dataset. According to Koppe, dataset less than 10,000 are "small" dataset [23]. The enhancement was made to cope with the drawback of ViT by performing feature map division rather than original image division to mitigate the destruction of image local continuity and to perform token selection through selecting the informative token rather than input all redundant tokens for further processing.

1.2 Motivation

ML has shown promising result in computer vision task including image classification when the technology first emerged. However, the accuracy often depends on large dataset, which owing to inadequate data preprocessing and generalization capabilities, affecting the classification precision [24]. Deep learning (DL), a subset of ML has continue to emerge as it can extract more precise features involving large amount of data [14]. In recent years, the application of DL in the AM sector has become increasingly widespread. DL especially CNN has shown excellent performance in AM defect classification. However, traditional deep learning methods for AM defect classification primarily rely on a single CNN architecture to analyzes image data. This method has effectively addressed the challenges of information extraction and low recognition efficiency in complex environments. While classification models based on CNN have shown good fitting ability, CNN do not have excellent global representation ability due to the limitation of receptive fields [25]. This approach often overlooks the global features of images which highlighting a significant drawback.

The introduction of a ViT helps models better understand the interactions between distant parts of an image. ViT models demonstrate impressive transfer learning capabilities tailored to specific domains such as AM when pre-trained on extensive datasets such as ImageNet. The utilization of ViT for AM defect classification reduces the reliance on expert identification, significantly enhancing detection speed and accuracy. In addition to using pure CNN for classification tasks, some researchers have explored combining CNN and transformers with the ability to capture long-range dependencies, aiming to fully integrate the advantages of both CNN and transformers to improve classification performance. Combining CNN with ViT allows models to not only capture detailed information but also understand how these details connect in the larger context. This hybrid approach enables models to process highly localized features and long-range dependencies, offering a robust method for tackling complex visual tasks. This combination reflects a trend in deep learning model development, which involves cross-domain borrowing and integrating different models' strengths to solve more complex challenges. However, during the training deep learning model, greater accuracy often has a large number of parameters. Therefore, finding a balance between accuracy and computational resources remains a significant challenge.

The motivation of this study is to develop a CNN-enhanced ViT model in order to balance between performance and computational efficiency. The data collection involved the AM dataset supplied by MARii, consist of two types of AM defects, crack and stringing and no defect product. Model development involve the original model of ViT developed by Dosovitskiy et al. [15]. The proposed model named as CNN-enhanced ViT use CNN in order to extract local features. The output from the CNN (feature map) is then set as the input to vision transformer as it will be process by patch embedding and positional encoding layer in ViT. Not like previous study, this study also saved the extracted features from the CNN as the class token. In addition, this study enhances the model by performing token selection in order to increase computational efficiency. The result in this study present in confusion matrix to measure the performance of the proposed model and compared with another model. From enhancement lead to more precise in classifying types of AM defects and differentiate with no defect product. Furthermore, the automation of identifying the AM defect features can reduce the time consumption when performing defect classification.

1.3 Problem Statement

Additive manufacturing (AM) is one of the key aspects in industrial revolution 4.0 (IR4.0), commonly utilized for quick prototyping products on a small scale. However, AM is still susceptible to various types of defects, which are affected by the material properties or process failure that lead to deterioration in the quality of printed products [3]. The defects that occur during the 3D-printing process led to the waste of material, time and resources [26], which does not fulfil the requirement of high-value industries [27]. Therefore, it is important to perform defect classification of AM products, to improve the quality inspection process and reduce the inspection time.

Conventional methods of inspecting AM products, such as visual inspection and NDT method were implemented to inspect the quality of AM parts. These methods depend greatly on the worker's expertise and unable to satisfy the continuously increasing quality standards in the manufacturing industry [28]. Furthermore, the inspection process can be time-consuming due to the intricate geometric design of the 3D-printed product, which necessitates a detailed inspection, and each printed product must be measured individually. In addition, NDT method only focused on defect identification, without classifying types of defects. For example, infrared imaging being used to detect defects but not classify the types of defects [29]. Thus, the demand for an advanced and automated approach in assessing the surface quality and classifying the defects has risen in order to cope with the problems.

The ML approach is more suitable for defect detection for AM product than conventional methods. ML capable in adapting complex data structures, effective feature extraction, perform data analysis and decision making without the assistance of special programmed software. There are five major algorithms that can be used to perform defect classification including regression, bayesian, support vector machine (SVM), decision tree (DT) and deep learning (DL). However, recent developed model including regression, bayesian, SVM and DT are unable to capture defects accurately, as it depends on manual parameter tuning and cannot adapt to non-linear data.

DL, a subset of machine learning has proven to be exceptionally successful in classification since it can adapt to the lighting, colour, shape, and size of the input images, as it is beneficial for detecting complex surface defects [30]. However, the existing models only focus on the defect classification without considering local and global feature extraction, which play an important role in determining the quality of the

AM product. The model used, particularly CNN architectures has capabilities in extracting local feature [31], as convolution operation correlate pixel within an area, but unable to capture global features [25]. While ViT serves as an alternative to extract global features through the self-attention mechanism and manages to analyze the connection between different parts of an image synchronously [32], it ignores local information of the input data. Motivate to use the strength of both architectures, combining both architectures into a hybrid model may provide satisfactory inductive bias and global modelling capabilities [25].

However, the original ViT divides the original input images into a series of non-overlapping patches in fixed size, in which this method destructs the image local-continuity in some degree, and also limits the ViT performance in visual tasks [19], despite leading to marginalization of information and loss of local tokens features [18]. In addition, the calculated amount of ViT increases exponentially as the number of tokens increases, and the final prediction result of ViT is often determined by some tokens with large amounts of information, whereas most of tokens are redundant [19]. On the other hand, previous study has performed token selection which involves ViT architecture only, but to the best of author's knowledge, no studies have performed CNN-ViT together with token selection, especially in AM scope. In addition, many developed models require large amounts of datasets to effectively learn the defect features, which in real world cases, there is limited number of datasets including various types of defects, since a particular defect merely occurs during printing.

This study aims to address the challenges by utilizing the advancements of deep learning, specifically developing a CNN-enhanced ViT model for classifying the defects of AM products, as it offers a promising way in handling the problems. The proposed model aims to work with small number of AM dataset, and the enhancement made to handle with the drawback of ViT by performing feature map division rather than original image division to mitigate destruction of image local-continuity and to perform token selection through selecting the informative token rather than input all redundant tokens for further processing. The developed model will be train and validate using the datasets of AM images and evaluate its performance using appropriate metrics. Then to determine the robustness, the performance of the developed model will be compared with original CNN-ViT and Swin transformer model using the same dataset and environment setting.

1.4 Research Objectives

The primary focus of this study is to develop a CNN-enhanced ViT model to classify defects in AM products according to crack, stringing and no defect class. The specific research objectives are as follows:

a) To develop an enhanced model of CNN-ViT model in extracting local and global features of the AM images and classify the defect according to crack, stringing and no defect product.

b) To train and validate the proposed enhance CNN-ViT model using the datasets of AM images and evaluate its performance using appropriate metrics (eg accuracy, precision, recall and F1-score).

c) To compare the performance of the proposed CNN-ViT model with original CNN-ViT and Swin transformer model using the same dataset and environment setting.

1.5 Significance of Study

This study proposed a hybrid deep learning model, CNN-enhanced ViT, for defect classification of additive manufacturing products, specifically used to classify between crack, stringing, and no defect products. The main significance of this study is to provide knowledge on how a hybrid deep learning model can effectively perform local feature extraction together with global feature extraction to improve the accuracy of defect classification of AM products. Through the proposed class token extracted from the CNN feature map and combined with the patch embedding and positional encoding in the ViT architecture, this study addresses the limitations reported in previous literature. Specifically, when ViT divides the original input images into a series of fixed-size non-overlapping patches, this process destructs the image local-continuity in some degree [19], leading to marginalization of information and loss of local tokens features [18]. In addition, the calculated amount of ViT increases exponentially as the number of tokens increases, and the final prediction result of ViT is often determined by some tokens with large amounts of information, whereas most of tokens are redundant [19]. Thus, this study introduces maxpooling before the transformer encoder to perform token selection through selecting the informative token rather than input all redundant tokens for further processing.

This proposed model reduces the reliance on human resources to classify the product according to the types of defects, saves time, and improves precision in identifying defects that humans can possibly miss. Researchers can use this finding as foundational knowledge and further make an improvement to be applied across different domains. Furthermore, this study can stand as a benchmark for accuracy to compare with other models in the future. This study gives insight on how a DL model as a part of AI can support Industry Revolution 4.0 and smart manufacturing practices. DL implements for defect classification of AM products align with the principle of the United Nations Sustainable Development Goals (SDG), specifically SDG 9 (Industry, Innovation, and Infrastructure), whereby deep learning used for defect classification during quality inspection directly supports reliable infrastructure, as it ensures the product complies with the standards, thereby standing as a key aspect for sustainable industrial methods.

1.6 Scope and Limitations of Study

This study focuses on improving defect classification in the AM process by developing a deep learning-based model that integrates CNN and ViT architectures into a single hybrid framework. The dataset supplied by Malaysia Automotive, Robotics, and IoT Institute (MARii) used in this study consists of two defect classes, crack and stringing, and no defect of AM products. The total dataset contains 1500 images, which are divided into training, validation, and testing sets. The study focuses specifically on this dataset and did not test using another AM image dataset. The developed model was tested offline and not tested in real-time applications. The performance of the proposed model is evaluated using standard classification performance metrics derived from the confusion matrix, including accuracy, precision, sensitivity, and F1-score. In addition, model efficiency was evaluated using accuracy versus epoch plots. These performance indicators provide a comprehensive assessment of both classification accuracy and computational efficiency. This study compares the performance of the proposed CNN-enhanced ViT model with selected baseline models, specifically the CNN-ViT hybrid model and the Swin Transformer. These models are selected due to their relevance as state-of-the-art and hybrid deep learning architectures in image classification tasks. The implementation, training, validation, and testing of all models are conducted using Google Colab, which provides a Python-based environment for deep learning model development.

1.7 Thesis Outline

This thesis is organized into five main chapters, structured to address the critical aspect in developing CNN with enhanced ViT model for defect classification of additive manufacturing products.

Chapter 1: This chapter introduce the background and motivation of the study, focusing on the importance of quality inspection and the challenges associated with deep learning model for AM defect classification. This chapter also outlines the problem statement, research objectives, scope of study, and significance of the study.

Chapter 2: This chapter provides a detailed literature review, which cover AM traditional inspection methods, conventional machine learning approaches, and recent developments in deep learning for quality inspection. The chapter also reviews CNN, ViT and hybrid CNN-ViT architectures.

Chapter 3: This chapter describe the methodology including data collection, framework of the proposed CNN-enhanced ViT architecture in detail. In addition, it presents the training configuration, hyperparameter settings, and performance evaluation metrics used to assess the model.

Chapter 4: This chapter presents the performance evaluation of the proposed model. It analyzes training performance, classification accuracy, precision, recall, F1-score, from confusion matrix results. The chapter compares the proposed model with baseline models such as CNN and standard Vision Transformer.

Chapter 5: This chapter concludes the key findings and evaluates the achievement of the research objectives. It highlights the main contributions of the proposed CNN-enhanced ViT model for AM defect classification. The chapter also discusses the limitations of the study and provides recommendations for future research.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter presents a comprehensive review of relevant literature within the scope of this study. Each subtopic is critically examined to establish the fundamental knowledge, justify the methodological choices, and connect previous research with the present work. The review begins by discussing AM and its associated defect challenges, followed by an exploration of conventional defect classification approaches, particularly non-destructive testing (NDT). It then examines machine learning models, specifically supervised learning including Regression, Bayesian, SVM, DT and DL methods for defect classification, with emphasis on CNN for local feature extraction and ViT for global feature extraction and how previous study made enhancement involving patch division and token selection. The discussion extends to hybrid CNN-ViT architectures, highlighting hierarchical feature learning in CNN, the self-attention mechanism in ViT, and the rationale for integrating both approaches to capture local and global features simultaneously. Finally, the chapter identifies the research gap and motivates the present study by proposing a CNN-enhanced ViT framework aims to improve accuracy, robustness, and generalization in defect classification for AM products.

2.2 AM Fabrication Process

3D-printing, also known as additive manufacturing (AM), refers to the making of a three-dimensional printed object using an additive process. The manufactured product is being constructed layer by layer with the help of computer-aided design (CAD) [3]. This process involves joining the material to form objects from 3D model data [33]. Specifically, the AM process consist of eight phases to be considered as depicted in Figure 2.1. All AM process begin by creating a digital 3D model using Computer-Aided Design (CAD) software or by scanning an existing object. By using CAD software, any CAD solid modelling software can be used as long as the final product represented as 3D surface or solid representation [2]. Otherwise, the process

can be done by scanning an existing object using reverse engineering tools like laser and optical scanning [34].

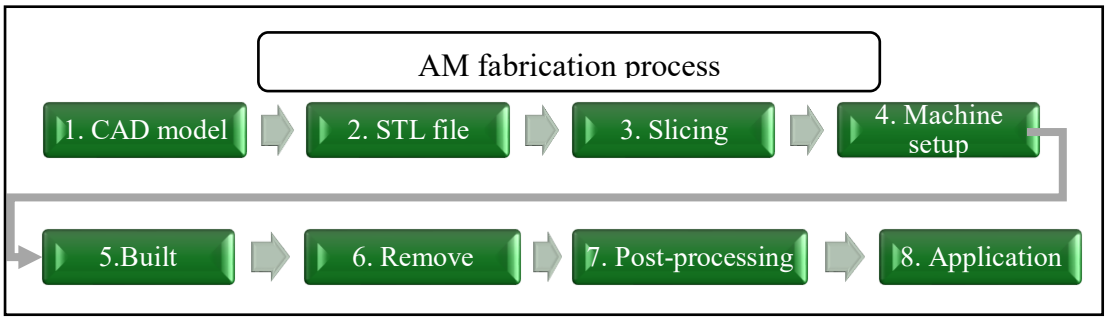


Figure 2.1 Flow of AM Fabrication Process

For the second phase, the CAD file is converted into a specialized file format, most commonly Standard Tessellation Language (STL) file, which defines the original CAD model's exterior closed surfaces and serves as the foundation for computing the slices. It is essential to take various parameters into account for effective printing of components when converting the CAD models into STL files. The parameter includes chord height (affecting surface smoothness), deviation (impacting accuracy), angle tolerance (for smoothness), poly count (influencing detail), mesh quality (overall mesh integrity), file size (transmission and processing speed), and slicing settings (printer-specific instructions) [2]. The parameters either configure manually by designer or optimized automatically using advanced AM system through considering materials and objectives of the print product. For the third phase involve slicing process, where the STL file is sliced into thin horizontal layers, and toolpaths are generated for each layer. In this phase, file editing can be performed to get the part the right size, location, and orientation for construction [35]. For the fourth phase, machine setup including construction parameters, such as material restrictions, energy supply, layer thickness, timings need to be configured before build the desired product. In the fifth phase to build the product, the printer deposits or solidifies material (like plastic, metal, or resin) layer by layer according to the sliced data. The printing process is automated, which require only supervision to ensure there is no problem with running out of material, power supply or software malfunction [36]. In the sixth phase, the support components need to be removed when the AM printer has finished printing process. The 3d-printer may have safety mechanisms to ensure that the temperatures in the chamber are appropriately low and no actively moving parts are present, which require user to wait for a while before removing the support. Next, in the seventh phase, post-processing

has to be performed. After printing, the part may require cleaning, curing, support removal, or finishing to achieve desired properties. A skilled physical manipulation is necessary to perform this post-processing. The priming and painting are required in order to have a satisfactory surface finish [37]. Then, the final part is checked for dimensional accuracy and mechanical performance during the inspection process before user can use the printed product.

2.2.1 Types of AM Processes

The printing process varies according to the type of AM processes. There are seven main categories of AM processes including material extrusion, vat polymerization, material jetting, binder jetting, sheet lamination, powder bed fusion and direct energy deposition [38]. Table 2.1 depict the example of AM process and example of the AM technology for each process accordingly.

Table 2.1
The AM Process, Principle of Operation and Example of AM Technology According for Each of the Process Accordingly.

AM process	Principle of operation	Example of AM technology
Material extrusion	Material is extruded layer by layer on the build plate through a nozzle	Fused Deposition Modelling (FDM), Fused Filament Fabrication (FFF)
Vat polymerization	A liquid photopolymer resin in a vat is cured by a light source (laser or projector) to form solid layers.	Stereolithography (SLA), Direct Light Processing (DLP)
Material jetting	Droplets of build material are selectively deposited and cured layer by layer.	Drop-On-Demand (DOD), Nanoparticle Jetting (NPJ)
Binder jetting	A liquid binding agent is selectively deposited to join powder particles layer by layer.	Binder jetting (BJT)
Sheet lamination	Sheets of material are bonded together and cut to shape.	Laminated Object Manufacturing (LOM)
Powder bed fusion	A thermal energy source (laser/electron beam) selectively fuses regions of a powder bed.	Selective Laser Sintering (SLS), Laser-based Powder Bed Fusion (LPBF)
Direct energy deposition	Focused thermal energy (laser, electron beam, plasma arc) melts material as it is being deposited.	Laser Engineered Net Shaping (LENS), Laser Based Metal Deposition (LBMD)

2.2.2 Type of AM defects

The significance of AM lies in its ability to create products with similar design and functionality at lower costs and shorter production times [39]. The AM process still susceptible to various defects due to variation in material properties and structure which lead to deterioration quality of the printed product [3]. The defects specifically occur in non-metallic AM processes, and the cause of defect is tabulated in Table 2.2, and the defects are illustrated accordingly in Figure 2.2.

Table 2.2
Non-Metallic AM Defects and Its Causes.

No.	Defect	Cause
1	Layer separation	Poor bed adhesion, high print speed, interference of cooling fan [40, 41]
2	Gaps between contours and infill and gaps into infill line	Low extrusion temperature, high print speed, low infill density [42]
3	Overheating effect	Inadequate cooling, elevated extrusion temperature, high speed/layer height [43]
4	Skipped layer	Loose belts, inconsistent filament diameter, motor issues, [44]
5	Whiskers	Poor retraction settings, insufficient cooling, incorrect temperature [44]
6	Gaps in thin walls	Low extrusion flow, poor adhesion, high print speed, inappropriate cooling [45]
7	Stringing	High temperature, insufficient cooling, poor retraction settings [3]
8	Warping	Poor adhesion, large surfaces, rapid temperature changes, high shrinkage materials [46]
9	Ringing	Sudden speed changes, printer vibrations, extrusion rate fluctuations [47]
10	Blobs and zits	Poor retraction settings, incorrect temperature, inadequate cooling [48]
11	Under extrusion	Clogged nozzle, incorrect filament diameter, filament tension [49]
12	Layer shifting	Loose belts/screws, external vibrations, motor issues, G-Code errors [11, 50]

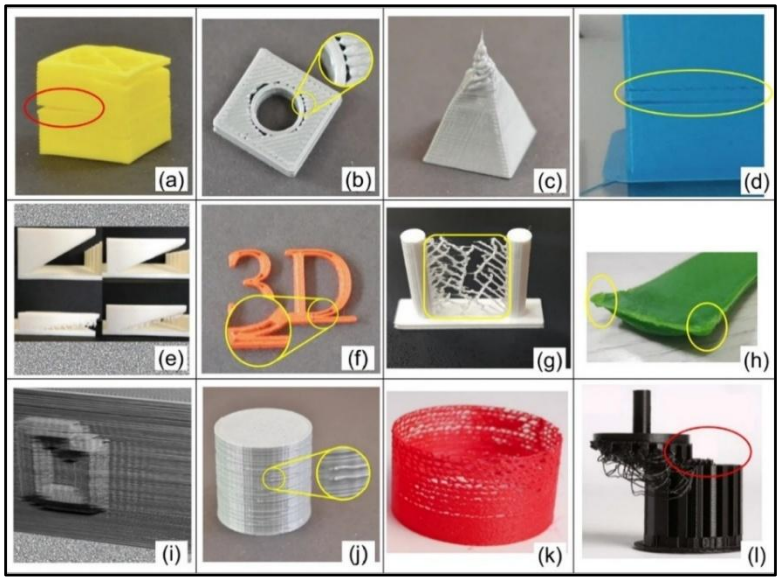


Figure 2.2 Types of AM Defects in Non-metallic AM Process, where (A) Layer Separation, (B) Gaps Between Contours and Infill and Gaps into Infill Line, (C) Overheating Effect, (D) Skipped Layer, (E) Whiskers, (F) Gaps in Thin Walls, (G) Stringing, (H) Warping, (I) Ringing, (J) Blobs and Zits, (K) Under Extrusion, and (L) Layer Shifting [4].

2.3 Conventional Defect Detection Methods in AM

The 3D-printing process are prone to various defects that lead the deterioration in the quality of the printed product [3]. Certain post-processing techniques are used to ensure the printed product has an acceptable surface finish, which require lots of time [4]. Conventional method including manual inspection, rule-based algorithms and non-destructive test has been adopted to inspect the AM products. NDT widely used in many industries include AM industries are ultrasonic testing, penetration testing, infrared imaging and eddy current testing [51].

Manual inspection is one of the oldest methods used to detect defects in AM products. Manual inspection is simple and does not require advanced equipment, as the inspector has to visually examined the printed parts to identify defect such cracks, pores, or surface irregularities. However, the defect inspection is highly subjective and greatly depends on expertise of inspector [52], where human fatigue and bias can lead to inconsistent results. Moreover, manual inspection is also tedious, ineffective use of manpower, time-consuming and not suitable for large-scale production [5], as well as possibly insufficient human resources to perform quality inspection [53].

Rule-based algorithms were introduced to automate defect classification. These algorithms use predefined rules to identify defects in images or sensor data. For

example, an edge detection algorithm may be used to identify cracks [54], while a thresholding algorithm may be used to detect porosity [55]. Rule-based methods are faster than manual inspection and reduce human error. However, they rely heavily on handcrafted features and predefined thresholds. This makes them less flexible and unable to handle complex defect patterns. Rule-based algorithms often fail when defects vary in shape, size, or intensity.

On the other hand, ultrasonic testing is a physical method that use the characteristics of ultrasonic waves in order to detect the defects on the surface, subsurface and inside the components in AM products [4].The transmitting probe propagates the ultrasonic wave over the printed product and reflect back the distinct signal for each scanned surface [4], including defects such as crack or void. The time lag between the transmitted and reflected signals is important to detect the size, location, and type of defects [56]. A study by Millon et al. has used a pulsed laser to generate ultrasonic waves projected to the surface of AM product and capable of identify crack defects [57]. Ultrasonic testing mainly uses to detect porosities, corrosion, cracks or inclusions has shown high accuracy as this test not affected by the complexity of the geometric shape, providing reliable results in thin or large components [58]. However, when defects are given in the planar direction parallel to the trajectory of ultrasonic waves can negatively affect the accuracy [59].

The penetration testing also being used to detect defects of AM product. The penetrant consist of fluorescent or vibrant coloured dyes is applied to the surface of AM products and the excess penetrant is removed, followed by applying developer on the surface, making the defects are visible under the light source [4]. The Marshall Space Flight Center (MSFC) of NASA has use this method to inspect the metal components printed using AM process [60]. While the penetration testing is considered effective in detecting surface defects, the application limited to visible surface defects only and subsurface defect cannot be detected [61].

Furthermore, infrared imaging with thermal cameras stands as a promising NDT method in traditional defect detection. The heat emitted from the printed part at a particular temperature [58], and the thermal response is recorded using the infrared cameras. The defect can be detected when the heat radiation between defective region and neighbouring regions are varies [4]. Schwerdtfeger et al. have used infrared imaging for each layer during the printing process and found that the radiation strength at the defect is higher [29]. This study shows that the defect can be detected as it affects the

thermal conductivity and results in temperature variation. Even though this method give advantages in detecting defects, the use of low quality IR camera provide lower accuracy, difficulty in interpreting the infrared images and variation in surrounding temperature, airflow or humidity can affect the accuracy of the result [62].

Finally, eddy current testing uses electromagnetic induction to detect defects through analyze the variations in the induced eddy currents. The alternating current generates a fluctuate magnetic field when passed through a primary coil [63]. When the primary coil near to the printed product, it induced eddy current and cause the variations in the flow of eddy current if defect exist [64]. A study by Gel'atko use eddy current testing to identify AM stainless steel parts and to examine the data obtained by eddy currents variation under the influence of various types of designed artificial defects[65]. This study shows eddy current test has significant potential to identify defects in AM product. This method is highly effective for real time monitoring but only applicable for conductive materials such as aluminium [58]. Table 2.3 summarize others conventional defect detection methods of AM products.

Table 2.3
Summary of Conventional Defect Detection Method of AM Products with its Limitation.

Author	Method	Description	Limitation
Dorafshan et al. [54] Rezaei et al. [55]	Rule-based algorithms	-Edge detection algorithm used to identify cracks. -Thresholding algorithm used to detect porosity.	Rely on handcrafted features and predefined thresholds
Millon et al. [57]	Ultrasonic testing	Identify crack defects	Equipment for ultrasonic testing is expensive
Waller et al. [60]	Penetration testing	Marshall Space Flight Center (MSFC) of NASA inspect the metal components printed using AM process	Limited to visible surface defects only
Schwerdtfeger et al. [29]	Infrared imaging	Detect defect for each layer during the printing process	Difficulty in interpreting the infrared images and variation in surrounding temperature, airflow or humidity can affect the accuracy
Gel'atko et al. [65]	Eddy current testing	Identify AM stainless steel parts	Applicable for conductive materials only

2.4 Machine Learning in Defect Classification

The machine learning (ML) approach are more suitable for defect detection for AM product than traditional procedures as it capable in adapting complex data structures and effective feature extraction [4]. Machine learning (ML), an artificial intelligence (AI) technology suitable to perform data analysis and decision making without the assistance of special programmed software [66]. Machine learning techniques divided into three categories, which are supervised, unsupervised and semi-supervised learning [67].

Supervised learning involved the use of labelled dataset as the input to train the model. The labelled dataset allow the model to learn the relationship between the input and the output [68]. A generalized relationship learned between the input and output allow accurate predictions of new unseen data [69]. A study by Park et al. have developed a supervised deep learning (DL) based model to assess the porosity of AM products [70]. The developed model shows the accuracy of trained model to assess the porosity closely match with the porosity measured using conventional scanning acoustic microscopy, proving the deep learning model is reliable for detecting defects. In addition, the use of supervised learning model achieves high accuracy for classification task since the model learned from the labelled data which provide reliable predictions. Supervised learning model such as CNN and ViT architectures have been applied in AM defect classification. CNN has been successfully used to detect and classify defects such as cracks and surface irregularities due to its ability to extract local image features automatically [71-73]. Recent studies have also applied Vision Transformer to classify defects by capturing global contextual relationships between image regions, improving classification accuracy and defect characterization [16, 74, 75]. These supervised learning approaches play a critical role in enabling automated quality inspection and intelligent defect classification systems in additive manufacturing.

Unsupervised learning uses unlabelled data to train the model. The model requires to learn the structure, patterns, and relationships of the input data without prior knowledge of the output labels [76]. Unsupervised learning is used to find hidden structure and cannot be used to predict a specific desired output. It is known as adaptive learning algorithm since no basic information provided to the network [77]. A study by Taherkhani et al. has implement Self-Organizing Map (SOM), an unsupervised machine learning algorithm to detect porosity induced during laser powder bed fusion (LPBF)

printing [78]. SOM capable of detecting defects without the need of user-defined threshold. It automatically clusters the photodiode intensity signal based on similarity, preserving the topology of the data. Another study by Song et al. proposes a two-stage unsupervised defect detection framework by using a CNN-based autoencoder [76]. A revised threshold method is used to detect anomalies during the wire and arc additive manufacturing (WAAM) process and achieves an F1-score of 86.3%. Based on these studies, unsupervised learning, shows capability in detecting defects, but it cannot predict a specific defect at the printed product.

Semi-supervised learning is the combination of supervised and unsupervised learning, where some data are labelled but massive unlabelled data are used to train the model [66]. This semi-supervised model helpful if dataset is scarce and need an ample sensor to process unlabelled data. However, this model is expensive and difficult to train. Manivannan has develop a semi-supervised deep learning model using CNN with pseudo-labelling to detect defects of powder bed defects in selective laser sintering (SLS) printing, achieve an accuracy of 98% [79]. The proposed method achieves satisfactory results with relative to the small amount of labelled data. However, incorrect prediction can be occur when the pseudo-labels are noisy [80], which may degrade the model performance compared to the use of only labelled data. Another study by Mattera et al. presents an approach based on semi-supervised learning for real-time anomaly detection in the production of Inconel 718 components via the Pulsed Transfer WAAM process [81]. The collected data employed to train an unsupervised learning approach based on a residual convolutional autoencoder by using solely good deposition data to learn hidden complex patterns of normality and hence detect anomalies based on deviations from these patterns. The proposed achieved an F-score of 0.895, representing a 30 % improvement over state-of-the-art methods that rely on time domain features for anomaly detection. However, the use of unsupervised learning only detects patterns or anomalies but cannot determine the exact types of defects. Table 2.4 summarize the strength and limitations for supervised, unsupervised and semi-supervised learning.

Table 2.4
Summary of Strength and Limitation of Different Machine Learning Categorries.

Category	Example	Strength	Limitation
Supervised	CNN, ViT	Provide reliable predictions, suitable for classification task.	Massive data labelling.
Unsupervised	SOM, CNN-based autoencoder	Learn the structure, patterns, and relationships of the input data without prior knowledge of the output labels.	Not capable of predicting a specific defect at the printed product.
Semi-supervised	CNN with pseudo-labelling, Residual convolutional autoencoder	Helpful if dataset is scarce.	Incorrect prediction can occur.

In this study, it focuses and discuss on supervised learning models which consist of five major algorithms that can be used to perform defect classification including regression, bayesian, support vector machine (SVM), decision tree (DT) and deep learning (DL).

2.4.1 Regression

Regression model used to find the mathematical relationship between one or more input variables and an output variable (dependent variable) so that the model can predict or explain the output based on the inputs. The architecture for a generalised regression neural network consists of input layer, pattern layer, summation layer and output layer as depicted in Figure 2.3. The input layer consists of number of input parameters that are equal to the number of neurons, the pattern layer consists of radial centres applied for all training points, as well as smoothing being applied to the radial basis function in order to minimize the prediction error, at the summation layer, it consists of two neurons, numerator adds up the weight values multiplied by the actual target value for each pattern neuron and denominator adds up the weight values coming from each of the hidden neurons, and at the output layer, the output neuron divides the numerator by the denominator to produce the predicted value [82].

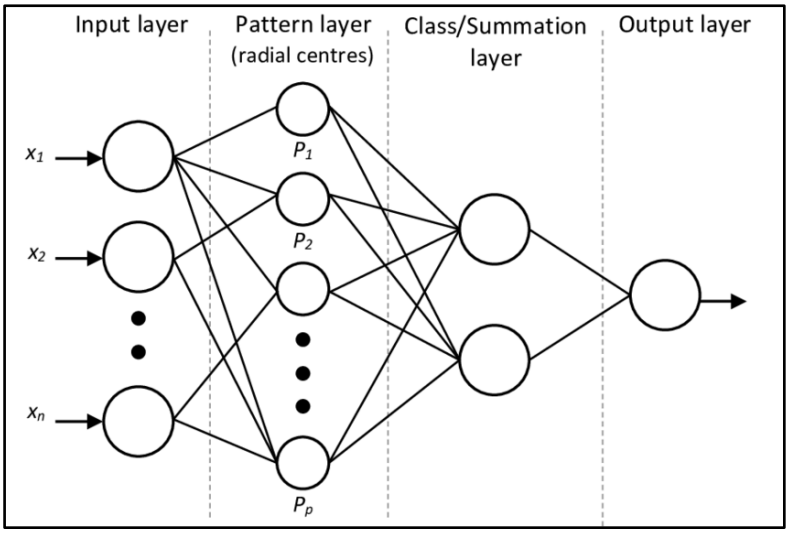


Figure 2.3 Generalised Regression Architecture [82].

The regression models aim to learn a functional relationship between a set of input variables (features) and a continuous output (target), by estimating the model parameters that minimize prediction errors. A study by Baturynska and Martinsen applies linear regression as an initial statistical method to detect and quantify geometric defects (dimensional deviations) in additively manufactured polymer parts produced by powder bed fusion [83]. However, the model assumes a linear relationship between input variables (part placement, orientation, STL properties) and dimensional deviation. As a result, many defect mechanisms cannot be accurately captured by a purely linear model. Another study by Tapia et al. has develop a Gaussian process-based model to learn and predict the porosity in metallic parts produced using selective laser melting (SLM - a laser-based AM process) [84]. The proposed model demonstrates that defects in AM products can be detected using probability method before manufacturing, by learning the relationship between process parameters and porosity. This shifts defect detection from post-process inspection to predictive quality control, enabling better parameter selection and reduced defect rates. However, the reliability of the proposed model is influenced by kernel and hyperparameter selection, as the model may underfit if it does not match with the true process behaviour. Another study by Garcia et al. shows that regression models can successfully predict dimensional quality including inner diameter and outer diameter in a tubing extrusion process using real production data [85]. It highlights the effectiveness of k-NN and support vector regression as practical tools for quality prediction and defect prevention in continuous manufacturing processes. However, k-NN and SVR model require carefully engineered input features,

which cannot directly process raw data such as images, acoustic signals, melt-pool videos or thermal maps. In summary, the regression models aim to learn a functional relationship between a set of input variables (features) and a continuous output (target), by estimating the model parameters that minimize prediction error. However, the regression models depend heavily on manual feature engineering, unable to discover hidden or abstract representations and require expert knowledge to select meaningful inputs.

2.4.2 Bayesian

Bayesian is another machine learning model which works based on Bayes' theorem which provides a probabilistic framework for learning from data, updating beliefs, and making predictions under uncertainty. Bayesian machine learning treats model parameters and predictions as random variables, rather than fixed values. Learning is performed by updating prior beliefs about these variables using observed data. Bayesian models require predefined probability distributions and Bayesian models struggle to capture highly nonlinear, hierarchical, or abstract patterns due to only few layers are used. In addition, the Bayesian model relies on manually selected variables, engineered features and domain-driven representations, in which the performance is constrained by the quality of feature design. Figure 2.4 illustrates the difference between neural networks (left) and Bayesian (right).

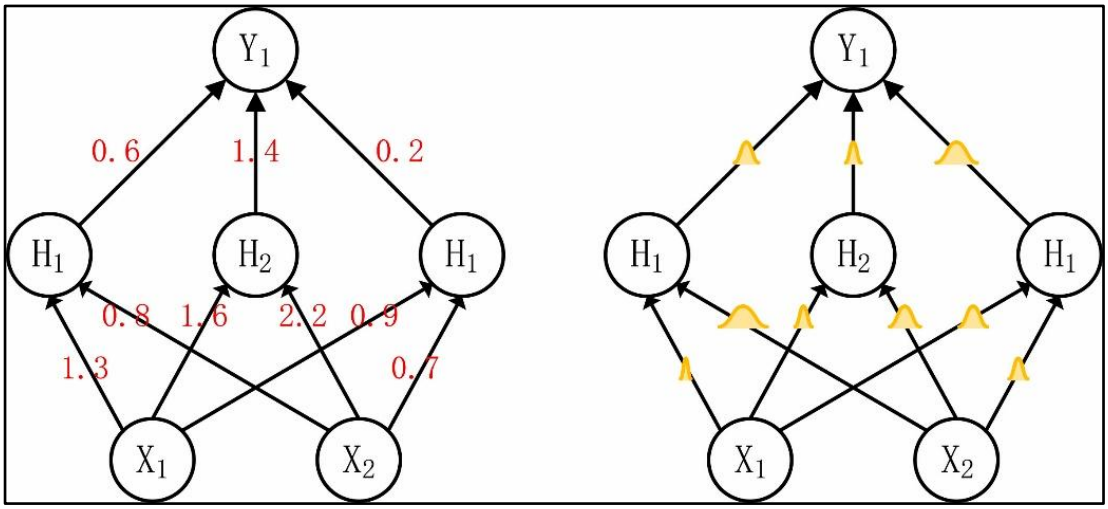


Figure 2.4 Determination Neural Networks, which Have Fixed Values for Each Parameter (Left) and Probabilistic Bayesian Assign a Distribution to Every Parameter (Right) [86].

A study by Ferreira et al. performed defect detection of AM product by modelling expected geometric deviations with a Bayesian neural network and identifying abnormal deviations that exceed the model's predicted mean and uncertainty, rather than using explicit defect labels [87]. While the proposed model can predict geometric deviation as well, the Bayesian relies on engineered input features, whereas deep learning can reduce this dependency. Another study by Aminzadeh and Kurfess develops an online monitoring system for quality of fusion and defect formation in every layer of the laser powder-bed fusion process using computer vision and Bayesian inference. Features are carefully selected based on physical intuition into the process and extracted from the images of the various types of builds [88]. However, the proposed model lacks automatic feature learning since it relies on manually designed features extracted from images, which means if defect appearance changes (due to different materials, machines, or lighting), the chosen features may lose discriminative power to perform classification accurately. Another study by Hertlein et al. predicting the quality of parts produced by selective laser melting (SLM) using a hybrid Bayesian network model [89]. The proposed model uses a hybrid Bayesian network with predefined structure, as relationships between variables must be manually specified (parent–child links). Thus, the dependency on predefined structure limits the performance due to process–quality relationships are extremely nonlinear or involve high-order interactions. From these models which use Bayesian to perform defect classification of AM products, it is clear that Bayesian models require predefined probability distributions and Bayesian models struggle to capture highly nonlinear, hierarchical, or abstract patterns due to only few layers are used. In addition, the Bayesian model relies on manually selected variables, engineered features and domain-driven representations, in which the performance is constrained by the quality of feature design.

2.4.3 Support Vector Machine

The basic principle of support vector machine (SVM) in machine learning is to find an optimal decision boundary that best separates data points of different classes. SVM chooses the hyperplane that maximizes the margin between classes, illustrated as in Figure 2.5. A larger margin generally implies a more robust classifier. When data is

not linearly separable, SVM uses kernel functions to map data into a higher-dimensional space where a linear separation becomes possible.

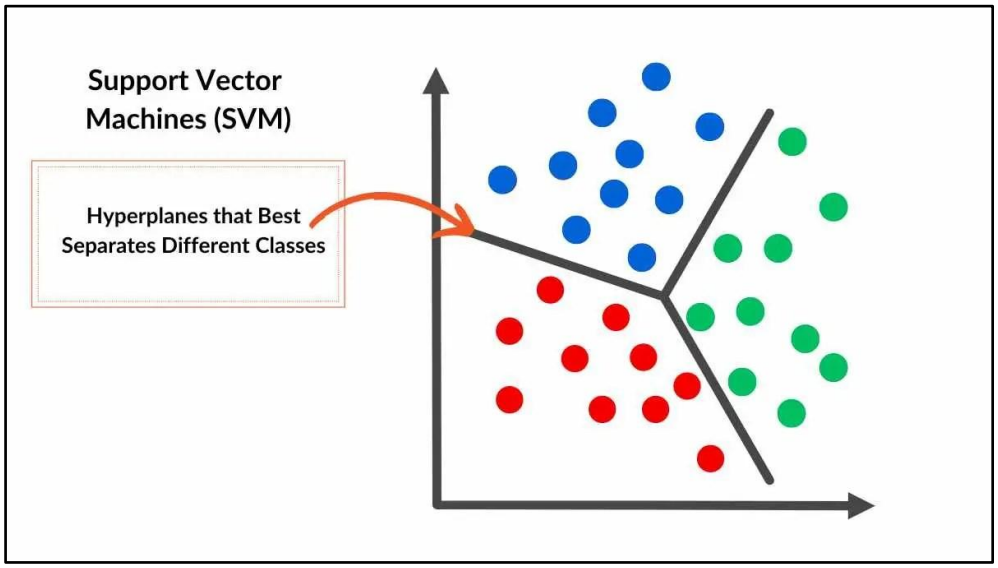


Figure 2.5 The Illustration of Hyperplanes that Best Separates Different Classes in Support Vector Machine.

SVM can be effectively used for defect classification in AM products by learning patterns that distinguish different defect types based on measured data. A study by Gui et al. investigates how internal defects in metal parts produced by electron beam powder bed fusion (PBF-EB) can be detected, classified, and avoided using surface morphology data and machine learning [90]. Five machine learning models including linear regression (LR), SVM, DT, XGBoost, naive bayes are compared. The results show that the SVM achieves the best classification performance of 95% accuracy. Thus, the SVM is used to construct a process map that predicts parameter ranges of non-defect parts. The RBF kernel used allowed SVM to model nonlinear decision boundaries between no defect and defective process conditions. However, SVM cannot automatically learn hierarchical or abstract features from raw data including image and signals data. Another study by Liu et al. uses SVM as the core classifier for real-time defect detection in laser additive manufacturing (LAM) [91]. SVM is applied after image-based feature extraction and dimensionality reduction. The results demonstrate that SVM can reliably distinguish normal and defective states during the LAM process. However, the study relies on handcrafted features (molten pool geometry, grayscale intensity, temperature-related features). In addition, performance is strongly limited by the completeness of selected features, human assumptions about which features represent defects. Thus, SVM cannot learn new or hidden features beyond those

explicitly defined. Another study by Wang et al. proposed an in-situ quality monitoring for selective laser melting (SLM) printing based on machine learning and improved variational modal decomposition (VMD) [92]. The artificial neural network (ANN) and support vector machine (SVM) were employed to perform quality prediction. The prediction achieves an accuracy of from 75.00% to 83.33% using ANN and SVM achieves 83.30% to 91.70%, clearly show that the SVM has better prediction accuracy and generalization performance in small sample datasets compared to ANN.

2.4.4 Decision Tree

Decision tree (DT) is a supervised machine learning algorithm used for classification and regression tasks. DT works based on splitting data into smaller subsets according to certain decision rules derived from the input features. DT divide a data set according to the certain conditions, which involves the collection of if-else statements, as true conditions will proceed to the next node connected to the results [93]. The goal is to create a model that predicts the target variable by learning simple decision rules from data features. The sequence of feature-based splits like 'inverted tree' flowchart, as the extraction begins with root node and finishes at the leaves (terminal mode). The tree starts from a root node, which represents the entire dataset. It chooses the best feature to split the data. Each split creates branches that divide the data into groups. This process continues recursively, forming decision node and terminal nodes (final predictions). In the classification case, each terminal node represents a class label, while in regression, it represents a numerical value. The algorithm keeps splitting until a stopping condition is reached. During prediction, a new data point travels down the tree, following the decision rules at each node until it reaches a leaf (terminal node), which provides the output. Figure 2.6 illustrates the architecture of decision tree.

A study by Trovato and Cicconi develop a DT classifier to evaluate the printability of metal parts for AM, specifically laser powder bed fusion (L-PBF), at an early design stage by identifying non-printable CAD geometries before manufacturing [94]. The trained model achieves an accuracy of approximately 91.00% in predicting printability. The study concludes that DT can effectively support early design for additive manufacturing (DfAM) decisions and that the method is suitable for rapid geometry validation before detailed process planning. The proposed model depends on

a knowledge-based tool to define which features are extracted from the CAD model. The quality and completeness of the prediction therefore depend heavily on expert-defined rules and selected parameters. Deep learning models reduce this dependency by performing automatic feature learning, which can uncover influential geometric patterns that are not explicitly defined by designers. Another study by Aziz et al. has performed automatic classification of typical defects found during metallographic examination of additively manufactured nickel alloys and the performance of two supervised machine learning methods, kth-nearest neighbours (kNN) and decision tree (DT) is studied [95]. The kNN and DT models rely on manually measured geometric descriptors. However, these handcrafted features capture only limited information from micrographs. Another study by Cheepu introduce an acquiring data system from the welding arc to identify defect in wire arc additive manufacturing (WAAM) [93]. Machine learning including DT, random forest (RF) and linear regression (LR) are used in this study. The analysis of the defects of the first layer and successive layers data has been used to train the models and predict the WAAM process defects of further layers. The DT model achieved 99.13% accuracy. While DT performed well, it showed slightly lower accuracy compared to random forest of 99.78 % accuracy. This study mainly focusses on voltage signals of arc voltage fluctuations, but other signal like current is not integrated in the ML models, which might lead missing defect information.

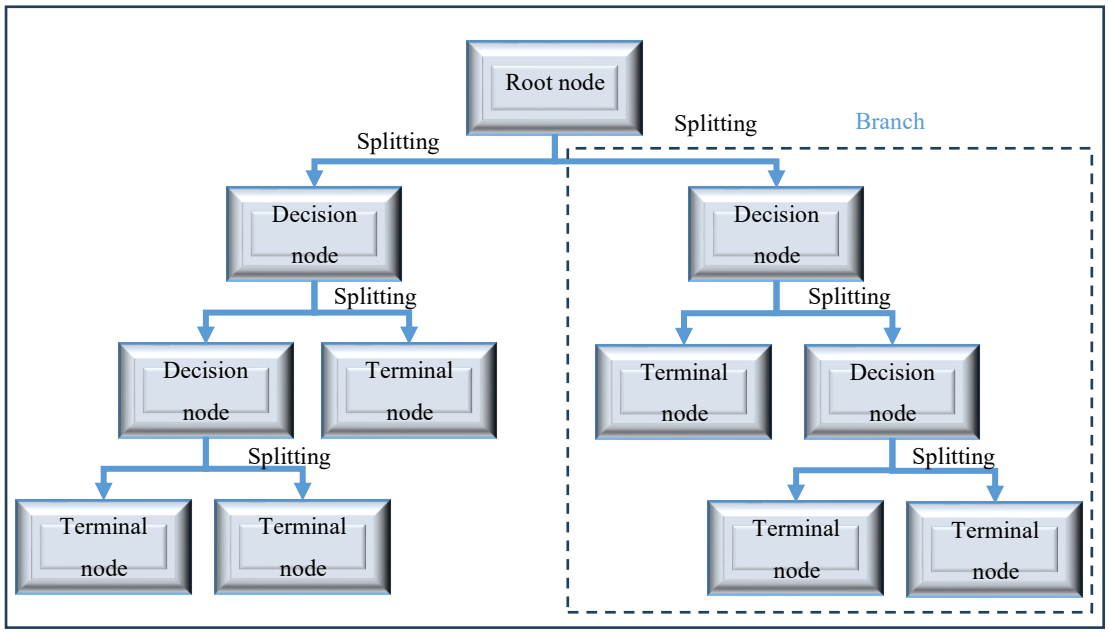


Figure 2.6 The Architecture of Decision Tree Consists of Root Node, Decision Node and Terminal Node.

While regression, bayesian, SVM and DT have shown good performance, these architectures are unable to capture defects accurately, as it depends on manual parameter tuning and cannot adapt to non-linear data. Deep learning (DL), a subset of machine learning has proven to be exceptionally successful in classification since it can adapt to the lighting, colour, shape, and size of the input images, as it is beneficial for detecting complex surface defects [30].

2.5 Deep Learning for Defect Classification

The focus of this study is to perform defect classification of additive manufacturing products using deep learning model. Supervised learning specifically deep learning model has been used to develop the AM defect classification system [70]. Deep learning model is applied in this study which aims to classify three types of AM class involving crack, stringing and no defect class only. The reason of choosing deep learning model is deep learning can analyze features of the input data. Note that input data used in this study is the AM images, while features refer to any measurable characteristic such as shape, edge, line, or segment from the image. Deep learning has proven to be exceptionally successful in classification since it can adapt to the lighting, colour, shape, and size of the input images, as it is beneficial for detecting complex surface defects [30]. The implementation of deep learning plays a significant factor while performing classification, as it may contribute to robust vision-based solutions while giving higher accuracy [96]. It also provides automated visual inspection (AVI) during quality control phases, since manual inspection cannot fulfil the increasing quality standard in the manufacturing process [28].

2.5.1 Convolutional Neural Network (CNN)

CNN has shown promising performance in computer vision tasks like image classification, segmentation and object detection [97]. CNN consists of convolutional layers, pooling layers and fully connected layers as depicted in Figure 2.7. At the convolutional layers, the learnable filters are used to extract meaningful features from the input image through performing a convolution operation and produce a feature map. At pooling layers, the average pooling or maxpooling method is applied to reduce the spatial dimension of the feature map, thereby keeping only important information to be

processed at the following layers. The fully connected layers are used to aggregate the extracted features and thus classify the input according to the class involved.

CNN stands as an effective model in computer vision systems, as CNN uses local receptive fields, weight sharing and spatial hierarchical principles to extract low-to-high level features from the image dataset [98]. In terms of local receptive field, the learnable filters in the convolutional layer slide over the input image intentionally to extract the features. This process allow the model to extract features between local pixels (known as local receptive field) [25]. The extracted local receptive field determines meaningful features from the images as it captures edges, textures, and more complex patterns.

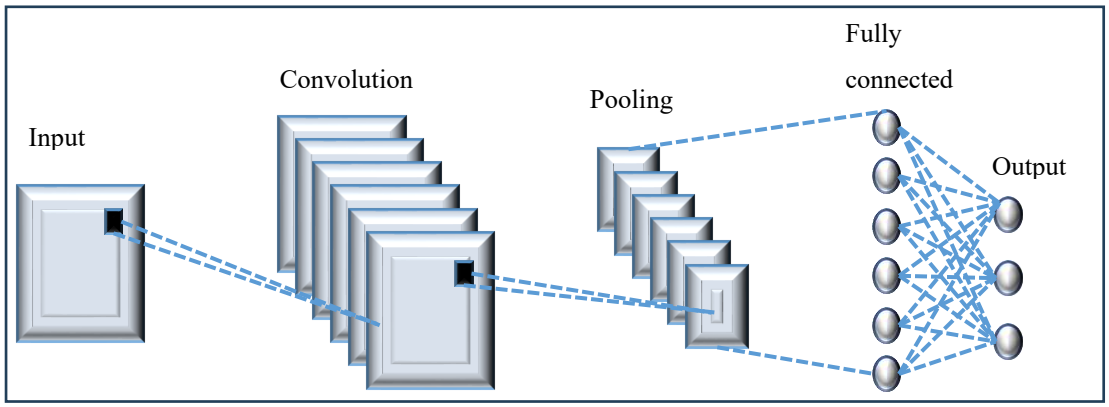


Figure 2.7 The Basic Architecture of CNN [99].

In terms of weight sharing principle, the same filter weights are applied across all different local regions of the image. This means the same filter weights are "shared" across the local regions. At different local regions, different elements of the feature map are obtained even with the use of the same filter weights. This principle shows that CNN is capable of adjusting the weight between elements in the feature map [100]. Note that if the value of elements in the feature map is the same, it means the model extract same features whereby the model can detect same meaningful features as edges from cracks at any location of the image (known as translation invariant). The weight sharing makes CNN an efficient model with fewer parameters used. This principle allow the network to extract visual characteristics through convolutional and pooling layers without relying on experience and professional knowledge [101].

In terms of spatial hierarchical principles, CNN is structured to automatically learn spatial hierarchical features from the input image using convolutional layers. The initial convolutional layer is used to extract low-level features like edges and corners.

Specifically, it helps to detect the sharp boundary of a crack or the fine filament of stringing. At the following convolutional layers, these layers combine the extracted low-level features into more complex patterns such as shapes, contours, or defect structures. This spatial hierarchical principle allow the learnable filter in the convolutional layer to be used to extract meaningful features from the images by capturing edges, textures, and more complex patterns in a hierarchical manner [102].

Deep learning particularly the CNN [103] is capable of processing complex visual data within AM scope, making it as valuable architecture for real-time defect detection and quality inspection [66]. CNN is preferrable compared to other deep learning models as it is known to be efficient such lower number of parameters [104, 105]. In defect classification, CNN can detect fine details of defects. A study by Westphal and Seitz proposed a CNN-based model by using VGG16 and the Xception CNN model with pretrained weights from the ImageNet dataset as initialization and an adapted classifier to classify good and defective image to perform automatic classification of powder bed defects in the selective laser sintering (SLS) process using very small datasets [71]. The VGG16 model architecture achieved the best results for accuracy of 95.80%. Yang et al. utilized a CNN classifier to classify four sizes of the melt pool from the L-PBF process yield a predictive performance of 91% with the melt pool images of two L-PBF-fabricated parts as the input [72]. The proposed model has been employed in-house AM software, outperform the traditional method by reducing the time required for melt pool image analysis by 90%. Also, another study by Zhang et al. uses CNN-based which is designed for predicting the porosity attribute, and demonstrates the classification accuracy of 91.20% for the direct laser deposition process [73]. This performance shows the DL-based model capable to detect defects of micropores up to 100 μm. The results from previous studies have demonstrated the effectiveness of CNN to perform classification tasks despite offering an alternative approach for non-destructive quality assurance of AM products.

The CNN architecture has shown strong performance in extracting features and capturing low-level details such as edges in initial layers and high-level patterns in later layers through the learnable filters in convolutional layers. However, the size of filter limits the extraction of a large receptive field thereby making the CNN unable to represent the large distance relationship. Even though the CNN has emphasized its capabilities in extracting local features, this architecture may struggle to capture long-distance contextual dependencies (global features), which can reduce the effectiveness

for task that require global information [106]. Global feature is important for detecting defects that span across regions. For example, a crack that spans across different regions may not fully recognized by CNN. These limitations motivate the use of other models such as vision transformers.

2.5.2 Vision Transformer (ViT)

Vision transformer (ViT) proposed by Dosovitskiy et al. [15], introduced a feature extraction method which entirely different from CNN-based model [107]. Vision transformer previously used for natural language processing (NLP) has been adopted for computer vision tasks including image classification task. The architecture of ViT is depicted in Figure 2.8. ViT process the input images as a sequence of image patches and utilize self-attention mechanism to extract the global contextual features [108]. The self-attention happen in transformer encoder learn the long-range dependencies between the image patches [109]. When the self-attention mechanism learn the long-range dependencies, every patch attends to every other patch, so that it can model the relationship between all pixels [110]. The self-attention mechanism weighs the importance of different patches when computing representations for each patch [111]. This mechanism allow ViT to model the relationship between any patches directly no matter the spatial distance in the original image [112].

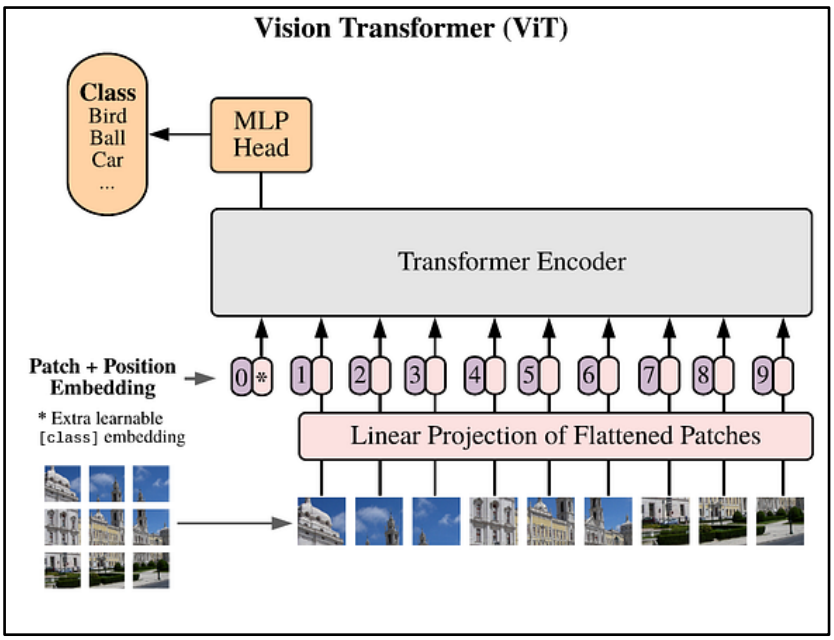


Figure 2.8 The Architecture of ViT [15].

ViT consist of three main modules which is patch embedding, transformer encoder and multilayer perceptron (MLP) classifier [25]. At patch embedding module, the input image is divided into 16x16 pixels of non-overlapping patches. The divided non-overlapping patches is flattened and linearly projected into a fixed-dimensional embedding to form a token sequence. Then, at the positional encoding layer, the learnable positional encodings are then added to these token sequences to retain spatial information. Additionally, a learnable classification token (known as the [CLS] token) is prepended to the token sequences. Token sequences with positional encoding and [CLS] token further input into repeated standard transformer encoder layers. The self-attention mechanism in the transformer encoder is used to process the token sequence by generating adaptive attention scores, enabling the comprehension of relationships between different regions of an image [13, 113]. Finally, the multi-layer perceptron (MLP) head is used for classification.

The feature extraction performed using ViT is contrary to the CNN-based model. ViT use self-attention mechanism to extract global features, while CNN-based model use filters in the convolutional layers to extract local features [14]. The use of self-attention mechanism can effectively solve the problem of limited local receptive field in convolutional layers in CNN model [114]. The CNN struggle to model long-range dependencies within an image due to convolutional layer focus on extracting features within neighbouring pixel [111]. The capability of ViT in extracting global features is beneficial in AM defect classification task, where ViT can relate distant regions of defects which span across regions.

Recently, vision transformer widely being used for image classification task. Maurício et al. found that the transformer-based architecture manage to perform better than the CNN-based model regarding the use of self-attention mechanism which allow the all features are accessible from highest to lowest layers [115]. Wang et al. proposed the ODP-Transformer[116], and Zhan et al. proposed IterationViT [117], where both models have utilized transformers to capture global features among images. Both have proven the use of ViT helps the model better understand the interactions between distant parts of an image. A study by Wang et al. has present a transformer-based approach to detect defects of fused filament fabrication (FFF) product [16]. The proposed model effectively classifies the printed product to different level of extrusion defect and slightly outperforms ResNet model under varying dataset sizes and noise levels. Another study by Singh et al. has use data-efficient image transformer (DeiT) combined

with weighted classification accuracy (WCA) to detect three types of defects which are warping, layer delamination and gaps in raster line achieve an accuracy of 99.30% [74]. The proposed model capable of detecting defects with high accuracy and capable to stop printing when defect occur. However, the performance on complex geometries is not verified. The proposed model may struggle to extract local features as complex geometries often have fine local irregularities. Another study by Uysalel et al. has implement ViT to recognize and classify the main types of AM defects based on their location, size, morphology, and type defects of Direct Metal Laser Sintering (DMLS) printing process [75]. The study uses ViT which the model has to split the microscopy image into patches for further extraction. However, this process may lose fine details of small defect of the printed product.

The self-attention mechanism in transformer solve the problem of limited local receptive fields in convolutional layers [114]. The self-attention mechanism generates adaptive attention scores to model non-local relations of the images [113]. Bhojanapalli et al. have studied different aspects to determine the robustness in ViT models, where ViT generally outperforms Resnets if trained with sufficient data and has fair robustness when removing the individual layer of ViT [118]. ViT has shown promising results in various classifications task including classification of brain tumors [119], traffic sign classification [120], and sign language recognition [121]. From these studies, ViT has provide a reliable option to perform defect classification using image as the input [74].

ViT has shown impressive result for computer vision task including image classification task within the use of a particular dataset with complex visual patterns or within the use on large-scale dataset [102]. ViT heavily rely on large-scale of dataset to achieve a comparable performance to CNN due to ViT lack of inductive bias [25]. When ViT model are pretrained on extensive dataset such as ImageNet, it capable of demonstrating impressive transfer learning capabilities tailored to specific domains [14]. ViT poses significant advantages regarding the nature of its architecture. ViT use self-attention mechanism, which allow every patch attend to another patches. This mechanism makes ViT capable of capturing global relationship across the entire image, and not just local textures. In defect classification, the cracks or stringing defect may span across distant regions of an AM product, where ViT capable to capture the defects globally. The transformer encoder in ViT can stabilize the gradient flow as it maintains relative positional information through encoding and decoding while training the model. It is mattered to maintain its stability in order to make them reliable for complex

task. However, ViT do have limitations, where ViT suffer to capture low-resolution and multi-scale features due to the use of fixed patch size [31]. In addition, the training of ViT model can be computationally expensive, rendering them less suitable for real-time deployment [110].

2.5.2.1 Enhancement Involves Token Selection

Another enhancement made in ViT involve the token selection mechanism including token selection based on convolution and pooling [19]. This method reduces the length of the token sequences using convolution and pooling operation aims to reduce redundant information and the computational cost. Pan et al. proposed hierarchical vision transformer (HVT) model, use the hierarchical pooling strategy to reduce the length of the token sequence [122]. The ViT blocks in this model first divided into several stages and maxpooling layer is added after the transformer block at each stage. This method capable of shrinking the sequence length when the network goes deeper. The proposed model may be beneficial on small datasets. However, this method causes underfitting or information loss on large-scale datasets. Another study by Chen et al. has proposed token pooling, selecting meaningful tokens as network goes deeper [123]. This study used convolution to adjust the token dimensions and pooling to reduce the number of tokens with enrich features of the token sequence to be processed.

2.5.2.2 Enhancement Involves Attention Mechanisms

Few studies have made an enhancement involving attention mechanisms operated in transformer encoder [124, 125]. The self-attention mechanism in original ViT model is characterized by feature collapse in deeper layers, resulting in the vanishing of low-level visual features, whereas the features can be helpful to accurately represent and identify elements within an image and increase the accuracy and robustness of vision-based recognition systems. A study by Diko et al. propose a novel residual attention learning method for improving ViT-based architectures, increasing their visual feature diversity and model robustness [124]. The proposed residual attention can capture and preserve significant low-level features, providing more details about the elements within the analyzed scene. Another study by Liu and Yan has propose a stripe depth-wise convolutional attention (SDCA) module [125]. This module

aggregates local convolution features at multiple scales as its attention map to effectively compensates for the limitations of self-attention mechanisms in handling object details.

2.5.3 Models Capable in Extracting Local and Global Features

A study by Dai et al. proposed CoAtNet which is capable of extracting both local and global features efficiently [20]. CoAtNet is a hybrid architecture that integrates convolutional operations with Transformer-based attention mechanisms for image classification. The model utilizes MBConv layers in the early stages to extract local spatial features, including edges and texture patterns. In the deeper stages, relative attention is applied to capture global contextual relationships and long-range dependencies. This combination enables CoAtNet to achieve strong generalization and high model capacity. However, the architecture still inherits the quadratic computational complexity of relative attention, leading to higher memory usage and computational cost.

Another study by Liu et al. develops swin transformer that uses window-based self-attention to extract local features and a shifted window being used for interaction between the extracted features [21]. Swin transformers construct hierarchical representations by starting from small-sized patches and gradually merging the patches in deeper layers. The proposed model is efficient for segmentation and detection tasks because it produces multi-scale feature maps similar to CNN algorithms. However, the reliance on window-based self-attention causes global information to propagate gradually across layers, which may limit the performance in image classification. Although the shifted window mechanism partially alleviates this issue by enabling interaction using cross-window, it does not provide global attention as in the ViT architecture.

2.5.4 The Hybrid CNN-ViT

As mentioned previously, ViT has limitation in capturing low-resolution and computationally expensive while training the ViT model. Hence, the aim of this research is to combine both CNN and ViT model into hybrid model that can complement each other as CNN excel in capturing local features while ViT excel in

capturing global features. A study has stated the hybrid model has potential in providing a more balanced approach for recognition tasks [126].

2.5.4.1 Enhancement Involves Patch Division from Feature Map

Previously, several studies have addressed the challenges associated with the original Vision Transformer, which divides input images into fixed-size non-overlapping patches. This process disrupts the local continuity of the image and limits the model's ability to effectively capture fine-grained visual features, thereby reducing its performance in visual tasks [19]. As a result, important local information may be marginalized, leading to the loss of meaningful local token features [18]. In addition, this process result the loss of fine-grained spatial information, which reduce the ability of the model in handling local details efficiently [110], thereby decrease the prediction capabilities [31]. In order to resolve the problems, there are few research have made enhancement involving patch division from original image to patch division from the feature map. The feature maps generated using a particular model turn as the input and divided into patches rather than traditional ViT use raw input image. This method makes each patch contain more meaningful information for further processing. The feature map either divided into non-overlap patches or overlap patches.

A study by Rajatha and Ashoka proposed a framework which leverage the strengths of EfficientNet-B4 and Vision Transformers for early detection and automated diagnosis of retinal disease [127]. The EfficientNetB4 reimagined as the multiscale feature encoder, creating discriminative feature maps preserving both local and intermediate contextual information and ViT incorporated to capitalize on the attention mechanisms to capture and model the global dependencies effectively. The proposed model achieved a significant advancement by scoring F1-score of 0.75 and the framework achieved a remarkable 5.17% increase in the overall score when compared to the previous cutting-edge technologies on the same task. There is another study by Ahn et al. proposed a hybrid deep learning framework for land use and land cover (LULC) classification using hyperspectral data, combining CNN and ViT which integrates a series of 1D convolutional layers for spatial-spectral feature extraction with a Transformer encoder for capturing long-range dependencies [128]. The developed model achieved an overall classification accuracy of 0.98, outperforming the SVM model's 0.72 and CNN model's 0.86. Both of these studies make each image first

passes through a CNN module, where local features are extracted. The output is then flattened and linearly transformed into a fixed-size vector (patch embedding) and positional encodings are added to retain the spatial order. However, both of the studies have no global information as reference as the positional encoding alone is not consistent across scale, whereby the global information is easily lost in the meantime of training or validating the model.

2.5.4.2 Enhancement Involves Feature Fusion Strategies Between CNN and ViT

A few studies had made enhancement by involving feature fusion strategies between CNN and ViT. A study by Rahman and Rahman proposes an Adaptive Attention-Augmented Convolutional Neural Network with Vision Transformer (AACNN-ViT), a hybrid framework that integrates local convolutional representations with global transformer embeddings through an adaptive attention-based fusion module for multiclass lung cancer classification from CT images, achieve 96.97% accuracy [129]. The CNN branch captures fine-grained spatial patterns, the ViT branch encodes long-range contextual dependencies, and the adaptive fusion mechanism learns to weight cross-representation interactions to improve discriminability. To reduce the impact of imbalance, a hybrid objective that combines focal loss with categorical cross-entropy is incorporated during training. Also another study by Yi et al. proposed CTANet, a hybrid CNN-ViT network for efficient diabetic retinopathy (DR) grading [130]. The CNN branch incorporates ODStarnet, leveraging multi-dimensional attention for precise local feature enhancement, while the ViT branch employs window-based attention to capture global structural context. The adaptive focus fusion (AFF) module dynamically integrates local and global features for optimal decision-making. CTANet achieves state-of-the-art accuracy of 90.08% on the APTOS 2019 dataset. Both of these studies made enhancements in fusion strategies between CNN and ViT in order to solve the static fusion inadequacy problem in existing methods. However, the enhancement at fusion strategy not necessary if the CNN-ViT make another enhancement can work well.

The motivation for choosing Hybrid CNN-ViT as the defect classification model is due to its ability to provide detailed feature extraction and contextual analysis, which overcome the traditional methods. This hybrid CNN-ViT offers a powerful approach to classifying the defects, making it the ideal choice to be implemented in this study. The

combination of CNN and ViT produce a hybrid model, which allow more comprehensive extraction of features across different scales and parts of defect structures. Hybrid models are particularly useful in AM, where defects can be both local (small pores, cracks) and global (large warping, stringing).

2.6 Summary of Machine Learning Models

Table 2.5 summarizes the machine learning models used in the defect detection and classification of AM products, with each model having strengths and limitations.

Table 2.5
Summary of Machine Learning Models Used in the Defect Detection Method of AM Products with Their Strengths and Limitations.

Author	Method	Description	Limitation
Baturynska and Martinsen [83]	Regression	-Detect and quantify geometric defects (dimensional deviations).	Depend heavily on manual feature engineering, unable to discover hidden representations.
Tapia et al. [84]		-Learn and predict the porosity in metallic parts.	
Garcia et al. data [85]		-Predict dimensional quality.	
Ferreira et al. [87]	Bayesian	-Modelling expected geometric deviations and identifying abnormal deviations.	Bayesian models require predefined probability distributions and Bayesian models struggle to capture highly nonlinear or hierarchical due to only few layers are used.
Aminzadeh and Kurfess [88]		-Monitoring system for quality of fusion and defect formation in every layer.	
Hertlein et al. [89]		-Predicting the quality of printed SLM parts.	
Gui et al. [90]	SVM	-Detected, classified, and avoided internal defects in metal parts produced by PBF-EB.	SVM cannot learn new or hidden features.
Liu et al.[91]		-Real-time defect detection in LAM.	
Wang et al. [92]		-In-situ quality monitoring for SLM.	
Trovato and Cicconi [94]	DT	-Evaluate the printability of metal parts for AM..	-Depend heavily on expert-defined rules and selected parameters.
Aziz et al. [95]		-Automatic classification of typical defects found during metallographic examination.	-Rely on manually measured geometric descriptors, only capture limited information from micrographs.
Cheepu [93]		-Identify defect in WAAM.	-Lead missing defect information.
Westphal and Seitz [71]	DL (CNN)	-Classify good and defective image from SLS.	CNN unable to represent the large distance relationship.
Yang et al. [72]		-Classify four sizes of the melt pool from the L-PBF.	
Zhang et al. [73]		-Predicting the porosity attribute	
Wang et al. [16].	DL (ViT)	-Detect defects of FFF products.	May struggle to extract local features.
Singh et al. [74]		-Detect three types of defects which are warping, layer delamination and gaps in raster line.	
Uysalel et al. [75]		-Classify the main types of AM defects based on their location, size, morphology.	

2.7 Summary

In summary, this chapter reviews the classification of AM products. Traditional defect detection methods, such as manual inspection, rule-based algorithms, and non-destructive testing, are discussed, proving their limitations in scalability and accuracy. Then, the chapter review the current method using the deep learning approaches. Classical ML methods improved upon rule-based systems but relied on handcrafted features. Deep learning introduced automatic feature extraction, specifically CNN performed well in extracting local features and ViT perform well in extracting global features. ViT enhancements, focusing on improvements at the patch embedding layer and token selection mechanisms. The enhancements address computational challenges, making them more suitable for complex tasks such as AM defect classification. Finally, this review considers the hybrid CNN-ViT architectures are highlighted as combining these strengths to achieve better accuracy, robustness, and generalization to classify defects of AM products. The chapter concludes by identifying the research gap: existing methods are limited when applied alone, motivating the proposed CNN-enhanced ViT model for improved defect detection.

CHAPTER 3
METHODOLOGY

3.1 Introduction

This chapter presents the methodology used to develop, train, and evaluate the proposed CNN-enhanced ViT model for AM defect classification. This section outlines the methodology employed to integrate CNN as local feature extraction with ViT as global feature extraction method in order to classify AM product according to crack, stringing and no defect class. There are 3 phases involved in developing the AM defect classification system, including the input phase, the model development phase, and the evaluation phase, as depicted in Figure 3.1. The dataset used in this study is discussed at the input phase. The model development of CNN-enhanced ViT developed in this research is described in detail in the model development phase. The performance indicators used for the classification task are described in the evaluation phase.

Figure 3.1 The Block Diagram for AM Defect Classification System Consists of Input Phase, Model Development Phase and Evaluation Phase.

3.2 Research Methodology

Figure 3.2 depicts the overall flowchart for this study. Initially, the AM images are collected from Malaysia Automotive, Robotics, and IoT Institute (MARii). The collected images consist of crack, stringing and no defect class. The collected data used as the input to the proposed model of CNN-enhanced ViT. The model first developed as CNN-ViT in order to extract local and global features of AM product. Then, the developed CNN-ViT being enhanced to address the limitations in single ViT architecture. The process of developing and enhancing the model involve objective 1. Next, the proposed CNN-enhanced ViT model further used to train and validate the model using MARii dataset and evaluate its performance using appropriate metrics in order to fulfil objective 2. Finally, the performance of the proposed model is compared

with original CNN-ViT and Swin transformer model using the same dataset and environment setting in order to fulfil the objective 3.

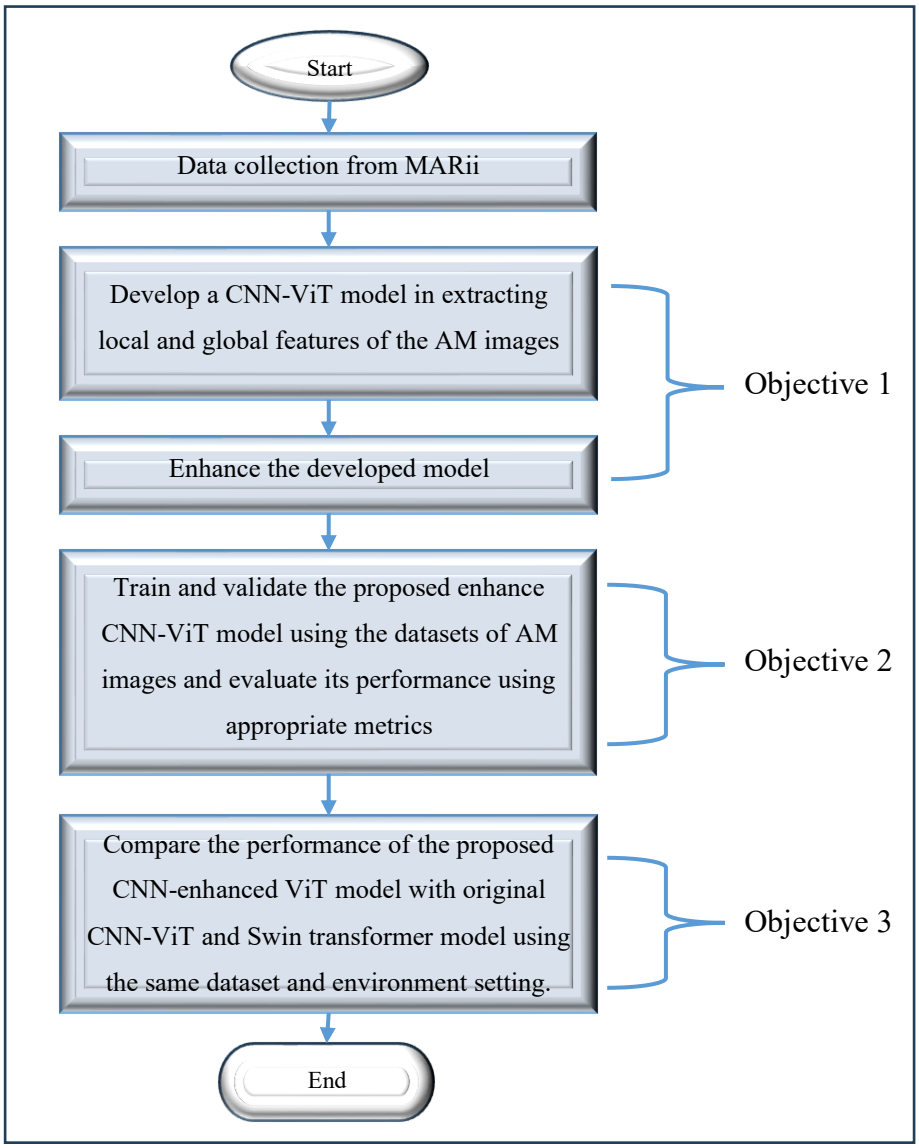


Figure 3.2 The Flowchart of the Whole Research Methodology for the AM Defect Classification System.

3.3 Input Phase

Data collection process is the main process at the input phase. Then, data augmentation was applied to the dataset in order to enhance the model robustness and mitigate imbalanced data issue, thereby increase variability and improve generalization. After augmentation, few image-preprocessing steps applied aims to standardize the dataset and optimize it for efficient training.

3.3.1 Data Collection

The dataset used in this study consists of addictive manufacturing (AM) images supplied by the Malaysia Automotive, Robotics, and IoT Institute (MARii), where the images were captured using a mobile phone camera with a resolution of 640 × 480 pixels under normal room lighting conditions. The dataset comprises 750 images as tabulated in Table 3.1 consist of three different classes, crack, stringing, and no defect, classify using identified characteristics described in previous literature [58, 131, 132]. This study only considered two types of defects, crack and stringing, as the dataset supplied by the MARii has no other types of defects involved.

Table 3.1
The Total of Raw Images According to Crack, Stringing and No Defect Class Before Augmentation Process.

Defect class	No. of images
Crack	256
Stringing	244
No defect	250
Total images	750

The examples of the AM images from crack, stringing and no defect classes s are shown in Figure 3.3 while Table 3.2 explains the description of each class.

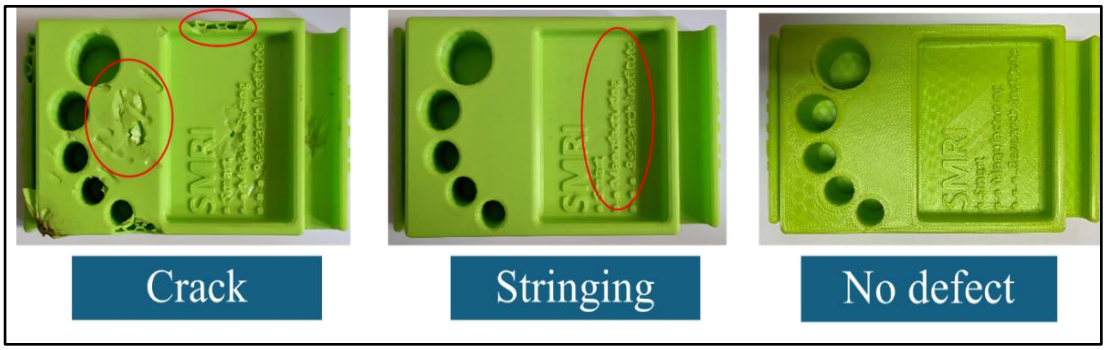


Figure 3.3 The Example of Defects on AM Products A) an Example of Crack Defect B) an Example of Stringing Defect C) an Example of No Defect.

Table 3.2
The AM Class Involved in this Study with the Description for Each Class.

AM Class	Description
Crack	The images are categorized as crack defects due to the present features with narrow openings in the body of the product [131], and have discontinuities or disconnections over the body due to residual stress occurrence [58].
Stringing	A feature like small or tiny strings from the same material are connected from one part to another part and remain on the surface of the printed product, which causes an aesthetic failure [131, 132].
No defect	The product falls under the no defect product if the printed product fulfils the quality standard set by the manufacturer and has no defect features as mentioned in the two listed types of defects.

3.3.2 Data Augmentation

This dataset exhibits an imbalanced distribution across three classes as some class imbalance persists due to the under-represented defects. Thus, the model could exhibit overfitting to most common class, resulting in poor performance for the under-represented samples [127]. To address this challenge and improve robustness of the model, data augmentation was carried out to ensure that the model focuses on relevant features, avoids overfitting, and handles subtle variations. Three basic geometric augmentation involving cropping, flipping and random rotation applied to all images in order to increase the data size and improve data diversity.

Initially, all images were cropped manually to reduce the background without cropping any parts of AM product. The cropped images were then manually rotated from 90° to 270° and flipped horizontally, vertically and both in order to increase data size and diversity. The rotation manually performed in this study, which mean the augmented images did not loss any significant amount of pixel information when rotating as stated by Alomar et al. [133], as the loss might occur when rotation perform through the coding. After applying these augmentation techniques, the total number of images increased to 1500 images with a balanced distribution across three classes. In addition, this study did not use any transfer learning or pre-trained models such as ImageNet as it is not the best solution to train model if the transfer learning scenario is different between the source domain and the target domain, leading the results of machine learning are not ideal [134]. It may not capture specific defect features of AM products in MARii dataset. The aim of this study is to evaluate the performance of the

proposed model from scratch and without external bias, where the pre-training has a little or no influence on the performance in a specific domain [135]. Table 3.3 summarizes the distribution of data which each classes consist of 500 images. Then, from the total images of the each classes, 70% are used for training, 20% for validating while remaining 10% are for testing, as the ratio obtain from this reference [136].

Table 3.3
The Data Distribution for Training, Validating and Testing for Each Defect Class: Crack, Stringing and No Defect of AM Products.

Class/Percentage	Training (%)	No. of images	Validating (%)	No. of images	Testing (%)	No. of images	Total images
Crack	70	350	20	100	10	50	500
Stringing	70	350	20	100	10	50	500
No defect	70	350	20	100	10	50	500
Total images							**1500**

3.3.3 Image Pre-processing

Image pre-processing were created using a simple procedure, as it is essential to refine the quality and intentionally make it suitable to integrate into a machine learning model [137]. The pre-processing process was carried out automatically by a self-programmed Python script as shown in Figure 3.4. All images, initially received in three color channels, red, green, and blue (RGB) are converted to grayscale images using the average grayscale method, calculated using Equation 3.1, to reduce large storage requirement and high computational costs by reducing colour information costs [138]. This method will calculate the average of each pixel's RGB value, then it will change to a grayscale value.

$$Average\ grayscale = \frac{R + G + B}{3} \tag{3.1}$$

where R is pixel value in red chanel, G is pixel value in green chanel and B is pixel value in blue channel.

All the grayscale images get sharpened by applying unsharp masking. The unsharp masking used to enhance the visibility of structures and details within in image [139]. The images were resized to 256x256 pixels, as image with higher resolution

enable the model to make more accurate predictions as it provide more detailed information [140].

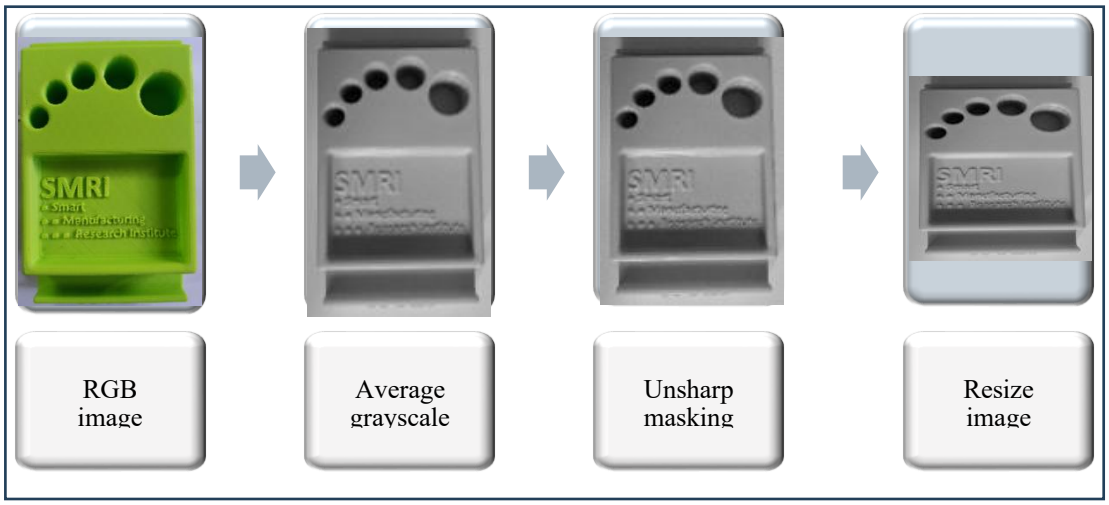


Figure 3.4 The Pre-processing Steps Applied to All Images Before Input to Model Development Phase, where the RGB Images are Converted to Grayscale Image using Average Grayscale, Follow with Unsharp Masking and Images are Resized to 256x256 Pixels.

3.4 Model Development Phase

The model development phase describes how the CNN is combined together with ViT architecture as both architectures require proper integration to make it works properly. Then, this section explains the details of full architecture developed in this study and how enhancement is made in enhanced ViT block as original ViT has limitation to overcome in order to make it suitable for computer vision task, specifically image classification task. This section also discusses the contribution of developed CNN-enhanced ViT which design to manage with small dataset of AM product. Finally, the overall architecture with its output shape per each layer is tabulated in table for future reference and future development. Note that this section performs feature extraction, as per discuss earlier, the extraction involves local and global feature extraction, as local features extracted using CNN and global features extracted using ViT.

3.4.1 CNN-enhanced ViT

The CNN-enhanced ViT refines the drawback of both standalone architectures. Figure 3.5(a) depicts the framework of CNN-ViT model proposed by Dosovitskiy, et

al. [15] while Figure 3.5(b) depicts the framework of the proposed CNN-enhanced ViT model.

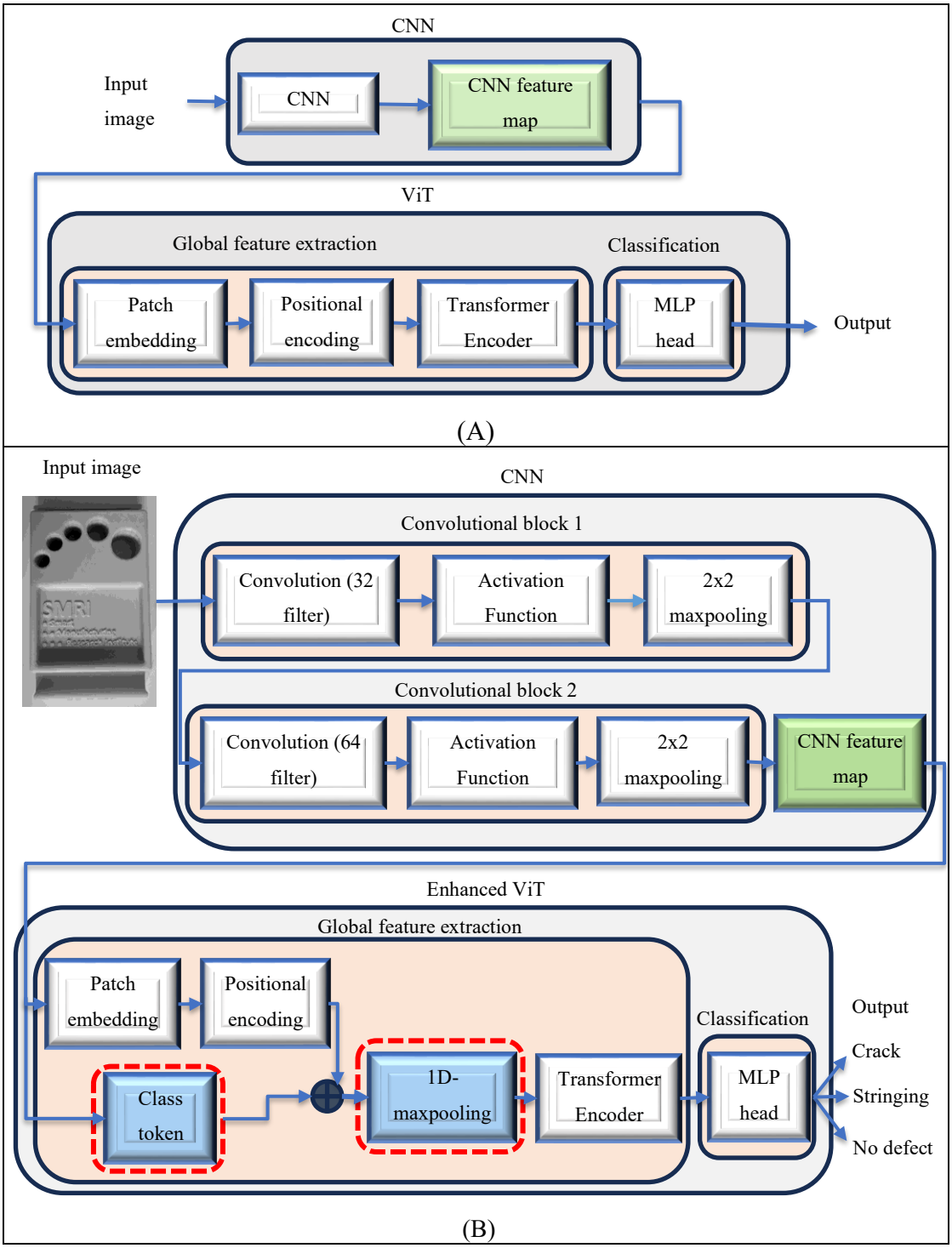


Figure 3.5 (A) The Framework of CNN-ViT Model Proposed by Dosovitskiy et al.[13], while (B) the Framework of the Proposed CNN-enhanced ViT Model.

CNN offer architectural flexibility and enable customization for a specific task. Based on Figure 3.5(b), a CNN comprises of two convolutional blocks are used to

extract local features and generate the feature map. Each convolutional block consists of a convolution layer, a non-linear activation function (ReLU), and a max-pooling layer. The difference between convolutional block 1 and convolutional block 2 is the convolution layer. The convolutional block 1 utilizes convolution layer with 32 filters, while the convolutional block 2 utilizes convolution layer with 64 filters. The output from the maxpooling layer of the second convolutional block, with height, width and dimension (64,64,64) will be treated as the CNN feature map since the fully connected layer of the CNN is removed. The configuration of simple network of CNN used in this study is to make it less prone to overfitting and generalizing better when facing small datasets as proven by Brigato and Iocchi [141].

The CNN feature map is then passed to the proposed enhanced ViT for global feature extraction and classification. The original ViT divides the original image into patches. In this study, the CNN feature map is divided into a 16x16 grid, producing 16 patches with each patches have 16384 dimensions. At the patch embedding layer, these 16 patches with 16384 dimensions are projected to 64 dimensions by using Equation 3.2, which produce 16 token sequence, with 64 features, denoted as (16,64).

$$Z_0 = TW_e + b_e \tag{3.2}$$

where Z_0 is the token sequence, W_e is projection matrix and b_e is the bias parameters.

However, the patch embedding layer leading to marginalization of information, loss of local tokens features [18], and cannot capture sequence ordering of input tokens and require positional encoding [142]. Thus, at the positional encoding layer, all token sequences are assigned with positional encoding with same dimension (16,64), respectively, to restore it before further processing in the following layers. This process is calculated using Equation 3.3.

$$Z = Z_0 + P \tag{3.3}$$

where Z is the output of patch embedding added with positional encoding, P is the positional encoding.

While positional encoding represents position for each token sequence, it has no global information as reference. The positional encoding alone is not consistent across scale, whereby the global information is easily lost in the meantime of training or validating the model. The feature collapse in deeper layers, resulting in the vanishing of low-level visual features [124]. Hence, in this study, a class token highlighted in a

dotted-red box as in Figure 3.5(b), is proposed which aims to support the positional encoding and to stabilize the positional encoding across scales, which aims to improve the robustness and generalization of image classification. The difference of this enhanced ViT is the class token, where the CNN feature map is flattened entirely, treated as one global context, as in Equation 3.4, which produce (1,64) output shape. This global context is used as the class token (known as [CLS] token). This [CLS] token is related to the CNN feature map rather than directly input external [CLS], which carries no information until it goes through the transformer encoder.

$$T = reshape(F) \tag{3.4}$$

where T is the global context which serves as class token, and F is the CNN feature map.

Both outputs from patch embedding and positional encoding and the [CLS] token are combined by concatenate into a single output, as in Equation 3.5. The introduction of this method helps to keep local and mid-level detail while allowing ViT to extract global features, making it have stronger local-global continuity.

$$Z_{hybrid} = Concat(\mathrm{Z}, \mathrm{T}) \tag{3.5}$$

where Z_{hybrid} is the combined token sequence.

Every pixel from the concatenated output is treated as an input to transformer encoder. However, if every pixel is treated as input to transformer encoder, redundant information is passed to the transformer encoder. The computation increases exponentially as the number of tokens increases, whereas most of tokens are redundant [19]. Token selection overcomes the computational cost by reducing the length of the token sequence to be processed by the transformer encoder [143], despite accelerating the speed of ViT [19]. This study fixed the problem by adding the token filtering using 1D-maxpooling highlighted in a dotted-red box as in Figure 3.5(b), to retain only informative token input to the transformer encoder. The selected informative token is calculated using Equation 3.6. The 1D-maxpooling with kernel size of 2 retain 1 token out of every 2, which reduces the sequence length by half. This process helps to reduce redundancy, without affecting the global context.

$$Z_{selected} = MaxPool1D(Z_{hybrid}) \tag{3.6}$$

where $Z_{selected}$ is the selected token, $MaxPool1D$ is 1D-maxpooling.

The selected token is then processed by a single transformer encoder, which also differ from original CNN-ViT proposed by Dosovitskiy et al.[15], which uses repeated number of transformer encoder, specifically 12 transformer encoders. A transformer encoder contains two sub-layers of multi-head self-attention (MHSA) and multi-layer perceptron (MLP) block. Each sub-layer is preceded with layer normalization (LN) and followed with residual connection. The first sublayer with MHSA is used to capture long-range dependencies and global context, as in Equation 3.7. The second sublayer with the MLP block is calculated using Equation 3.8.

$$Z_l' = MHSA\big(LN(Z_{selected})\big) + Z_{selected} \tag{3.7}$$

$$Z_l = MLP\big(LN(Z')\big) + Z_l' \tag{3.8}$$

where Z_l' is the output from the first sub-layer of transformer encoder and Z_l is the output from the second sub-layer.

Finally, the output is passed through a MLP head to classify additive manufacturing (AM) images into crack, stringing or no defect. The classification of the ViT model is performed by the MLP head, which consists of a fully connected layer (dense layer) with 512 units and the rectified linear unit (ReLU) activation function. The [CLS] token after passing through the transformer encoder is transformed into a richer representation for classification. A dropout layer with a retention rate of 0.5 randomly drops 50% of neurons to prevent overfitting during training with regard to the small dataset used. The softmax activation function is applied at the final classification layer to determine the class output based on the probability score. The detailed architecture of CNN-enhanced ViT is tabulated in Table 3.4, where the two lines shaded in blue represent the class token and 1D-maxpooling proposed in this study.

Table 3.4
The Details of CNN-enhanced ViT Architecture.

Phase	Layer (type)	Output shape	Setting / Name
Input	Input 2D data	(256, 256, 1)	AM images
Local feature extraction	2D-Convolution-1	(256, 256, 32)	32 filter, kernel size = 3x3
	Activation function-1		ReLU
	2D-Maxpooling-1	(128, 128, 32)	Kernel size = 2x2
	2D-Convolution-2	(128, 128, 64)	64 filter, kernel size = 3x3
	Activation function-2		ReLU
	2D-Maxpooling-2	(64, 64, 64)	Kernel size = 2x2
Global feature extraction	Reshape-1	(16, 16384)	
	Dense-1	(64)	
	Dense-2	(16, 64)	Patch embedding
	Add-1	(16, 64)	Positional encoding
	Flatten-1	(262144)	
	Reshape-2	(1, 64)	Class token
	Concatenate-1	(17, 64)	
	1D-Maxpooling-1	(8, 64)	Pool size= 2, stride=2
	Multi-head self-attention	(8, 64)	Number of head =8
	Dropout-1	(8, 64)	0.1
	Add-2	(8, 64)	
	Layer normalization-1	(8, 64)	
	Dense-3	(8, 512)	
	Dense-4	(8, 64)	
	Dropout-2	(8, 64)	0.1
	Add-3	(8, 64)	
	Layer normalization-2	(8, 64)	
Classification	Dense-5	(512)	
	Activation function-3		ReLU
	Dropout-3	(512)	0.5
	Dense-5	(3)	
	Activation function-4		Softmax

3.5 Evaluation Phase

The performances of the developed CNN-enhanced ViT model were evaluated based on diverse evaluation measures, including the confusion matrix and performing a comparative analysis between CNN-enhanced ViT, CNN-ViT, and Swin transformer models for the MARii dataset.

3.5.1 Confusion Matrix

The confusion matrix is being used as the primary evaluation metric, as it provides detailed performance analysis for each class involved beyond measuring

accuracy only. By quantifying true positives, true negatives, false positives, and false negatives, it enables the calculation of precision, recall, and F1-score, which are essential for assessing the reliability of defect classification. This ensures that the model is not only accurate but also balanced in detecting all defect types, making the evaluation more robust and meaningful for practical deployment.

The output from the classification process is categorized into three classes: crack, stringing, and no defect. The performance of the developed CNN-enhanced ViT is then analyzed by calculating the accuracy, precision, sensitivity and F1-score of the model, as shown in Equation 3.9 – Equation 3.12.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{3.9}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.10}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.11}$$

$$F1 - score = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision} \tag{3.12}$$

where TP is the number of true positive, TN is the number of true negative, FP is the number of false positive and FN is the number of false negative [144].

In this study, the crack defect is considered as the positive condition, whereas the no defect and stringing defect are considered as negative condition. Specifically, TP represents the correct identification of crack. TN represents the correct identification of no defect and stringing. FP represents the incorrect identification of no defect or stringing as crack defect, while FN represents the incorrect identification of crack as no defect or stringing. This approach is applied separately for no defect and stringing to calculate the respective performance metrics.

3.5.2 Comparative Analysis

The proposed CNN-enhanced ViT model is compared with recent hybrid model including the original CNN-ViT and Swin transformer [21]. Swin transformer chosen as the comparison since it capable of extracting local and global features similar like

CNN-ViT. In Swin transformer, the local features extracted through window-based self-attention and global features extracted via shifted window mechanism and hierarchical design. It constructs hierarchical representation by starting from small size patches and gradually merging the patches in deeper layer. It is suitable to make the Swin transformer as the baseline model due to its capabilities in extracting local and global features.

3.6 Experimental Setup

The ratio of 70%, 20%, and 10% with respect to training, validating, and testing is the exact ratio used in this study within the MARii dataset. All these experiments utilize the Tensorflow framework, Python 3 as the computing language, and Google Colab as the development environment. The parameter setting for each model is set to the default value as developed in the original model.

3.7 Summary

This chapter presents the development of a hybrid deep learning model, CNN-enhanced ViT, designed to classify additive manufacturing (AM) products into three categories, which are crack, stringing, and no defect. The development process begins with the input phase, where the data collection, image preprocessing and data augmentation occur. The input phase is used to enhance the feature representation and ensure suitability for deep learning input. Then, at the model development phase, this study proposed a hybrid model that integrates two architectures. The CNN is the first architecture used for local feature extraction and ViT is the second architecture for global feature extraction. A class token is added in ViT architecture to cope with loss of information at the patch embedding layer, and the addition of 1D maxpooling added before transformer encoder as the only token sequence with rich of information being processed at the following layer to increase the computational efficiency. The model performance is subsequently evaluated using unseen data to assess its generalization capability. The proposed approach may effectively classify AM defects across varying spatial scales, from localized to widespread defects. Moreover, this CNN-enhanced ViT enhances classification accuracy while reducing processing time and offers a more efficient and reliable solution for the AM defect classification system.

The developed model can be extended to other defect types by replacing the current dataset with a new defect dataset. Users only need to change the number of classes and update the training and testing dataset paths according to their local directory. The main algorithm does not need to be modified because the model can automatically extract relevant features from the input images. The model was also tested on the NEU surface defect dataset, a hot-rolled steel strip surface defect dataset created by Song and Yan [145], and achieved good performance in terms of accuracy, precision, sensitivity, and F1-score. Therefore, as long as the dataset is sufficient and well-balanced across all classes, the model can be applied to different defect classification tasks beyond the AM scope. In addition, the source code will be publicly available on GitHub for future use and extension by other researchers.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Introduction

This study presents a CNN-enhanced ViT architecture designed to effectively capture both local and global features for accurate defect classification of AM products. This chapter reports the experimental outcomes corresponding to the methodologies outlined in Chapter 3. The effectiveness of the proposed model for defect classification of AM products using the MARii dataset was evaluated by comparing the CNN-enhanced ViT model with the CNN-ViT and Swin transformer models. All the models were evaluated with the same machine and environment. The results are analyzed comprehensively, including accuracy-epoch plot and evaluation metrics such as the confusion matrix. These outcomes are interpreted in the context of key architectural components including the patch embedding receiving input from the CNN feature map, the addition of class token, the combination of the token sequence and the pooling mechanism.

4.2 Dataset

In this study, a total 1500 MARii dataset being used to test the developed model, CNN-enhanced ViT and another model of CNN-ViT and Swin transformer. The features from the input image may not degrade through the utilization of grayscale images since the defect features influence by the structure, but not the hue of the colour, which not discards any discriminative features.

4.3 Accuracy-epoch Plot

The accuracy against epochs plot for the CNN-enhanced ViT models attained during the training over 15 epochs are depicted in Figure 4.1. The accuracy against epoch plot reveals the effectiveness for this proposed model to learn the features over the epochs.

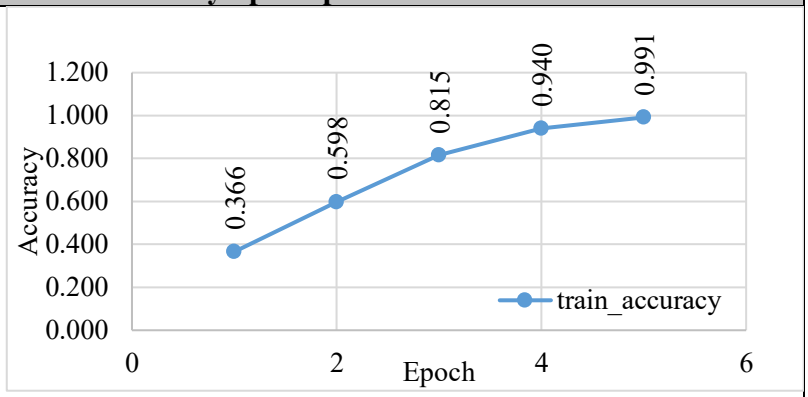
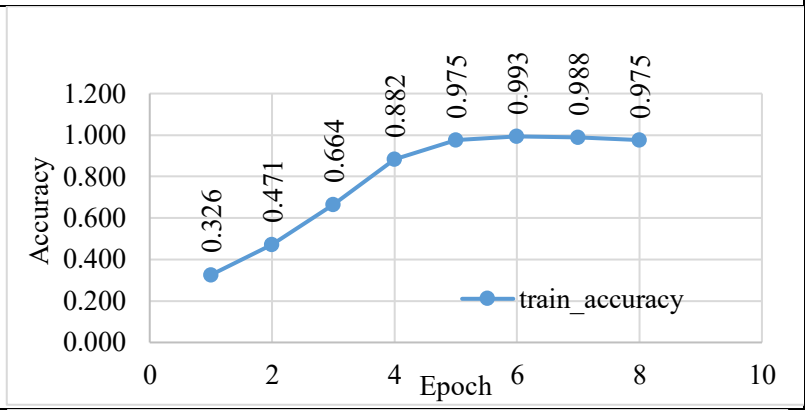
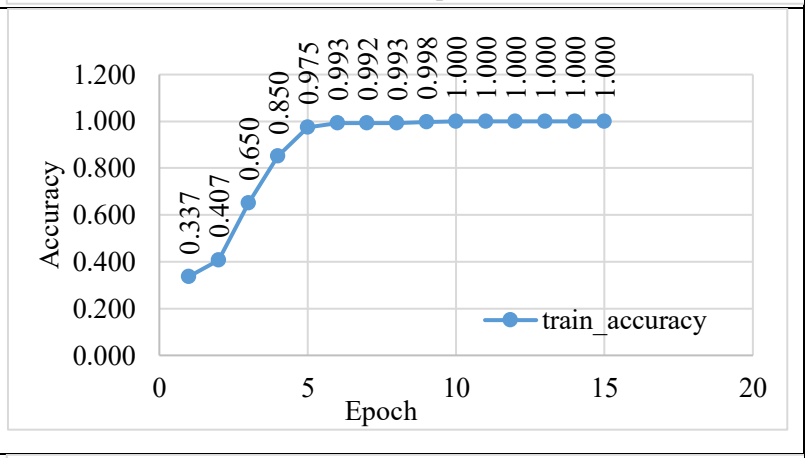
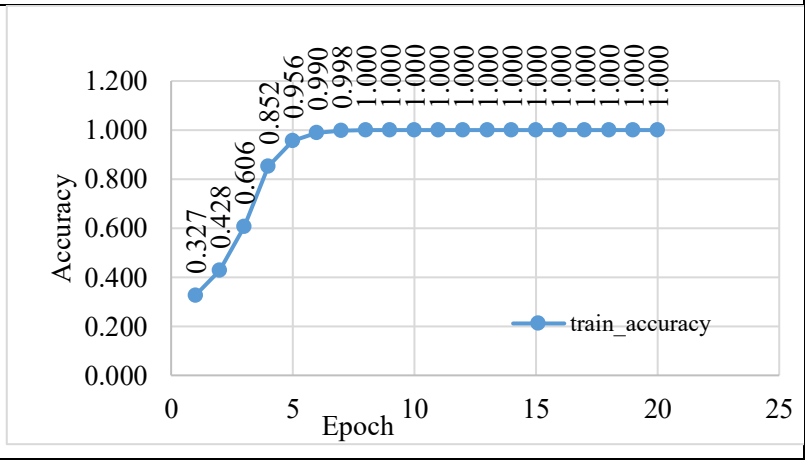
Number of epoch	Accuracy-epoch plot for CNN-enhanced ViT
5	
8	
15	
20	

Figure 4.1 The Accuracy-epoch Plot for CNN-enhanced ViT for 4 Different Epochs.

Based on Figure 4.1, the training accuracy of CNN-enhanced ViT increases rapidly from 0.37 to 0.99, and by 8 epochs, the training accuracy reaches 0.98. The CNN-enhanced ViT achieves 1.00 accuracy at 15 and 20 epochs respectively and shows no further improvement as the model has already saturated with maximum accuracy. The proposed model also shows consistency across runs as the starting accuracy values range from 0.33 to 0.370 and similar across different epochs setting. This result shows a stable initialization while learning the AM defect features. The proposed model is effective on the tested MARii dataset, as it learns discriminative features quick despite reach near to perfect classification within 10 epochs. Since the model has stable accuracy at 15 epochs, training with 20 epochs may increase the computation without giving any benefit, which suggest the optimal epochs to stop is around 8 to15 epochs. Hence, in this CNN-enhanced ViT model, the epochs is chosen to be 15 epochs.

4.4 CNN Filter

The configuration of CNN filters in the convolutional blocks plays an important role in determining feature extraction capacity. The number of filters evaluated is (8,16), (32,64), and (64,128) across two convolutional blocks, and the performance is tabulated in Table 4.1.

Table 4.1
The Variation Number of Filters Tested for Two Convolutional Blocks in CNN Architecture.

CNN filters	Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
8,16	Crack	96.00	94.00	94.00	94.00
	No defect	92.00	97.50	78.00	86.67
	Stringing	92.00	81.67	98.00	89.09
	Average	93.33	91.06	90.00	89.92
32,64	Crack	96.67	95.92	94.00	94.95
	No defect	96.00	97.83	90.00	93.75
	Stringing	95.33	89.09	98.00	93.33
	Average	**96.00**	**94.28**	**94.00**	**94.01**
64,128	Crack	96.00	94.00	94.00	94.00
	No defect	93.33	95.45	84.00	89.36
	Stringing	93.33	85.71	96.00	90.57
	Average	94.22	91.72	91.33	91.31

Based on Table 4.1, the experimental result demonstrates that the (32,64) filter configuration achieves the most balanced and superior performance across all four evaluation metrics in terms of accuracy, precision, recall, and F1-score compared to other tested filter settings. This configuration provides balanced performance in extracting low to high-level features regarding the dataset used. The average F1-score of the (32,64) filter is 94.01%, outperforming the average F1-score in the (8,16) filter of 89.92%, indicating that it is balanced in handling, these three classes. In addition, the (32,64) demonstrates superior generalization compared to (8,16) or (64,128), indicate it is the optimal setting for the CNN-enhanced ViT model tested on the MARii dataset.

4.5 Number of Convolutional Blocks

The comparative experiments to determine the number of convolutional blocks used in this study were also measured using a confusion matrix. Table 4.2 shows the comparative experiments between two convolutional blocks and four convolutional blocks.

Table 4.2
The Comparative Experiments Between Two Convolutional Blocks and Four Convolutional Blocks.

CNN filters	**Class**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Two convolutional blocks (32,64)	Crack	96.67	95.92	94.00	94.95
	No defect	96.00	97.83	90.00	93.75
	Stringing	95.33	89.09	98.00	93.33
	Average	96.00	94.28	94.00	94.01
Four convolutional blocks (32,64,128,256)	Crack	98.67	100.00	96.00	97.96
	No defect	95.33	93.88	92.00	92.93
	Stringing	94.00	88.68	94.00	91.26
	Average	96.00	94.19	94.00	94.05

Based on Table 4.2, the comparative experiments between two convolutional blocks (32,64 filters) and four convolutional blocks (32,64,128,256 filters) reveal that increasing network depth does not yield significant improvements in overall performance across the four measured metrics (accuracy, precision, recall, and F1-score). the average accuracy of two convolutional block is 96.00%, same as the average accuracy in four convolutional blocks, as well as average precision, sensitivity and F1-score are nearly identical. These results indicate that increasing deeper layer of CNN

may not provide greater performance. In addition, adding more convolutional blocks might increases computational cost, training time, and risk of overfitting. therefore, this study chooses to use two convolutional blocks of CNN as it has already achieved similar average performance to four convolutional blocks.

4.6 Confusion Matrix

The performance of the proposed model is assessed by using the confusion matrix. In this research, the performance of the proposed CNN-enhanced ViT model were compared with the CNN-ViT and Swin transformer as mentioned in objective 3. The confusion matrix depicted in Figure 4.2 highlights the comparison between the actual class and the predicted class for CNN-enhanced ViT, CNN-ViT and Swin transformer for each of the AM defect classes which are crack, stringing and no defect classes.

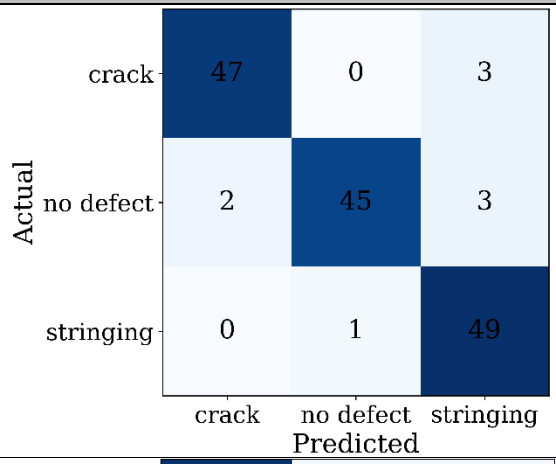
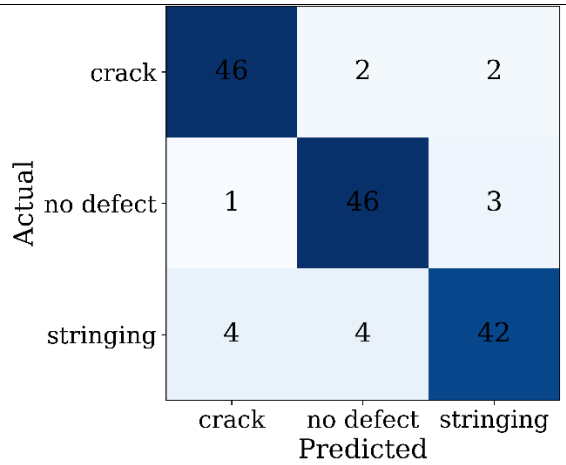
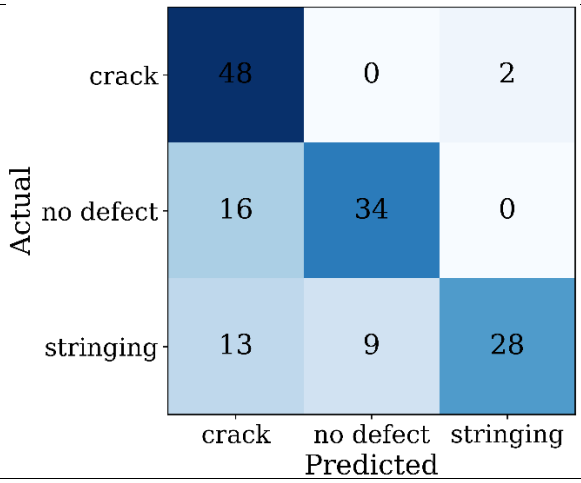
Model	Confusion Matrix
CNN-enhanced ViT	
CNN-ViT	
Swin transformer	

Figure 4.2 The Confusion Matrix of Actual Class and the Predicted Class for CNN-Enhanced ViT, CNN-ViT and Swin Transformer.

Based on Figure 4.2, the proposed model of CNN-enhanced ViT correctly classifies 47 out of 50 images for crack defects while only 3 images are misclassified as stringing, indicating that the proposed model effectively captures crack features. For no defect class, 45 images were correctly classified as no defect, 2 images were misclassified as crack, and 3 images were misclassified as stringing. Stringing defects

often appear in small, localized regions (thin filament strands), which can resemble the smooth background of no defect images. As of this feature overlap, the model sometimes confuses stringing with the no-defect class. Crack defects, on the other hand, introduce strong, high-contrast discontinuities that are visually and structurally distinct from no defect surfaces. Therefore, the model is less likely to misclassify no-defect images as cracks, but more likely to misclassify them as stringing.

In terms of stringing class, 49 images were correctly classified, and only 1 image was misclassified as no defect class. This superior performance of the proposed model shows that the feature map serves as the input to patch embedding and has preserved the continuity of local features. In addition, when the feature map is stored in the proposed class token, the model manages to mitigate feature collapse in deeper layers, as the class token serves global information as reference. Furthermore, the token selection proposed in this study enables only informative tokens to be processed by the transformer encoder. This process allows the transformer encoder to focus on important defect features, which can reduce misclassification.

Next, the CNN-ViT model also shows good performance, but slightly lower than the proposed model. In terms of crack class, 46 images were correctly classified as crack, but 2 images were misclassified as no defect and 2 images were misclassified as stringing. For no defect class, 46 images were correctly classified, while 1 image was misclassified as crack and 3 images were misclassified as stringing. Then, for stringing class, 42 out of 50 images were correctly classified as stringing, 4 images were misclassified as crack, and 4 images were misclassified as no defect. The performance is slightly low due to the absence of token selection. The transformer encoder has to process all tokens, including less informative tokens, so the model cannot focus only on important defect features.

Next, the Swin transformer shows weak performance, especially in the no defect and stringing classes. For crack class, 48 images were correctly classified as crack, and 2 images were misclassified as stringing class, which indicates this model is good at detecting crack defects. However, for no defect class, only 34 images were correctly classified, but another 16 images were misclassified as crack defects. In terms of stringing defects, only 28 images were correctly classified as stringing defects, while 13 images were misclassified as cracks and 9 images were misclassified as no defect class. Swin transformer that uses window-based self-attention to extract local features and shifted window being used for interaction between the extracted features. The use

of window-based self-attention in swin transformer limits its ability to capture global context. The attention is restricted to local windows, where distant features only interact gradually through shifted windows. Due to these limitations, it limits the model's ability to distinguish the defect patterns well.

4.7 Overall Performance

Furthermore, the overall performance of each of the model which are the proposed CNN-enhanced ViT, CNN-ViT and Swin transformer were compared and tabulated in Table 4.3 for each of the metrics - accuracy, precision, sensitivity, and F1-score, for each of the defect classes.

Table 4.3
The Overall Performance for the Three Models which are CNN-Enhanced ViT, CNN-ViT, and Swin Transformer Based on the Testing Images Are Presented in the Form of Accuracy, Precision, Sensitivity, and F1-Score for the Crack, Stringing, and No Defect Classes.

Model	Defect Class	Accuracy (%)	Precision (%)	Sensitivity (%)	F1-Score (%)
CNN-enhanced ViT	Crack	96.67	95.92	94.00	94.95
	No defect	96.00	97.83	90.00	93.75
	Stringing	95.33	89.09	98.00	93.33
	Average	96.00	94.28	94.00	94.01
CNN-ViT	Crack	94.00	90.20	92.00	91.09
	No defect	93.33	88.46	92.00	90.20
	Stringing	91.33	89.36	84.00	86.60
	Average	92.89	89.34	89.33	89.29
Swin transformer	Crack	79.33	62.34	96.00	75.59
	No defect	54.67	79.07	68.00	73.12
	Stringing	84.00	93.33	56.00	70.00
	Average	72.67	78.25	73.33	72.90

Based on Table 4.3, the proposed CNN-enhanced ViT model delivered an improvement in average accuracy of 96.00% which outperforms the CNN-ViT model of 92.89% average accuracy and Swin transformer model of 72.67% average accuracy across all the defect classes. The CNN-enhanced ViT model achieved the highest performance for all the metrics including accuracy, precision, sensitivity, and F1-score, in which all parameters exceed more than 94.00% for all classes. It shows that the CNN feature map serves as the input to patch embedding has preserved the continuity of local features. Furthermore, when the feature map is stored in the proposed class token, the

class token serves global information as reference. This makes the model manages to mitigate feature collapse in deeper layers. In addition, transformer encoder can focus on extracting important defect features as no redundant token being processed due to token selection has been applied.

On the other hand, the CNN-ViT model achieved an average accuracy of 92.89% for all the three defect classes. Unlike the CNN-enhanced ViT model, the CNN-ViT did not consist of the class token, as the model heavily relies on patch embedding only, making it misses some defect and thus reduce the accuracy.

Furthermore, the Swin transformer model did not perform very well with an average accuracy of 72.67% across all the three defect classes, due to some information would be lost during patch merging modules[146].

In terms of precision, the CNN-enhanced ViT model achieves a result of 94.28% across all defect classes with no defect class were precisely classify with a result of 97.83%. These results are higher than the other two models due to the class token helps the model to have global information as reference, as the features might collapse in deeper layers, resulting in the vanishing of low-level visual features [124]. In contrast, the CNN-ViT model delivers a precision of 89.34% across the three defect classes which indicate more false positive case occur especially in stringing classification. This reduction occurs because of the local features were not reinforced with the global context as in CNN-enhanced ViT model.

For the Swin transformer, the average precision of 78.25% for all the three classes is moderate but unstable indicates that many false positive occur. In terms of sensitivity, the CNN-enhanced ViT model delivers a promising result of 94.00% across all the three defects indicates that patch embedding extracted from the CNN is capable of extracting and preserving local details, so that the model rarely misses the defects as it can classify crack and stringing consistently.

On the other hand, the CNN-ViT model achieves 89.33% of the sensitivity, showing that the feature may collapse in deeper layers, resulting in the vanishing of low-level visual features [124]. Finally, the Swin transformer model produced an average of sensitivity of 73.33% across all classes but perform a very good result of 96% of sensitivity for cracks defect class.

The positional encoding alone is not consistent across scale, whereby the global information is easily lost in the meantime of training or validating the model.

Hence, based on the results that depicted in Table 4.3, it shows that the proposed CNN-enhanced ViT model is capable in classifying the defect classes and performed a significantly excellent results and performance as compared to the CNN-ViT model and Swin transformer model.

Next, Figure 4.3 shows the average performance of F1-score based on the CNN-enhanced ViT, CNN-ViT and Swin transformer models for all three defect classes which are crack, stringing and no defect classes.

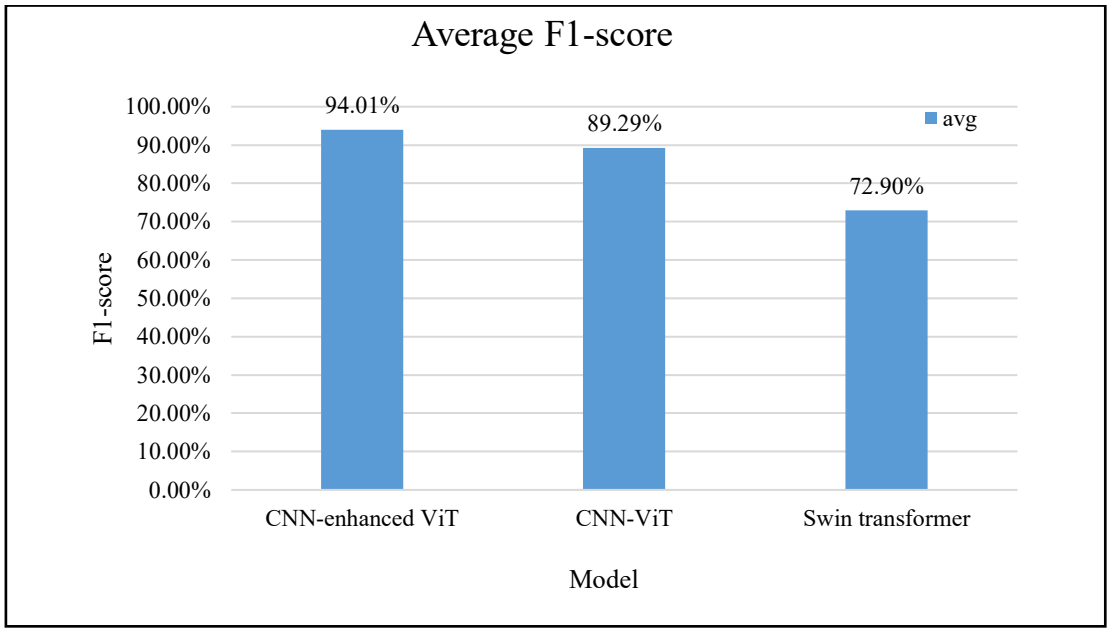


Figure 4.3 Average Performance Metrics Across Three Classes Involved According to Three Models Tested.

In terms of F1-score, the CNN-enhanced ViT model achieves the highest F1-score of 94.01% as depicted in Figure 4.3, indicates that it has the best balance between precision and sensitivity, leading to the least misclassification occurrence. The utilization of the CNN-enhanced ViT model ensures the most important local features are well preserved before input to the transformer, which explains the balance in the F1 score. In contrast, the average F1-score of CNN-ViT is 89.29%, proving that the hybrid structure improves the local feature extraction. However, some information may be lost due to the reliance on the flattening process. Finally, the Swin transformer achieves a result of F1-score of 72.90% as the precision and sensitivity is unbalanced due to some information would be lost during patch merging modules[146].

4.8 Summary

This chapter has presented a detailed discussion of the results, focusing on the main findings which alligned with the research objectives. It explained the performance of the proposed CNN-enhanced ViT model that is capable of extracting local and global features which significantly produced an excellent result in classifying the AM defect classes. Furthermore, the results explained the impact of using the CNN feature map as input to the patch embedding layer, the proposed class token to support global feature extraction, the combination of token sequences from patch embedding and positional encoding and the class token, and the impact of 1D-maxpooling in reducing the redundant token to be processed in the following layer. The performance of the proposed CNN-enhanced ViT model is compared with the CNN-ViT model and Swin transformer model using the confusion matrix and performance metrices. Lastly, the findings in this section were then concluded in Chapter 5.

CHAPTER 5
CONCLUSION

5.1 Conclusion

The application of DL in the AM sector has become increasingly widespread in recent years. DL especially CNN knowns for its excellent performance in AM defect classification. However, conventional deep learning methods for AM defect classification primarily rely on a single CNN architecture to analyzes features of the images. While classification models based on CNN have shown good performance in classifying defects, its capabilities limit to local feature extraction only. This approach often overlooks the global features of images which highlighting a significant drawback. In contrast, ViT helps the model in extracting global features due to capabilities of better understanding the interactions between distant parts of an image. In addition of using pure CNN for classification tasks, some researchers have explored combining CNN and Transformers to fully integrate the advantages of both architectures. The combination allows the model to capture detailed information and understand how these details connect in the larger context. However, during the training deep learning model, greater accuracy often has a large number of parameters. Therefore, finding a balance between accuracy and computational resources remains a significant challenge.

The primary focus of this study is to develop an enhanced model of CNN-ViT model in extracting local and global features of the AM images and classify the defect according to crack, stringing and no defect product. In addition, the developed model aims to balance between performance and computational efficiency. The data collection involved the AM dataset supplied by MARii, consist of two types of AM defects, crack and stringing and no defect product. The proposed model named as CNN-enhanced ViT use CNN architecture to extract local features. The output from the CNN (feature map) is then set as the input to vision transformer as it will be process by patch embedding and positional encoding layer in ViT. Not like previous study, this study also saved the extracted features from the CNN as the class token. In addition, this study enhances the model by performing token selection in order to increase computational efficiency.

The result in this study present in confusion matrix to measure the performance of the proposed model and compared with another model. The CNN-enhanced ViT model has demonstrated a strong performance in classifying no defect and two types of defects, crack and stringing in the AM product. Results were compared with CNN-ViT and Swin transformer models where the CNN-enhanced ViT model demonstrated highest average accuracy of 96.00%, precision of 94.28%, sensitivity of 94.00% and F1-score of 94.01% accordingly across all the three defect classes. On the other hand, for CNN-ViT and Swin transformer models, the results from the confusion matrix showed lower performance in terms of average accuracy, precision, sensitivity, and F1-score ranging between 72.67% to 92.89%. From the enhancement, the results lead to more precise in classifying types of AM defects and differentiate with no defect product. Furthermore, the automation of identifying the AM defect features using the proposed model possibly can reduce the time consumption when performing defect classification.

In this study, all research objectives were successfully achieved. A CNN-enhanced ViT model was developed to effectively extract local and global features of AM images and classify defects into crack, stringing, and no defect categories. The model was trained and validated using AM datasets, with performance evaluated through accuracy, precision, recall, and F1-score, showing strong results. Furthermore, comparative analysis confirmed that the proposed CNN-enhanced ViT model outperformed the original CNN-ViT and Swin Transformer under the same dataset and experimental settings, demonstrating its robustness and suitability for defect classification in additive manufacturing.

5.2 Research Contribution

This study develops a CNN-enhanced ViT architecture that aims to improve the AM defect classification system through the combination of local and global feature extraction with additional of class token and perform token selection mechanism. The placement of CNN before ViT enables the model to capture both low-level texture and capture global relationships through global feature extraction. The main contribution of CNN-enhanced ViT relies on its architecture. The proposed class token extracted from the CNN feature map may serve as a reference for global information in order to support the positional encoding and to stabilize the positional encoding across scales. The class token is important, as the positional encoding (a layer in ViT) alone is not consistent

across scale, whereby the global information is easily lost in the meantime of training or validating the model.

In addition, this study further contributes to enhance the ViT architecture by performing token selection using 1D maxpooling layer. This 1D maxpooling layer is used to filter redundant or noisy token before further processing using the transformer encoder. The token sequences need to filter the information, as if every pixel is treated as input to transformer encoder, redundant information is passed to the transformer encoder. The redundant information will constrain the following transformer encoder to process the input, leading to high memory usage and high computation. The addition of 1D maxpooling aims to retain only the most informative token input to the transformer encoder.

The proposed CNN-enhanced ViT model contributes to produce a lightweight architecture by employing a single transformer encoder block instead of the standard twelve-layer transformer encoder in ViT. This architectural design helps the model to reduce the inference time despite reducing model complexity, making it suitable for real-time deployment in industrial quality inspection regarding AM defect classification.

5.3 Limitations and Future Recommendations

Although the developed model of CNN-enhanced ViT is able to produce a very good result, but for this study the model was implemented on a limited dataset, which inevitably limits the generalizability of the findings. Cross-dataset validation or testing on external datasets would provide stronger evidence of robustness across different imaging conditions and defect types. Furthermore, this study uses grayscale images within a relatively small dataset, so the utilization of a colour dataset may restrict the model's ability.

Moreover, a total of 1500 images considered as a good number of dataset to train deep learning model especially ViT. This study leverages CNN-enhanced ViT architectures that incorporate inductive biases from convolution, which are known to perform better in small dataset. The aim of this study is not to claim scalability to very large datasets, but to demonstrate that with appropriate architectural choices, meaningful performance can still be achieved even under constrained data availability. It is suggested that extensive data augmentation should be applied in order to improve

diversity and reduce the chance for overfitting despite larger datasets should be tested in the future in order to improve generalization.

Furthermore, this study employed data augmentation by rotating within the range of 90° to 270° to expand the dataset. While this technique effectively increases the number of samples, it primarily introduces only minor variations from the original images. Consequently, the diversity of the dataset remains limited. To achieve greater variability and improve the robustness of the model, it is advisable to incorporate advanced augmentation strategies. For instance, generative adversarial network (GAN) can be utilized to synthesize realistic and diverse images, which is particularly beneficial when the original dataset is small or difficult to obtain.

A modified vision transformer (ViT) architecture has been used in this study, utilizing a single transformer encoder block instead of the standard configuration comprising 12 identical transformer encoders. This adjustment results in a significantly more lightweight model, making it suitable for resource-constrained environments. Hence, for future work, it is recommended to explore the impact of replacing the ViT backbone with alternative lightweight architectures such as MobileViT, TinyViT, or EfficientFormer, which are specifically designed for efficient edge deployment.

Although this study focused on crack, stringing, and no defects arising from the 3D printer, the CNN-enhanced ViT model can be extended to handle a wide range of defect types, such as porosity and layer shifting, as well as defects arising in other 3D printing methods, such as selective laser melting (SLM) or stereolithography (SLA). By utilizing the CNN-enhanced ViT across diverse datasets, the CNN-enhanced ViT should maintain its performance in preserving local features despite the global context modelling.

As the model has shown good performance when tested offline, it is suggested to test the developed model in real-time application to measure the robustness and efficiency within real manufacturing data. In addition, this measure is important to confirm the model is suitable for edge deployment as well as to ensure the model is accurate, fast, and reliable outside the controlled experiment settings.

REFERENCES

[1] C. M. S. Vicente, M. Sardinha, L. Reis, A. Ribeiro, and M. Leite, "Large-format additive manufacturing of polymer extrusion-based deposition systems: review and applications," *Progress in Additive Manufacturing,* vol. 8, no. 6, pp. 1257-1280, 2023.

[2] K. Kanishka and B. Acherjee, "Revolutionizing manufacturing: A comprehensive overview of additive manufacturing processes, materials, developments, and challenges," *Journal of Manufacturing Processes,* vol. 107, pp. 574-619, 2023.

[3] V. Kadam, S. Kumar, A. Bongale, S. Wazarkar, P. Kamat, and S. Patil, "Enhancing surface fault detection using machine learning for 3D printed products," *Applied System Innovation,* vol. 4, no. 2, p. 34, 2021.

[4] V. V. Bhandarkar, H. Y. Shahare, A. P. Mall, and P. Tandon, "An overview of traditional and advanced methods to detect part defects in additive manufacturing processes," *Journal of Intelligent Manufacturing,* vol. 36, no. 7, pp. 4411-4446, 2025.

[5] N. A. Othman, N. S. Damanhuri, M. S. Mazalan, S. A. Shamsuddin, M. H. Abbas, and B. C. Meng, "Automated water quality monitoring system development via LabVIEW for aquaculture industry (Tilapia) in Malaysia," *Indones. J. Electr. Eng. Comput. Sci,* vol. 20, no. 2, pp. 805-812, 2020.

[6] N. A. Othman, N. S. Damanhuri, N. M. Ali, B. C. C. Meng, and A. A. Abd Samat, "Plant leaf classification using convolutional neural network," in *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2022, vol. 1: IEEE, pp. 1043-1048.

[7] N. S. M. Sauki *et al.*, "Classification Type of Asynchrony Breathing Image Using 2-Dimensional Convolutional Neural Network," in *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*, 3-6 July 2023 2023, pp. 2281-2286.

[8] J. Qin *et al.*, "Research and application of machine learning for additive manufacturing," *Additive Manufacturing,* vol. 52, p. 102691, 2022.

[9] Y. Zhang *et al.*, "Design of an intelligent grading system for Chinese water chestnuts utilizing advanced artificial intelligence methods," *Engineering Applications of Artificial Intelligence,* vol. 160, p. 112070, 2025.

[10] J. Chen, J. Lei, Z. Chen, and Z. Zeng, "Application of Convolutional Neural Network (CNN) to Optimize Quality Inspection System for Filamentary Materials," *Procedia Computer Science,* vol. 261, pp. 450-457, 2025.

[11] K. Kozhay, S. Turarbek, T. Asselbekova, M. H. Ali, and E. Shehab, "Convolutional Neural Network-Based Defect Detection Technique in FDM Technology," *Procedia Computer Science,* vol. 231, pp. 119-128, 2024.

[12] F. V. Senhora, H. Chi, Y. Zhang, L. Mirabella, T. L. E. Tang, and G. H. Paulino, "Machine learning for topology optimization: Physics-based learning through an independent training strategy," *Computer Methods in Applied Mechanics and Engineering,* vol. 398, p. 115116, 2022.

[13] Y. Hadhoud *et al.*, "From Binary to Multi-Class Classification: A Two-Step Hybrid CNN-ViT Model for Chest Disease Classification Based on X-Ray Images," *Diagnostics,* vol. 14, no. 23, p. 2754, 2024.

[14] M. Zeng, S. Chen, H. Liu, W. Wang, and J. Xie, "HCFormer: A lightweight pest detection model combining CNN and ViT," *Agronomy,* vol. 14, no. 9, p. 1940,

2024.
[15] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929,* 2020.
[16] X. Q. Wang, Z. Jin, B. Zheng, and G. X. Gu, "Transformer-based approach for printing quality recognition in fused filament fabrication," *npj Advanced Manufacturing,* vol. 2, no. 1, p. 15, 2025.
[17] V. Vasan, N. V. Sridharan, S. Vaithiyanathan, and M. Aghaei, "Detection and classification of surface defects on hot-rolled steel using vision transformers," *Heliyon,* vol. 10, no. 19, 2024.
[18] J. Chen, P. Wu, X. Zhang, R. Xu, and J. Liang, "Add-Vit: CNN-Transformer Hybrid Architecture for small data paradigm processing," *Neural Processing Letters,* vol. 56, no. 3, p. 198, 2024.
[19] T. Zhou, Y. Niu, H. Lu, C. Peng, Y. Guo, and H. Zhou, "Vision transformer: To discover the “four secrets” of image patches," *Information Fusion,* vol. 105, p. 102248, 2024.
[20] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in neural information processing systems,* vol. 34, pp. 3965-3977, 2021.
[21] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012-10022.
[22] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357-366.
[23] G. Koppe, A. Meyer-Lindenberg, and D. Durstewitz, "Deep learning for small and big data in psychiatry," *Neuropsychopharmacology,* vol. 46, no. 1, pp. 176-190, 2021.
[24] L. C. Ngugi, M. Abelwahab, and M. Abo-Zahhad, "Recent advances in image processing techniques for automated leaf pest and disease recognition–A review," *Information processing in agriculture,* vol. 8, no. 1, pp. 27-51, 2021.
[25] S. Li, C. Wu, and N. Xiong, "Hybrid architecture based on CNN and transformer for strip steel surface defect classification," *Electronics,* vol. 11, no. 8, p. 1200, 2022.
[26] G. A. Sampedro, D. J. Agron, R. G. Kim, D.-S. Kim, and J.-M. Lee, "Fused deposition modeling 3D printing fault diagnosis using temporal convolutional network," in *2021 1st International Conference in Information and Computing Research (iCORE)*, 2021: IEEE, pp. 62-65.
[27] F. Han, S. Liu, S. Liu, J. Zou, Y. Ai, and C. Xu, "Defect detection: Defect Classification and Localization for Additive Manufacturing using Deep Learning Method," in *2020 21st International Conference on Electronic Packaging Technology (ICEPT)*, 12-15 Aug. 2020 2020, pp. 1-4, doi: 10.1109/ICEPT50128.2020.9202566.
[28] X. Zheng, S. Zheng, Y. Kong, and J. Chen, "Recent advances in surface defect inspection of industrial products using deep learning techniques," *The International Journal of Advanced Manufacturing Technology,* vol. 113, pp. 35-58, 2021.
[29] J. Schwerdtfeger, R. F. Singer, and C. Körner, "In situ flaw detection by IR-imaging during electron beam melting," *Rapid Prototyping Journal,* vol. 18, no. 4, pp. 259-263, 2012.
[30] P. M. Bhatt *et al.*, "Image-Based Surface Defect Detection Using Deep

Learning: A Review," *Journal of Computing and Information Science in Engineering,* vol. 21, no. 4, 2021.
[31] A. Kanadath, J. A. A. Jothi, and S. Urolagin, "CViTS-Net: A CNN-ViT Network with Skip Connections for Histopathology Image Classification," *IEEE Access,* 2024.
[32] M. R. Hasan, "Transformer and Convolutional Neural Network Based Hybrid Approaches in Medical Image Classification, Caption Generation, and Retrieval Processes," Morgan State University, 2024.
[33] C. W. J. Lim, K. Q. Le, Q. Lu, and C. H. Wong, "An Overview of 3-D Printing in Manufacturing, Aerospace, and Automotive Industries," *IEEE Potentials,* vol. 35, no. 4, pp. 18-22, 2016.
[34] T. Vaneker, A. Bernard, G. Moroni, I. Gibson, and Y. Zhang, "Design for additive manufacturing: Framework and methodology," *CIRP Annals,* vol. 69, no. 2, pp. 578-599, 2020.
[35] M.-A. de Pastre, Y. Quinsat, and C. Lartigue, "Effects of additive manufacturing processes on part defects and properties: A classification review," *International Journal on Interactive Design and Manufacturing (IJIDeM),* vol. 16, no. 4, pp. 1471-1496, 2022.
[36] A. P. Fagundes, J. O. de Brito Lira, N. Padoin, C. Soares, and H. G. Riella, "Additive manufacturing of functional devices for environmental applications: A review," *Journal of Environmental Chemical Engineering,* vol. 10, no. 3, p. 108049, 2022.
[37] T. Miki, "Self-support topology optimization considering distortion for metal additive manufacturing," *Computer Methods in Applied Mechanics and Engineering,* vol. 404, p. 115821, 2023.
[38] P. Lakkala, S. R. Munnangi, S. Bandari, and M. Repka, "Additive manufacturing technologies with emphasis on stereolithography 3D printing in pharmaceutical and medical applications: A review," *International journal of pharmaceutics: X,* vol. 5, p. 100159, 2023.
[39] N. B. A. Karna, M. A. P. Putra, S. M. Rachmawati, M. Abisado, and G. A. Sampedro, "Toward Accurate Fused Deposition Modeling 3D Printer Fault Detection Using Improved YOLOv8 With Hyperparameter Optimization,", *IEEE Access,* Article vol. 11, pp. 74251-74262, 2023, doi: 10.1109/access.2023.3293056.
[40] N. Dhakal, X. Wang, C. Espejo, A. Morina, and N. Emami, "Impact of processing defects on microstructure, surface quality, and tribological performance in 3D printed polymers," *journal of materials research and technology,* vol. 23, pp. 1252-1272, 2023.
[41] N. Jyeniskhan, A. Keutayeva, G. Kazbek, M. H. Ali, and E. Shehab, "Integrating Machine Learning Model and Digital Twin System for Additive Manufacturing," *IEEE Access,* vol. 11, pp. 71113-71126, 2023.
[42] M. Moradi *et al.*, "Correlation between infill percentages, layer width, and mechanical properties in fused deposition modelling of poly-lactic acid 3D printing," *Machines,* vol. 11, no. 10, p. 950, 2023.
[43] R. Ranjan, Z. Chen, C. Ayas, M. Langelaar, and F. Van Keulen, "Overheating control in additive manufacturing using a 3D topology optimization method and experimental validation," *Additive Manufacturing,* vol. 61, p. 103339, 2023.
[44] M. Galati *et al.*, "A methodology for evaluating the aesthetic quality of 3D printed parts," *Procedia CIRP,* vol. 79, pp. 95-100, 2019.
[45] C. G. Amza, A. Zapciu, G. Constantin, F. Baciu, and M. I. Vasile, "Enhancing

mechanical properties of polymer 3D printed parts," *Polymers,* vol. 13, no. 4, p. 562, 2021.
[46] S. Amithesh, B. Shanmugasundaram, S. Kamath, S. Adhithyan, and R. Murugan, "Analysis of dimensional quality in FDM printed Nylon 6 parts," *Progress in additive manufacturing,* vol. 9, no. 4, pp. 1225-1238, 2024.
[47] Z. Zhang, I. Fidan, and M. Allen, "Detection of material extrusion in-process failures via deep learning," *Inventions,* vol. 5, no. 3, p. 25, 2020.
[48] G. Cicala, D. Giordano, C. Tosto, G. Filippone, A. Recca, and I. Blanco, "Polylactide (PLA) filaments a biobased solution for additive manufacturing: Correlating rheology and thermomechanical properties with printing quality," *Materials,* vol. 11, no. 7, p. 1191, 2018.
[49] L. Rettenberger, N. Beyer, I. Sieber, and M. Reischl, "Fault detection in 3D-printing with deep learning," in *2024 IEEE International Conference on Consumer Electronics (ICCE)*, 2024: IEEE, pp. 1-4.
[50] G. Lipkowitz, I. Coates, N. Krishna, E. S. Shaqfeh, and J. M. DeSimone, "Methods for modeling and real-time visualization of CLIP and iCLIP-based 3D printing," *Giant,* vol. 17, p. 100239, 2024.
[51] Y. Chen, X. Peng, L. Kong, G. Dong, A. Remani, and R. Leach, "Defect inspection technologies for additive manufacturing," *International Journal of Extreme Manufacturing,* vol. 3, no. 2, p. 022002, 2021.
[52] K. Imoto, T. Nakai, T. Ike, K. Haruki, and Y. Sato, "A CNN-Based Transfer Learning Method for Defect Classification in Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing,* vol. 32, no. 4, pp. 455-459, 2019.
[53] S. Mandapaka, C. Diaz, H. Irisson, A. Akundi, V. Lopez, and D. Timmer, "Application of Automated Quality Control in Smart Factories - A Deep Learning-based Approach," in *2023 IEEE International Systems Conference (SysCon)*, 17-20 April 2023 2023, pp. 1-8.
[54] S. Dorafshan, R. J. Thomas, and M. Maguire, "Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete," *Construction and Building Materials,* vol. 186, pp. 1031-1045, 2018.
[55] F. Rezaei, H. Izadi, H. Memarian, and M. Baniassadi, "The effectiveness of different thresholding techniques in segmenting micro CT images of porous carbonates to estimate porosity," *Journal of Petroleum Science and Engineering,* vol. 177, pp. 518-527, 2019.
[56] Z. Zhenggan and S. Guangkai, "New progress of the study and application of advanced ultrasonic testing technology," *Journal of Mechanical Engineering,* vol. 53, no. 22, pp. 1-10, 2017.
[57] C. Millon, A. Vanhoye, A.-F. Obaton, and J.-D. Penot, "Development of laser ultrasonics inspection for online monitoring of additive manufacturing," *Welding in the World,* vol. 62, pp. 653-661, 2018.
[58] I. S. Ramírez, F. P. G. Márquez, and M. Papaelias, "Review on additive manufacturing and non-destructive testing," *Journal of Manufacturing Systems,* vol. 66, pp. 260-286, 2023.
[59] M. Y. Khalid *et al.*, "Developments in chemical treatments, manufacturing techniques and potential applications of natural-fibers-based biodegradable composites," *Coatings,* vol. 11, no. 3, p. 293, 2021.
[60] J. M. Waller, R. L. Saulsberry, B. H. Parker, K. L. Hodges, E. R. Burke, and K. M. Taminger, "Summary of NDE of additive manufacturing efforts in NASA,"

in *AIP Conference Proceedings*, 2015, vol. 1650, no. 1: American Institute of Physics, pp. 51-62.
[61] S. K. Dwivedi, M. Vishwakarma, and A. Soni, "Advances and researches on non destructive testing: A review," *Materials Today: Proceedings,* vol. 5, no. 2, pp. 3690-3698, 2018.
[62] R. Usamentiaga, P. Venegas, J. Guerediaga, L. Vega, J. Molleda, and F. G. Bulnes, "Infrared thermography for temperature measurement and non-destructive testing," *Sensors,* vol. 14, no. 7, pp. 12305-12348, 2014.
[63] E. I. Todorov, "Measurement of electromagnetic properties of powder and solid metal materials for additive manufacturing," in *Nondestructive Characterization and Monitoring of Advanced Materials, Aerospace, and Civil Infrastructure 2017*, 2017, vol. 10169: SPIE, pp. 44-54.
[64] J. C. Muñoz, F. G. Márquez, and M. Papaelias, "Railroad inspection based on ACFM employing a non-uniform B-spline approach," *Mechanical Systems and Signal Processing,* vol. 40, no. 2, pp. 605-617, 2013.
[65] M. Geľatko, M. Hatala, F. Botko, R. Vandžura, and J. Hajnyš, "Eddy current testing of artificial defects in 316L stainless steel samples made by additive manufacturing technology," *Materials,* vol. 15, no. 19, p. 6783, 2022.
[66] S. Inayathullah and R. Buddala, "Review of machine learning applications in additive manufacturing," *Results in Engineering,* p. 103676, 2024.
[67] H. Baumgartl, J. Tomas, R. Buettner, and M. Merkel, "A deep learning-based model for defect detection in laser-powder bed fusion using in-situ thermographic monitoring," *Progress in Additive Manufacturing,* vol. 5, no. 3, pp. 277-285, 2020.
[68] K. Chen *et al.*, "A review of machine learning in additive manufacturing: Design and process," *The International Journal of Advanced Manufacturing Technology,* vol. 135, no. 3, pp. 1051-1087, 2024.
[69] S. I. Kampezidou, A. Tikayat Ray, A. P. Bhat, O. J. Pinon Fischer, and D. N. Mavris, "Fundamental components and principles of supervised machine learning workflows with numerical and categorical data," *Eng,* vol. 5, no. 1, pp. 384-416, 2024.
[70] S.-H. Park, S. Choi, and K.-Y. Jhang, "Porosity evaluation of additively manufactured components using deep learning-based ultrasonic nondestructive testing," *International Journal of Precision Engineering and Manufacturing-Green Technology,* vol. 9, no. 2, pp. 395-407, 2022.
[71] E. Westphal and H. Seitz, "A machine learning method for defect detection and visualization in selective laser sintering based on convolutional neural networks," *Additive Manufacturing,* vol. 41, p. 101965, 2021.
[72] Z. Yang, Y. Lu, H. Yeung, and S. Krishnamurty, "Investigation of deep learning for real-time melt pool classification in additive manufacturing," in *2019 IEEE 15th international conference on automation science and engineering (CASE)*, 2019: IEEE, pp. 640-647.
[73] B. Zhang, S. Liu, and Y. C. Shin, "In-Process monitoring of porosity during laser additive manufacturing process," *Additive Manufacturing,* vol. 28, pp. 497-505, 2019.
[74] M. Singh, P. Sharma, S. K. Sharma, and J. Singh, "A novel real-time quality control system for 3D printing: A deep learning approach using data efficient image transformers," *Expert Systems with Applications,* vol. 273, p. 126863, 2025.
[75] C. Uysalel, J. Cotrina, and M. Ghazinejad, "Automated defect characterization

and physics-based performance modeling of additively manufactured metals with vision transformers," *MRS Advances,* pp. 1-7, 2025.
[76] H. Song, C. Li, Y. Fu, R. Li, H. Zhang, and G. Wang, "A two-stage unsupervised approach for surface anomaly detection in wire and arc additive manufacturing," *Computers in Industry,* vol. 151, p. 103994, 2023.
[77] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, "Unsupervised learning based on artificial neural network: A review," in *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, 2018: IEEE, pp. 322-327.
[78] K. Taherkhani, C. Eischer, and E. Toyserkani, "An unsupervised machine learning algorithm for in-situ defect-detection in laser powder-bed fusion," *Journal of Manufacturing Processes,* vol. 81, pp. 476-489, 2022.
[79] S. Manivannan, "Automatic quality inspection in additive manufacturing using semi-supervised deep learning," *Journal of Intelligent Manufacturing,* vol. 34, no. 7, pp. 3091-3108, 2023.
[80] M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah, "In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning," *arXiv:2101.06329,* 2021.
[81] G. Mattera, J. Polden, A. Caggiano, L. Nele, Z. Pan, and J. Norrish, "Semi-supervised learning for real-time anomaly detection in pulsed transfer wire arc additive manufacturing," *Journal of Manufacturing Processes,* vol. 128, pp. 84-97, 2024.
[82] H. Brooks and N. Tucker, "Electrospinning predictions using artificial neural networks," *Polymer,* vol. 58, pp. 22-29, 2015.
[83] I. Baturynska and K. Martinsen, "Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms," *Journal of Intelligent Manufacturing,* vol. 32, no. 1, pp. 179-200, 2021.
[84] G. Tapia, A. H. Elwany, and H. Sang, "Prediction of porosity in metal-based additive manufacturing using spatial Gaussian process models," *Additive Manufacturing,* vol. 12, pp. 282-290, 2016.
[85] V. García, J. S. Sánchez, L. A. Rodríguez-Picón, L. C. Méndez-González, and H. d. J. Ochoa-Domínguez, "Using regression models for predicting the product quality in a tubing extrusion process," *Journal of Intelligent Manufacturing,* vol. 30, no. 6, pp. 2535-2544, 2019.
[86] Y. Chen, L. Zhao, and T. Han, "Uncertainty-guided Bayesian active learning for cost-effective fault diagnosis with minimal labeled data," *Engineering Applications of Artificial Intelligence,* vol. 160, p. 111945, 2025.
[87] R. d. S. B. Ferreira, A. Sabbaghi, and Q. Huang, "Automated geometric shape deviation modeling for additive manufacturing systems via Bayesian neural networks," *IEEE Transactions on Automation Science and Engineering,* vol. 17, no. 2, pp. 584-598, 2019.
[88] M. Aminzadeh and T. R. Kurfess, "Online quality inspection using Bayesian classification in powder-bed additive manufacturing from high-resolution visual camera images," *Journal of Intelligent Manufacturing,* vol. 30, no. 6, pp. 2505-2523, 2019.
[89] N. Hertlein, S. Deshpande, V. Venugopal, M. Kumar, and S. Anand, "Prediction of selective laser melting part quality using hybrid Bayesian network," *Additive Manufacturing,* vol. 32, p. 101089, 2020.
[90] Y. Gui, K. Aoyagi, H. Bian, and A. Chiba, "Detection, classification and prediction of internal defects from surface morphology data of metal parts

fabricated by powder bed fusion type additive manufacturing using an electron beam," *Additive Manufacturing,* vol. 54, p. 102736, 2022.
[91] T. Liu, L. Huang, and B. Chen, "Real-time defect detection of laser additive manufacturing based on support vector machine," in *Journal of Physics: Conference Series*, 2019, vol. 1213, no. 5: IOP Publishing, p. 052043.
[92] H. Wang, B. Li, and F.-Z. Xuan, "Acoustic emission for in situ process monitoring of selective laser melting additive manufacturing based on machine learning and improved variational modal decomposition," *The International Journal of Advanced Manufacturing Technology,* vol. 122, no. 5, pp. 2277-2292, 2022.
[93] M. Cheepu, "Machine learning approach for the prediction of defect characteristics in wire arc additive manufacturing," *Transactions of the Indian Institute of Metals,* vol. 76, no. 2, pp. 447-455, 2023.
[94] M. Trovato and P. Cicconi, "A Decision Tree approach for an early evaluation of 3D models in Design for Additive Manufacturing," *Procedia CIRP,* vol. 128, pp. 96-101, 2024.
[95] U. Aziz, A. Bradshaw, J. Lim, and M. Thomas, "Classification of defects in additively manufactured nickel alloys using supervised machine learning," *Materials Science and Technology,* vol. 39, no. 16, pp. 2464-2468, 2023.
[96] S. A. Singh and K. A. Desai, "Automated surface defect detection framework using machine vision and convolutional neural networks," *Journal of Intelligent Manufacturing,* vol. 34, no. 4, pp. 1995-2011, 2023.
[97] A. Ma'arif, W. Rahmaniar, H. I. K. Fathurrahman, and A. Z. K. Frisky, "Understanding of Convolutional Neural Network (CNN): A Review," *International Journal of Robotics & Control Systems,* vol. 2, no. 4, 2022.
[98] T. Mowbray, "A Survey of Deep Learning Architectures in Modern Machine Learning Systems: From CNNs to Transformers," *Journal of Computer Technology and Software,* vol. 4, no. 8, 2025.
[99] M. T. Fuad *et al.*, *Recent Advances in Deep Learning Techniques for Face Recognition*. 2021.
[100] K. Gao, F. Han, P. Dong, N. Xiong, and R. Du, "Connected vehicle as a mobile sensor for real time queue length at signalized intersections," *Sensors,* vol. 19, no. 9, p. 2059, 2019.
[101] H. Cheng, Z. Xie, Y. Shi, and N. Xiong, "Multi-step data prediction in wireless sensor networks based on one-dimensional CNN and bidirectional LSTM," *IEEE Access,* vol. 7, pp. 117883-117896, 2019.
[102] N. Lethanh, T. A. Trinh, and M. T. Hossain, "An Investigation on Prediction of Infrastructure Asset Defect with CNN and ViT Algorithms," *Infrastructures,* vol. 10, no. 5, p. 125, 2025.
[103] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems,* vol. 25, 2012.
[104] M. Jeon and Y.-S. Jeong, "Compact and accurate scene text detector," *Applied Sciences,* vol. 10, no. 6, p. 2096, 2020.
[105] T. Vu, C. Van Nguyen, T. X. Pham, T. M. Luu, and C. D. Yoo, "Fast and efficient image quality enhancement via desubpixel convolutional neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0-0.
[106] A. Zeynali, M. A. Tinati, and B. M. Tazehkand, "Hybrid cnn-transformer architecture with xception-based feature enhancement for accurate breast cancer

classification," *IEEE Access,* 2024.
[107] I. Pacal, B. Ozdemir, J. Zeynalov, H. Gasimov, and N. Pacal, "A novel CNN-ViT-based deep learning model for early skin cancer diagnosis," *Biomedical Signal Processing and Control,* vol. 104, p. 107627, 2025.
[108] C. Hu, N. Cao, H. Zhou, and B. Guo, "Medical Image Classification with a Hybrid SSM Model Based on CNN and Transformer," *Electronics,* vol. 13, no. 15, p. 3094, 2024.
[109] J. W. Kim, A. U. Khan, and I. Banerjee, "Systematic Review of Hybrid Vision Transformer Architectures for Radiological Image Analysis," *medRxiv,* p. 2024.06. 21.24309265, 2024.
[110] H. Yunusa, S. Qin, A. H. A. Chukkol, A. A. Yusuf, I. Bello, and A. Lawan, "Exploring the synergies of hybrid CNNs and ViTs architectures for computer vision: A survey," *arXiv :2402.02941,* 2024.
[111] R. Tavasoli, A. VarastehNezhad, and H. Farbeh, "Hybrid Vision Transformer for Detection of Dentigerous Cysts in Dental Radiography Images," in *2024 14th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2024: IEEE, pp. 143-148.
[112] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.
[113] W. Huang, Y. Deng, S. Hui, Y. Wu, S. Zhou, and J. Wang, "Sparse self-attention transformer for image inpainting," *Pattern Recognition,* vol. 145, p. 109897, 2024.
[114] X. Zhang, "A Hybrid CNN-Transformer Architecture for Image Classification on Small Datasets," in *2025 5th International Conference on Sensors and Information Technology*, 2025: IEEE, pp. 324-331.
[115] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences,* vol. 13, no. 9, p. 5521, 2023.
[116] S. Wang, Q. Zeng, W. Ni, C. Cheng, and Y. Wang, "ODP-Transformer: Interpretation of pest classification results using image caption generation techniques," *Computers and Electronics in Agriculture,* vol. 209, p. 107863, 2023.
[117] B. Zhan, M. Li, W. Luo, P. Li, X. Li, and H. Zhang, "Study on the tea pest classification model using a convolutional and embedded iterative region of interest encoding transformer," *Biology,* vol. 12, no. 7, p. 1017, 2023.
[118] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10231-10241.
[119] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling," *Current Oncology,* vol. 29, no. 10, pp. 7498-7511, 2022.
[120] Y. Zheng and W. Jiang, "Evaluation of vision transformers for traffic sign classification," *Wireless Communications and Mobile Computing,* vol. 2022, no. 1, p. 3041117, 2022.
[121] V. Singla, S. Bawa, and J. Singh, "Enhancing Indian sign language recognition through data augmentation and visual transformer," *Neural Computing and Applications,* vol. 36, no. 24, pp. 15103-15116, 2024.
[122] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, "Scalable vision transformers with hierarchical pooling," in *Proceedings of the IEEE/cvf international conference*

on computer vision, 2021, pp. 377-386.
[123] B. Chen *et al.*, "Psvit: Better vision transformer via token pooling and attention sharing," *arXiv:2108.03428,* 2021.
[124] A. Diko, D. Avola, M. Cascio, and L. Cinque, "ReViT: Enhancing vision transformers feature diversity with attention residual connections," *Pattern Recognition,* vol. 156, p. 110853, 2024.
[125] Y. Liu and T. Yan, "Vision transformer enhanced with convolutional attention and graph convolution for semantic segmentation," *Image and Vision Computing,* vol. 161, p. 105633, 2025.
[126] K. Alomar, H. I. Aysel, and X. Cai, "RNNs, CNNs and Transformers in Human Action Recognition: A Survey and A Hybrid Model," *arXiv:2407.06162,* 2024.
[127] D. Ashoka, "Effivit: Hybrid cnn-transformer for retinal imaging," *Computers in Biology and Medicine,* vol. 191, p. 110164, 2025.
[128] J. M. Ahn, W. Jeong, H. Lee, and K. Kim, "Hyperspectral imaging-based land use classification using a hybrid Convolutional Neural network-vision transformer model," *ENVIRONMENTAL TECHNOLOGY & INNOVATION,* vol. 39, 2025.
[129] M. I. Rahman and A. Rahman, "AACNN-ViT: Adaptive Attention-Augmented Convolutional and Vision Transformer Fusion for Lung Cancer Detection," *Journal of Imaging,* vol. 12, no. 2, p. 62, 2026.
[130] S. Yi, Y. Ren, and D. Shao, "CTANet: A hybrid CNN and ViT network with adaptive focus fusion for diabetic retinopathy grading," *Biomedical Signal Processing and Control,* vol. 113, p. 109106, 2026.
[131] T. Lalitha, N. Anushkannan, S. Shreepad, S. Sasireka, H. Anandaram, and S. Razia, "Deep learning-based automatic 3D printer anomaly detection during the printing process," in *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, 2022: IEEE, pp. 1343-1348.
[132] K. Paraskevoudis, P. Karayannis, and E. P. Koumoulos, "Real-Time 3D Printing Remote Defect Detection (Stringing) with Computer Vision and Artificial Intelligence," *Processes,* vol. 8, no. 11, p. 1464, 2020.
[133] K. Alomar, H. I. Aysel, and X. Cai, "Data augmentation in classification and segmentation: A survey and new strategies," *Journal of Imaging,* vol. 9, no. 2, p. 46, 2023.
[134] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," *Advances in neural information processing systems,* vol. 27, 2014.
[135] Z. Dai, G. Lai, Y. Yang, and Q. Le, "Funnel-transformer: Filtering out sequential redundancy for efficient language processing," *Advances in neural information processing systems,* vol. 33, pp. 4271-4282, 2020.
[136] P. Lattanzi, T. Mendo, A. Galdelli, and A. N. Tassetti, "Machine learning-based prediction of passive gears from vessel tracking data in small-scale multi-gear fisheries," *Ecological Informatics,* p. 103670, 2026.
[137] N. S. a. M. Sauki *et al.*, "Convolutional long short-term memory neural network integrated with classifier in classifying type of asynchrony breathing in mechanically ventilated patients," *Computer Methods and Programs in Biomedicine,* vol. 263, p. 108680, 2025.
[138] Z. N. Khudhair *et al.*, "Color to Grayscale Image Conversion Based on Singular Value Decomposition," *IEEE Access,* vol. 11, pp. 54629-54638, 2023.
[139] S. Rizwana, R. Das, and V. Bhateja, "HVS-Based Modified Nonlinear Unsharp Masking Framework for Mammographic Image Enhancement Using Bilateral

Filter," *Arabian Journal for Science and Engineering,* vol. 50, no. 15, pp. 12337-12355, 2025/08/01 2025, doi: 10.1007/s13369-024-09726-8.

[140] A. Benyahia, A. Benammar, A. B. Azzououm, I. Karmal, A. Belaout, and I. E. Araar, "Impact of Image Resolution on Convolutional Neural Networks for Alzheimer Disease Dataset Classification: A ResNet Study," in *2024 10th International Conference on Computing, Engineering and Design (ICCED)*, 2024: IEEE, pp. 1-6.

[141] L. Brigato and L. Iocchi, "A close look at deep learning with small data," in *2020 25th international conference on pattern recognition (ICPR)*, 2021: IEEE, pp. 2490-2497.

[142] K. Jiang, P. Peng, Y. Lian, and W. Xu, "The encoding method of position embeddings in vision transformer," *Journal of Visual Communication and Image Representation,* vol. 89, p. 103664, 2022.

[143] G. Wang *et al.*, "P 2fevit: Plug-and-play cnn feature embedded hybrid vision transformer for remote sensing image classification," *Remote Sensing,* vol. 15, no. 7, p. 1773, 2023.

[144] M. Fahmy Amin, "Confusion matrix in three-class classification problems: A step-by-step tutorial," *Journal of Engineering Research,* vol. 7, no. 1, pp. 0-0, 2023.

[145] K. Song and Y. Yan, "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects," *Applied Surface Science,* vol. 285, pp. 858-864, 2013.

[146] K. Wang, B. Suo, W. Qian, G. Zhang, T. Wang, and Y. Yang, "Small object detection by synchronous attention swin transformer with channel granularity adaptive mechanism," *International Journal of Computational Intelligence Systems,* vol. 18, no. 1, p. 255, 2025.

AUTHOR'S PROFILE

Nur Najiha Binti Kamarulzaman obtained Bachelor of Engineering (Hons.) Electric and Electronic Engineering in 2024 from University Teknologi MARA (UiTM) Cawangan Pulau Pinang. Her research interests include the deep learning and image processing.

LIST OF PUBLICATIONS:

Kamarulzaman N. N., N. S. Damanhuri, N. A. M. Salleh, N. A. Othman, B. C. C. Meng, and A. N. Handayani, "Defect Classification of Additive Manufacturing Product using Deep Learning Method," in *2025 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, 2025, vol. 1, pp. 396-401.

In review:

Kamarulzaman N. N., N. S. Damanhuri, N. H. M. Asri, N. A. Othman, N. A. M. Salleh, B. C. C. Meng, and A. N. Handayani, "Automated System for Defect Classification Image of a 3D-printed Additive-manufactured Product", International Journal of Automotive and Mechanical Engineering (IJAME).
-Accepted-will be published June 2026

Kamarulzaman N. N., N. S. Damanhuri, N. A. M. Salleh, N. A. Othman, B. C. C. Meng, and A. N. Handayani, “Application of machine learning models in defect classification of additive manufacturing product – A review”, ESTEEM Academic Journal.

Kamarulzaman N. N., N. S. Damanhuri, N. A. M. Salleh, N. A. Othman, B. C. C. Meng, N. S. M. Sauki and A. N. Handayani, “Defect Classification of Additive Manufacturing Product using CNN-Enhanced ViT”, IFAC Journal of Systems and Control.