Detection of Counterfeit Coins Using Multimodal GPT-4 and Vision Transformer

Dina Omidvar Tehrani

A Thesis
in
The Department
of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Master of Computer Science at
Concordia University
Montréal, Québec, Canada

September 2024

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **Dina Omidvar Tehrani**

Entitled: **Detection of Counterfeit Coins Using Multimodal GPT-4 and Vision Transformer**

and submitted in partial fulfillment of the requirements for the degree of

Master of Computer Science

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Jinqiu Yang

_____ Examiner
Dr. Jinqiu Yang

_____ Examiner
Dr. Yuhong Yan

_____ Thesis Supervisor
Dr. Ching Y . Suen

Approved by _____
Dr. Charalambos Poullis, Graduate Program Director

September 20, 2024 _____
Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Detection of Counterfeit Coins Using Multimodal GPT-4 and Vision Transformer

Dina Omidvar Tehrani

The proliferation of counterfeit coins poses a substantial threat to the integrity of monetary systems and the stability of financial markets. Advanced counterfeiting techniques allow these fraudulent coins to closely mimic genuine ones, complicating the detection process and necessitating robust methods capable of discerning minute differences between genuine and fake coins. This thesis addresses the problem of counterfeit coin detection by introducing a diverse dataset comprising high-resolution images of both Danish and Chinese coins, categorized into genuine and counterfeit sets across multiple years.

To tackle the detection task, we employ two advanced approaches: a Vision Transformer (ViT) model and a multimodal GPT-4 model. The ViT model leverages its self-attention mechanisms to capture intricate patterns and details within the coin images, while the GPT-4 model integrates both visual and textual data, utilizing various prompting techniques to enhance its performance. Our results show that the ViT model outperforms previous methods and the state-of-the-art in terms of accuracy and robustness, achieving a remarkable 99.31% accuracy. The GPT-4 model, although primarily designed for natural language processing, demonstrates promising capabilities in counterfeit detection, particularly with advanced prompting strategies like Chain-of-Thought and Generated Knowledge.

This research advances the current state-of-the-art in counterfeit coin detection and highlights the potential of few-shot learning and transfer learning in achieving high accuracy with limited training data.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Ching Yee Suen, whose support and guidance were instrumental in the completion of this thesis. His insightful advice and encouragement have been invaluable throughout this research journey. It has been an honor to work under his mentorship, and I am deeply grateful for the opportunities and learning experiences he has provided.

I also extend my thanks to my colleagues and friends at CENPARMI at Concordia University, whose collaboration and motivation have significantly contributed to this work. Special thanks go to Mr. Nicola Nobile, for their excellent technical support and assistance.

I am also grateful to my thesis committee members for their time and effort in evaluating this work. Their constructive feedback and suggestions are invaluable and highly appreciated.

Lastly, I want to express my heartfelt appreciation to my family, whose unwavering support and encouragement have been a constant source of strength. Their belief in my abilities has inspired me to persevere and achieve my goals.

Table of Content

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The detection of counterfeit coins is a critical issue for maintaining the integrity of monetary systems and ensuring the stability of financial markets. The proliferation of counterfeit currency, including coins, can lead to significant economic disruptions. It devalues genuine currency, contributes to inflation, and can potentially destabilize entire economic systems. On the other hand, as counterfeiters increasingly leverage advanced technology, the production of fake coins has become more sophisticated, posing significant challenges for detection. These challenges demand more advanced and robust methods to distinguish genuine coins from counterfeits.

Figure 1: Comparison of a genuine Canadian two-dollar coin (left) with a counterfeit version (right) [1], highlighting the subtle differences that can make detection challenging without advanced methods.

Historically, counterfeit coins have been a persistent problem. However, recent technological advancements have significantly exacerbated the issue. Criminal organizations now have access to high-tech equipment and materials that enable them to produce counterfeit coins with an alarming degree of accuracy. A recent operation in Quebec and Ontario [1], for example, uncovered thousands of fake toonies (Fig. 1), underscoring the ongoing urgency of this issue. Canadian authorities have reported that advances in counterfeiting methods are making it increasingly easier for criminals to produce high-quality fakes. This national threat underscores the necessity for continuous improvement in detection technologies to keep pace with the evolving methods of counterfeiters. Given these challenges, developing effective counterfeit coin detection methods is of paramount importance. This thesis aims to contribute to this field by introducing advanced detection techniques using Vision Transformer (ViT) models [14, 15] and multimodal approaches with GPT-4 [4]. By leveraging high-resolution datasets and state-of-the-art machine learning algorithms, this research seeks to enhance the accuracy and robustness of counterfeit coin detection, thereby supporting the broader goal of securing financial systems against the threat of counterfeiting.

1.2 Objectives

The primary objective of this research is to develop and validate robust methodologies for detecting counterfeit coins by using advanced machine learning models that achieve high accuracy with minimal data. This goal addresses the critical issue of data scarcity [7], a common challenge in counterfeit detection, where obtaining large, diverse datasets can be difficult. Traditional approaches often require extensive datasets to train machine learning models effectively. However, in many real-world scenarios, such comprehensive datasets are not available, highlighting the need for models that can perform well even with limited training data. To bridge this gap, this research explores few-shot learning

techniques [12, 16, 19], enabling models to generalize from a small number of examples. In parallel, this research also aims to develop and evaluate a representative dataset comprising high-resolution images of both genuine and counterfeit coins. Ensuring that this dataset is balanced between the two categories is crucial for fairness and representativeness in the training process. High-quality images are essential for allowing the models to extract intricate details and critical information necessary for distinguishing between genuine and counterfeit coins. To achieve these objectives, two approaches are investigated: a Vision Transformer model, known for its ability to capture intricate patterns and contextual information through self-attention mechanisms, and a multimodal GPT-4 model, which integrates visual and textual data along with various prompting techniques to enhance performance.

Through rigorous experimentation and comparative analysis, this research aims to assess the effectiveness of these models in detecting counterfeit coins, highlighting their advantages over traditional methods. We want to contribute to the field by providing innovative solutions that improve counterfeit coin detection, ensuring the security and reliability of financial systems, even in the face of limited data availability. By addressing the challenges of data scarcity and leveraging state-of-the-art machine learning techniques, this research aspires to set new benchmarks in counterfeit coin detection, offering practical and scalable solutions for real-world applications.

1.3 Challenges

As mentioned earlier, counterfeit coin detection is essential for maintaining the integrity of monetary systems and ensuring the stability of financial markets. However, the detection process is fraught with several challenges that complicate the identification of counterfeit coins. These challenges arise from the physical characteristics of coins, the variability in their design, and the practical difficulties in acquiring a comprehensive dataset for training

detection models. Addressing these challenges is crucial for developing robust and reliable counterfeit detection methods. Below are some of the key challenges encountered in this research:

1. **Durability and Degradation:** Despite being made of durable alloys, coins are susceptible to corrosion and rust over time, particularly ancient ones. This degradation can significantly alter the physical appearance of coins, impacting features such as edges and inscriptions. Such alterations make it challenging to analyze their images accurately. Although this research does not primarily focus on edge-based features, the presence of damaged coins remains a significant challenge. The advanced pattern recognition capabilities of the Vision Transformer model help mitigate some of these issues, but variability due to coin degradation persists as a hurdle.

2. **Diversity in Coin Design:** A robust detection method must recognize a wide variety of coin types, which differ in size, design, characters, and the language of inscriptions. This diversity adds complexity to the detection process, as the models must generalize well across different coin types. The proposed methods employed in this research are designed to handle such variability, leveraging their advanced architectures to adapt to the diverse features and characteristics present in coins from various origins.

3. **Illumination Variation:** Shiny coins can exhibit significant glare and reflection, distorting the features captured in images. Consistent lighting conditions during the scanning process are crucial to minimize these effects. In this research, high-resolution imaging and controlled illumination conditions are used to enhance the quality and consistency of the dataset, thereby improving the accuracy of the detection models.

4. **Data Scarcity:** Access to counterfeit coins and their images is restricted due to security concerns, making it difficult to obtain a sufficiently large and diverse dataset for training machine learning models. This scarcity is a significant challenge in counterfeit coin detection research. To address this issue, the research employs few-shot learning techniques, enabling models to generalize from a limited number of examples. By maximizing detection accuracy with minimal data, this approach addresses one of the most critical challenges in this field.

1.4 Contributions

This research makes several significant contributions to the field of counterfeit coin detection, employing advanced machine learning techniques to overcome challenges related to data scarcity, coin diversity, and imaging inconsistencies. The key contributions of this thesis are outlined below:

1. **Development of a Dataset of Danish and Chinese Coins:** A major contribution of this research is the creation of a high-resolution image dataset comprising both genuine and counterfeit Danish and Chinese coins. This dataset includes coins from various years, ensuring a balanced representation of genuine and counterfeit specimens. The images were captured using a Keyence 3D scanner [5], which provides detailed surface data with a resolution of 0.1 µm. The scanner's rotational capability allows for comprehensive imaging of each coin, capturing intricate details that are crucial for accurate counterfeit detection. The high quality of the images allows the models to extract critical features, enhancing the reliability of the detection process. This dataset serves as a valuable resource for training and evaluating machine learning models in the context of counterfeit coin detection.

2. **Application of Vision Transformer (ViT) Models:** This research implements and

fine-tunes Vision Transformer models for the task of counterfeit coin detection. ViT models are known for their ability to capture intricate patterns and contextual information through self-attention mechanisms. The fine-tuned ViT model in this research achieved a remarkable accuracy of 99.31%, significantly outperforming traditional methods and existing state-of-the-art techniques. This demonstrates the effectiveness of ViT models in handling the complexities of counterfeit coin detection.

3. **Exploration of Multimodal GPT-4 Model:** This thesis explores the capabilities of the multimodal GPT-4 model in the domain of counterfeit coin detection. By integrating visual and textual data and employing various prompting techniques [6], such as Zero-Shot [36], Few-Shot [16], Chain-of-Thought [34, 35], and Generated Knowledge prompting [33], the research evaluates the performance of GPT-4 in discerning genuine coins from counterfeit ones. The findings highlight the potential of multimodal approaches in enhancing counterfeit detection accuracy, despite the model's primary design for natural language processing tasks.

4. **Utilization of Few-Shot Learning Techniques:** Addressing the challenge of data scarcity, this research leverages few-shot learning techniques to enable models to generalize from a small number of training examples. This approach is particularly important in counterfeit coin detection, where access to extensive datasets is often limited due to security concerns. By employing few-shot learning, the models in this research achieve high accuracy even with minimal data, demonstrating the viability of this approach in practical applications.

5. **Comparative Analysis and Validation:** This research conducts a rigorous comparative analysis of the proposed methods against existing state-of-the-art techniques. The analysis investigates the capabilities of the Vision Transformer and GPT-4 models, showcasing their performance in terms of precision, recall, and F-measure across

various coin types and years. This comparative validation highlights the significant contributions of this research to the field of counterfeit coin detection.

1.5 Outline

The remainder of this thesis is organized as follows:

In Chapter 2, a comprehensive literature review is provided, discussing related research and existing studies on counterfeit coin detection, with a specific focus on few-shot learning, transfer learning, and state-of-the-art models, including GPT-4 and Vision Transformers (ViT). The chapter introduces the architecture, capabilities, and functionality of these models, setting the foundation for the methodologies used in this research. Chapter 3 delves into the dataset preparation necessary for this research, detailing the creation and preparation of the dataset utilized for counterfeit coin detection.

The methodologies we utilized and the proposed methods are discussed in detail in Chapter 4 and Chapter 5, followed by results and discussions that showcase the effectiveness of the models. Chapter 4, focuses on the multimodal coin authentication methodology using GPT-4, applying various prompting techniques for coin authentication, and discussing the results achieved. Chapter 5 presents the Vision Transformer-based coin authentication framework. This chapter details the implementation of the Vision Transformer for coin classification and its performance in counterfeit coin detection.

Finally, in Chapter 6, the study concludes with a summary of the primary contributions of this work and outlines potential future research directions.

Chapter 2

Literature Review

2.1 Introduction

Coins have been an essential medium of exchange for goods and services for centuries, playing a crucial role in economic transactions. Despite the rise of digital payment methods and credit cards, coins remain integral to everyday life. They are widely accepted in various settings, including retail stores, gas stations, and public transportation systems. The ubiquitous use of coins underscores their continued importance in modern economies. However, the prevalence of counterfeit coins poses a significant threat to the integrity of monetary systems. Counterfeit coins can have a detrimental impact on the economy by undermining public trust in the currency and causing financial losses. The challenge of detecting counterfeit coins has become increasingly complex with advancements in counterfeiting techniques.

Recent statistics indicate a troubling rise in the circulation of counterfeit coins. According to the Deutsche Bundesbank [8], around 115,900 counterfeit euro coins were detected in German payment transactions in 2023, a significant increase from the 73,400 detected in

2022. This increase highlights ongoing issues with counterfeit currency in Europe, particularly with 2-euro coins being frequently targeted by counterfeiters. Additionally, more than 56,600 counterfeit euro banknotes were identified, demonstrating a broader issue with counterfeit currency across different denominations. Given the critical nature of this issue, governments and financial institutions worldwide have taken significant measures to prevent the production and distribution of counterfeit coins. These efforts include the adoption of advanced technologies for coin production and the support of research initiatives aimed at developing more effective detection methods. For instance, the introduction of sophisticated coin designs and the use of innovative materials and minting techniques are part of the strategy to combat counterfeiting. The detection of counterfeit coins is also of paramount importance for museums and collectors, especially concerning ancient coins, which can have significant historical and monetary value. Museum curators and collectors are increasingly urged to use advanced technologies to authenticate coins and ensure the integrity of their collections. The high stakes involved in the trade of ancient coins necessitate meticulous verification processes to prevent the acquisition of counterfeit items. In summary, the detection of counterfeit coins remains a pressing issue with substantial economic implications. The ongoing advancements in counterfeiting techniques demand equally sophisticated detection methods. This research aims to contribute to the field by exploring innovative machine learning approaches to enhance the accuracy and reliability of counterfeit coin detection, thereby supporting the broader goal of maintaining the integrity of monetary systems.

2.2 Related Works

Counterfeit coins are typically crafted to closely mimic genuine coins, with the intent to deceive. In recent years, numerous studies have focused on identifying counterfeit coins through various detection methods.

Traditionally, research has concentrated on the electromagnetic, frequency, and physical characteristics of coins—techniques commonly employed by vending machines, game machines, and parking meters to authenticate coins. These systems typically operate via electromagnetic mechanisms and authenticate coins using techniques such as X-ray fluorescence. For instance, a patented method proposed in [9] uses an oscillation coil to pass a signal through a coin. The coin's characteristics, when analyzed in discrimination mode, determine if they fall within a predefined reference range of minimum and maximum values, thereby automating the process of separating genuine coins from counterfeit ones. As technology has progressed, researchers have increasingly turned to image-based and machine learning-based methods to enhance counterfeit coin detection. These modern approaches offer capabilities beyond traditional electromagnetic techniques, enabling more sophisticated analyses. Consequently, counterfeit coin detection methods can now be broadly categorized into two main approaches: image-based methods and machine learning-based methods.

Image-Based Methods:

Image-based methods focus on analyzing the visual characteristics of coins, such as their shape, the position of letters and numbers, and other intricate details.

One example of image-based approach is presented in [22], where the researchers propose a method to detect fake coins based on the characteristics of coin images. This approach involves computing the dissimilarity between coin images using local key points identified by the Difference of Gaussians (DOG) detector and described by the Scale-Invariant Feature Transform (SIFT) descriptor. Each comparison between a test image and a predefined image is stored as a vector in dissimilarity space, and a Support Vector Machine (SVM) is then used to classify the coins into genuine or counterfeit categories. Another study,

detailed in [20], describes a method to detect counterfeit two-euro coins using images digitized by an optical mouse sensor. However, this method faces challenges related to coin rotation and vulnerability to distortions. Also, research in [21] focuses on detecting counterfeit Danish coins based on their image characteristics, specifically by analyzing edge features such as width, thickness, and edge counts. However, it relies heavily on a large dataset, which necessitates data augmentation to achieve optimal results.

Sharifi Rad et al. [24] developed a blob detector image-based method for automatically detecting counterfeit coins using fuzzy association rules mining. This method involves preprocessing coin images with a blob detector to extract all relevant features, followed by extracting effective fuzzy rules via fuzzy association rules mining and classifying the coin image data. Despite its sensitivity, the system struggles with degraded coins, often classifying them randomly. In [26], the authors utilize a 3D scanner to extract height and depth information from coin images to differentiate between genuine and counterfeit coins. They convert circular coin images to linear rectangular images using straightening algorithms and process the images to address the issue of shiny surfaces. Another study [25] employs a three-dimensional image-based approach to examine the precipice-borders of coin surfaces and trains an ensemble classification system to extract critical features from the images. While 3D scanning offers resistance to low-quality coins, the lengthy processing time remains a significant drawback.

Machine Learning-Based Methods:

One of the initial applications of deep learning techniques to the problem of counterfeit coin detection involved adapting a pre-trained neural network through transfer learning to assess the features on coins. Hmood and Suen [28] applied an ensemble technique that combined outcomes from three classifiers, providing a more robust and reliable detection system.

Furthermore, the innovative counterfeit detection methodology proposed by Bavandsavadkouhi et al. [27] employs an autoencoder-based technique that is notable for its training exclusively on genuine coins, thus avoiding the need for counterfeit samples. This method leverages anomaly detection, where a trained autoencoder assesses reconstruction errors to identify counterfeit coins. Moreover, in another research, Sharifi Rad et al. [23] developed a novel counterfeit coin detection approach using a Pruned Fuzzy Associative (PrFA) classifier that incorporates fuzzy logic and associative classification. This method optimizes rule selection for efficiency and accuracy, enhancing the detection process with a focus on interpretability and robustness against variable coin features.

However, a primary drawback of these methods is their reliance on large datasets and data augmentation techniques [10, 11]. In real-world scenarios, obtaining such extensive and diverse datasets can be challenging, as access to counterfeit coins and their images is often restricted due to security concerns. This limitation underscores the importance of developing methods that can achieve high accuracy with limited data, which is the primary focus of this thesis.

2.3 Few-Shot Learning

As discussed before, one of the persistent challenges in counterfeit coin detection is the scarcity of data. Most techniques discussed in the literature heavily rely on large datasets to train models effectively [23, 25, 26, 27, 28]. This dependence on extensive datasets is particularly problematic in this field, as acquiring counterfeit coins and their images is often difficult due to security concerns and the rarity of counterfeit specimens. Additionally, many existing methods resort to data augmentation to artificially expand datasets, which can introduce biases and fail to capture the true variability of real-world data.

Few-shot learning (FSL) [12, 13] is a machine learning approach designed to address the problem of limited data. Unlike traditional methods that require large amounts of labeled

data, FSL aims to train models to perform tasks with only a few examples. This approach is particularly useful in scenarios where data is scarce or expensive to obtain. Few-shot learning models typically employ techniques such as meta-learning, where the model is trained on a variety of tasks to learn a generalizable representation that can quickly adapt to new tasks with minimal data. Another common approach is to use pre-trained models on large datasets and fine-tune them on the small dataset available for the specific task. This allows the model to leverage the knowledge acquired from the large dataset while being tailored to the nuances of the specific application. FSL alleviates the challenges posed by traditional supervised learning methods in several key ways:

- **Reduction in Data Requirements:** FSL eliminates the need for large volumes of labeled data, which are often costly and difficult to obtain, particularly in niche fields like counterfeit coin detection.
- **Computational Efficiency:** By extending a pre-trained model to new categories without the need to re-train from scratch, FSL saves significant computational resources.
- **Adaptability to Rare Categories:** FSL models can effectively learn about rare or newly identified categories with exposure to only limited prior information, making it ideal for detecting rare counterfeit coin types.
- **Handling Domain Shifts:** Even if the model has been pre-trained on a statistically different distribution of data, FSL can adapt to new domains as long as the support and query sets are coherent.

In the context of counterfeit coin detection, few-shot learning can be instrumental. Traditional image-based and machine learning-based methods have shown promising results but are hampered by their reliance on large, augmented datasets. By employing few-shot

learning, it becomes feasible to develop robust models that perform well even with limited genuine and counterfeit coin images. Our research is pioneering in applying few-shot learning to the task of counterfeit coin detection using real datasets, without relying on data augmentation. This approach not only mitigates the data scarcity issue but also enhances the model's ability to generalize from limited examples, making it more practical and applicable in real-world scenarios. By leveraging advanced techniques such as Vision Transformers and the multimodal capabilities of GPT-4, our models can effectively learn and adapt to the subtle differences between genuine and counterfeit coins with minimal data.

In summary, few-shot learning represents a significant advancement in addressing the challenges posed by data scarcity in counterfeit coin detection. Our work contributes to the field by demonstrating the effectiveness of this approach, setting a new benchmark for future research and applications in detecting counterfeit coins with limited data.

2.4 Transfer Learning

Transfer learning [37, 38] is a powerful technique in machine learning that leverages knowledge gained from one task to enhance the performance of a model on a related, but different, task. This method is particularly advantageous in scenarios where data is scarce, as it allows models to benefit from pre-existing data and learned features, thereby reducing the need for extensive new data collection and training. The process of transfer learning generally involves the following steps:

1. **Selection of a Pre-Trained Model:** A model that has already been trained on a large dataset for a specific task is chosen as the starting point. This pre-trained model, often trained on datasets such as ImageNet, has learned to recognize a wide array of features relevant to many tasks.

2. **Freezing the Base Layers:** The initial layers of the pre-trained model, which capture general features, are often frozen. This means their weights are not updated during subsequent training, preserving the learned information.

3. **Adding New Layers:** New layers are added on top of the pre-trained model. These layers are trainable and can adapt to the specific features of the new task.

4. **Fine-Tuning:** The model is fine-tuned using the dataset for the new task. Fine-tuning adjusts the weights of the newly added layers and, optionally, some of the base layers to improve the model's performance on the target task.

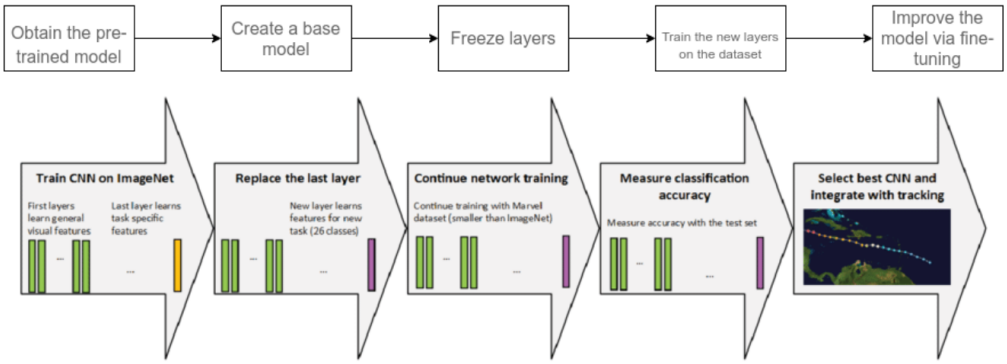

Figure 2: Transfer Learning process [2]

Transfer learning offers several advantages:

1. **Efficiency:** It significantly speeds up the training process because the model starts with pre-learned features, reducing the amount of time and computational resources needed.

2. **Performance:** Models often achieve better performance on the new task because they can leverage previously acquired knowledge, which helps in learning the new task more effectively.

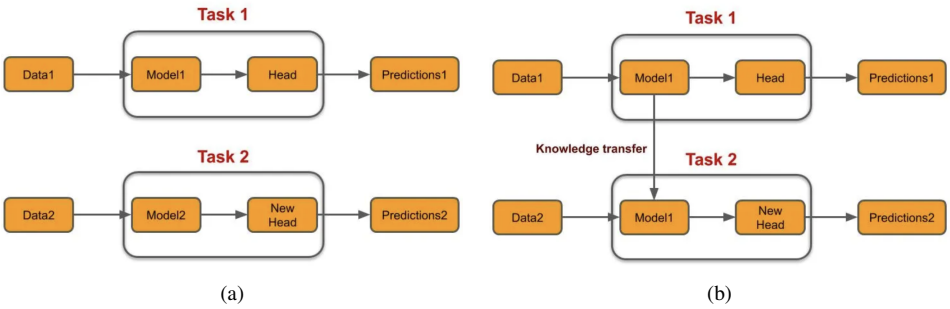

Figure 3: Comparison of Traditional Learning and Transfer Learning [3]: In traditional learning (a), separate models are trained independently for each task, requiring large datasets for each new task. In transfer learning (b), a pre-trained model on Task 1 is reused and fine-tuned with new data for Task 2, leveraging existing knowledge to improve performance and efficiency with limited data.

3. **Data Utilization:** It is particularly useful when there is a limited amount of labeled data for the new task. The pre-trained model's knowledge helps mitigate overfitting issues that commonly arise with small datasets.

Transfer learning is widely used in various domains, including computer vision, where pre-trained models like VGG, ResNet, and Inception are commonly used for tasks such as image classification, object detection, and segmentation. In natural language processing (NLP), models such as BERT, GPT, and word2vec, pre-trained on vast corpora [17, 18], are fine-tuned for specific NLP tasks like sentiment analysis, translation, and text classification. As mentioned earlier, transfer learning is one of the key techniques in few-shot learning. As a result, in the context of counterfeit coin detection, transfer learning is particularly relevant because it helps mitigate the data scarcity problem. Our research applies transfer learning to develop robust Vision Transformers for counterfeit coin detection. By leveraging pre-trained ViTs, we can fine-tune the models to recognize the subtle differences between genuine and counterfeit coins. This approach not only enhances the detection accuracy but also reduces the dependency on extensive new datasets.

2.5 Introduction to GPT-4 Multimodal Model

GPT-4 is the fourth iteration of the Generative Pre-trained Transformer (GPT) series developed by OpenAI. Unlike its predecessors, GPT-4 is a multimodal model, which means it can process and integrate both visual and textual data. This capability allows GPT-4 to perform tasks that require a combination of language understanding and visual perception, making it particularly powerful for a wide range of applications, including coin authentication.

2.5.1 Architecture of GPT-4 Multimodal

GPT-4 is built upon the Transformer architecture, which is a neural network design originally introduced in the paper "Attention is All You Need" by Vaswani et al [39]. The Transformer architecture is designed to handle sequential data and is highly effective at capturing dependencies between different elements in a sequence, making it well-suited for tasks involving natural language. The key innovation in GPT-4 is its multimodal capability. Traditional language models like GPT-3 are solely text-based, meaning they can only process and generate text. GPT-4, on the other hand, can take in both images and text as input and generate outputs that can also include text or other forms of data. This is achieved by integrating visual encoders into the Transformer architecture, allowing the model to process images alongside text. The GPT-4 architecture is generally composed of:

Visual Encoder:

The visual component of GPT-4 uses a vision encoder, which is typically a convolutional neural network (CNN) like ResNet or Vision Transformers. The encoder processes input images, extracting features that are then converted into a form that can be integrated with the text data in the Transformer layers. These features might include patterns, textures, and other visual characteristics that are important for understanding the image content.

Textual Encoder:

The textual data is processed through the standard Transformer layers, where the model uses self-attention mechanisms to capture the relationships between different words or tokens in the input text. This part of the model functions similarly to GPT-3, where it excels at tasks involving natural language understanding and generation. However, in GPT-4, these layers now interact more dynamically with the visual data, enabling richer contextual comprehension.

Integration Mechanism:

The strength of GPT-4 lies in its ability to seamlessly integrate visual and textual data into a coherent representation. The model employs attention mechanisms that align features extracted from images with corresponding textual data, enabling it to perform complex tasks that require both types of input. In the context of coin authentication, for example, the model can analyze an image of a coin while simultaneously considering accompanying descriptive text to determine whether the coin is genuine or counterfeit. This ability to cross-reference and draw insights from multiple data streams sets GPT-4 apart from previous models.

2.5.2 Capabilities of GPT-4 Multimodal Model

The GPT-4 multimodal model represents a significant advancement in AI [40, 41], primarily due to its ability to process and integrate both visual and textual data. This integration opens up a wide range of capabilities, making GPT-4 a versatile tool in various domains. Below are some of the key capabilities of GPT-4's multimodal model:

1. **Image Captioning and Understanding:** GPT-4 can generate descriptive text for images, effectively summarizing the content of an image or providing detailed descriptions of specific elements within it [42]. For example, it can describe the layout

of objects in a scene, identify key features, or explain the humor in a visually-based joke. One practical application of this is in accessibility tools like the collaboration between GPT-4 and "Be My Eyes", where GPT-4 helps visually impaired users by describing images, objects, or even offering navigation assistance.

2. **Visual Question Answering (VQA):** The model can answer questions related to an image by understanding the visual content and connecting it with the query.

3. **Data Interpretation and Visualization:** GPT-4 can process complex visual data, such as graphs and charts, and provide detailed breakdowns and insights. This capability is particularly useful in academic and professional settings where data visualization plays a crucial role. For example, GPT-4 can analyze a plot, identify trends, and even make inferences based on the visual data, though it's important to note that it still requires human oversight to ensure the accuracy of these interpretations.

4. **Image Classification:** By integrating visual and textual data, GPT-4 can classify images into categories, which is crucial for tasks like object detection and recognition [46, 47].

5. **Contextual Analysis:** GPT-4 can analyze images within a broader context, considering both the visual elements and related text, which is particularly useful in fields like medical imaging, document analysis, and, as in this research, counterfeit coin detection.

6. **Optical Character Recognition (OCR):** GPT-4 is capable of performing OCR tasks, which involves extracting text from images. It can handle both simple and complex text recognition tasks, making it useful for digitizing documents, reading signs in images, or even solving math problems that involve reading and interpreting handwritten equations.

These features of GPT-4, combined with its ability to integrate visual and textual data seamlessly, have inspired researchers to explore its potential in various image-based tasks, including image classification and object recognition. The successful application of GPT-4 in these areas motivated us to evaluate its accuracy and effectiveness in the specific task of detecting counterfeit coins.

2.5.3 Prompting Techniques in GPT-4

Prompting [6] refers to the process of designing and inputting specific instructions or questions into an AI model like GPT-4 to elicit a particular type of response and produce desired outputs. It is a crucial concept in the use of large language models (LLMs) such as GPT-4. Since GPT-4 is a generative model trained on vast amounts of text data, the way a prompt is framed can significantly impact how the model interprets the query and generates its output. The effectiveness of prompting directly influences the quality and relevance of the model's responses. To optimize model performance for different tasks, various prompting techniques have been developed, including Zero-Shot, Few-Shot, Chain-of-Thought, and Generated Knowledge Prompting. This section provides an in-depth exploration of some of these techniques.

1. **Zero-Shot Prompting:** Zero-Shot Prompting [36] is a technique where the model is asked to perform a task without any examples or prior context provided. The model relies entirely on its pre-trained knowledge to generate a response. The term "zero-shot" implies that the model has received zero examples of how to complete the task within the current interaction.

    - **Advantages:** The main advantage of zero-shot prompting is its simplicity and efficiency. It does not require any examples or setup, making it quick and easy to use. It can also be surprisingly effective for tasks that are within the model's general knowledge base.

- **Limitations:** Because it relies entirely on the model's existing knowledge, it can sometimes produce less accurate or relevant results for complex or specialized tasks.

2. **Few-Shot Prompting:** Few-Shot Prompting [16] involves providing the model with a few examples of the task you want it to perform before asking it to generate a response. This technique helps the model better understand the task by showing it how similar tasks have been completed in the past. The term "few-shot" refers to the fact that only a small number of examples (usually 1-5) are provided.

    - **Advantages:** Few-shot prompting is particularly powerful for tasks that require more context or where the model's output can vary significantly depending on the task. It helps guide the model towards the desired output by providing it with relevant examples.
    - **Limitations:** The downside is that this method requires carefully chosen examples, and the quality of the output is heavily dependent on the quality of the examples provided.

3. **Chain-of-Thought Prompting:** Chain-of-Thought Prompting [34, 35] encourages the model to break down its reasoning process into a sequence of intermediate steps before arriving at a final answer. This technique is designed to improve the model's performance on complex reasoning tasks by prompting it to consider each part of the problem systematically.

    - **Advantages:** This technique is particularly useful for tasks that require logical reasoning or problem-solving. By breaking down the problem, the model is less likely to make mistakes that arise from skipping steps or misunderstanding the problem.

- **Limitations:** The main limitation is that chain-of-thought prompting can be more time-consuming and may require more computational resources due to the longer and more detailed responses it generates.

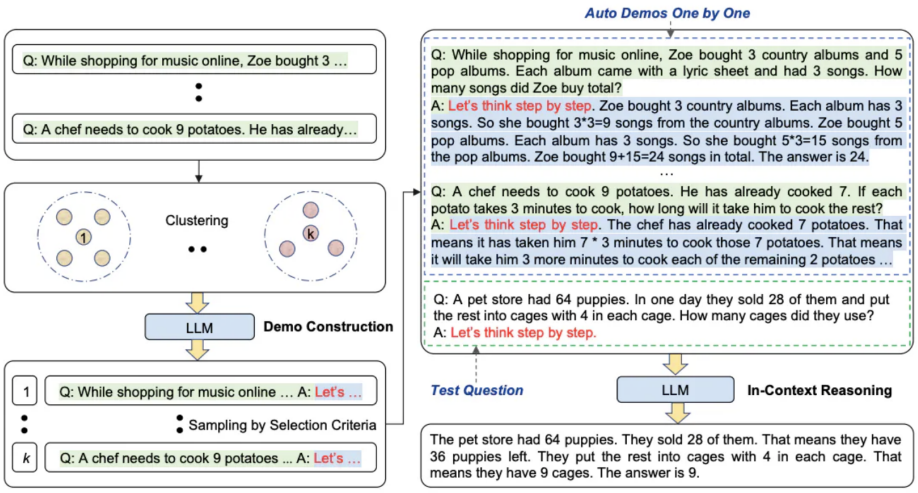

Figure 4: Overview of Chain-of-Thought Prompting [34].

4. **Generated Knowledge Prompting:** Generated Knowledge Prompting [33] involves providing the model with carefully crafted instructions or information that encourages the generation of content aligned with particular knowledge areas, facts, or understanding.

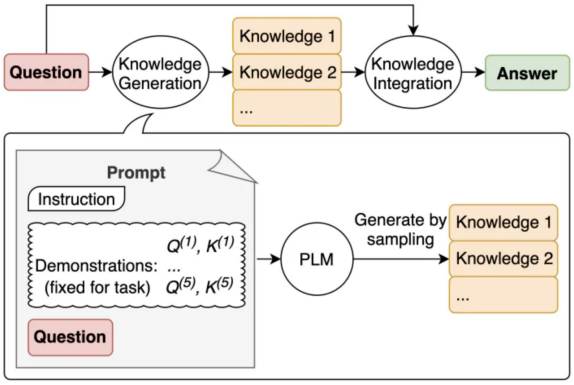

Figure 5: Example of Generated Knowledge Prompting [33].

- **Advantages:** Generated knowledge prompting can significantly improve the

quality of the model's output by ensuring it has all the necessary context before attempting the task. It is especially useful for complex tasks that require a deep understanding of the subject matter.

- **Limitations:** This technique can be more resource-intensive and may require careful orchestration to ensure that the generated knowledge is accurate.

2.6 Introduction to Vision Transformers (ViT)

Vision Transformers represent a transformative approach in the field of computer vision, diverging significantly from the traditional Convolutional Neural Networks (CNNs) [29] that have dominated the field for years. Introduced by Dosovitskiy et al. in their groundbreaking paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," [15] ViTs adapt the Transformer architecture, originally developed for natural language processing, to handle visual data. This section provides a comprehensive overview of Vision Transformers, their architecture, and how they function.

2.6.1 Architecture of Vision Transformers

The architecture of Vision Transformers is built upon the same foundational principles as the Transformer models used in NLP, such as BERT and GPT, but adapted to process images. Here are the key components that define the architecture of a ViT:

1. **Patch Embedding:**

    - Unlike CNNs, which process images through a series of convolutions and pooling layers [30, 31], Vision Transformers divide an input image into fixed-size patches. For instance, a typical image of 224x224 pixels might be split into 16x16 patches, resulting in a grid of 14x14 patches.

- Each of these patches is then flattened into a vector and linearly embedded into a lower-dimensional space, typically of size 768 or 1024 dimensions, which serves as the input to the Transformer model. This process is analogous to the tokenization step in NLP, where sentences are split into words or subwords.

2. **Positional Embedding:**

   - Since Transformers lack an inherent understanding of the spatial structure of images (unlike CNNs, which have spatial locality through convolution), positional embeddings are added to the patch embeddings to retain information about the relative positions of patches within the image. This step is crucial for preserving the spatial relationships between patches, which is vital for tasks like object detection and image classification.

3. **Transformer Encoder:**

   - The core of the Vision Transformer consists of several Transformer encoder layers, each comprising multi-head self-attention mechanisms and feedforward neural networks. The self-attention mechanism allows the model to weigh the importance of different patches relative to each other, capturing both local and global dependencies in the image. These dependencies are aggregated layer by layer, enabling the model to progressively build a richer, more nuanced representation of the image as it passes through each layer of the encoder.
   - The attention mechanism works by computing attention scores between each pair of patches, which determine how much focus the model should place on each patch when processing the current one. This results in a set of attention maps that capture the relationships between different regions of the image.

4. **Classification Head:**

   - At the beginning of the sequence of patch embeddings, a special [CLS] token is added, which, after passing through the Transformer layers, is used by the classification head to predict the final output (e.g., the class label in image classification tasks). The final layer typically consists of a feedforward network that maps the output of the [CLS] token to the desired number of classes.

5. **MLP Head:**

   - The final output of the Transformer is passed through a Multi-Layer Perceptron (MLP) head that produces the classification logits. This MLP head usually consists of one or more fully connected layers and is responsible for making the final decision based on the learned representations.

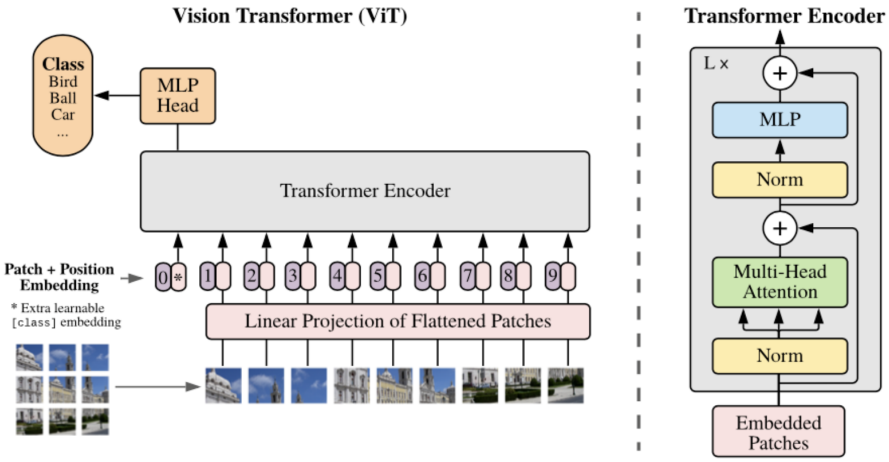

Figure 6: Vision Transformer Architecture [15].

Key Advantages:

- **Global Context Understanding:** ViTs excel at capturing long-range dependencies and global context due to their self-attention mechanisms, which contrasts with the

local receptive fields of CNNs.

- **Scalability:** ViTs are highly scalable, as they can be trained on extremely large datasets, such as ImageNet, and easily adapted to various tasks.

- **Transfer Learning:** The architecture of ViTs allows them to be pre-trained on large datasets and then fine-tuned on smaller, task-specific datasets, making them highly versatile.

2.6.2 How Vision Transformers Work

Vision Transformers represent a novel approach to image processing by leveraging the strengths of the Transformer architecture, traditionally used in natural language processing. The process begins with the transformation of an input image into a sequence of patches, each of which is flattened and linearly embedded into a high-dimensional space. This transformation allows the model to treat the image patches similarly to how words are treated in a sentence within a language model. After the image has been tokenized into patches, positional embeddings are added to retain the spatial information of the image. These embeddings ensure that the model understands the relative positions of the patches within the original image, which is crucial for tasks that require spatial awareness, such as object detection or image classification.

The core of the Vision Transformer is the Transformer encoder, which consists of multiple layers of self-attention mechanisms and feedforward networks. The self-attention mechanism enables the model to weigh the importance of each patch relative to others, capturing both local and global dependencies in the image. This allows the Vision Transformer to consider the entire context of the image rather than focusing solely on local features, which is a significant departure from how traditional CNNs operate. As the image patches pass through the Transformer layers, the model progressively integrates information from

all patches to form a comprehensive understanding of the image. This process of global context integration allows the Vision Transformer to identify relationships between distant patches that might be critical for tasks such as identifying subtle patterns or anomalies in counterfeit coin detection. Once the Transformer layers have processed the patches, the output corresponding to the [CLS] token is used as the aggregate representation of the entire image. This representation is then passed through a Multi-Layer Perceptron head, which maps the features extracted by the Transformer into a final prediction, such as classifying the image as a genuine or counterfeit coin. The Vision Transformer is typically pre-trained on large datasets, allowing it to learn general features that can be applied to various tasks. After pre-training, the model can be fine-tuned on a smaller, task-specific dataset, making it particularly well-suited for applications like counterfeit coin detection where the availability of labeled data may be limited.

In summary, Vision Transformers work by transforming an image into a sequence of patches, embedding these patches into a high-dimensional space, and processing them through a series of Transformer layers that capture both local and global dependencies. The output is a comprehensive representation of the image, which is then used for classification or other tasks. This approach offers several advantages over traditional CNNs, particularly in tasks that require an understanding of global context and long-range dependencies within images, making ViTs especially powerful in domains like counterfeit coin detection where such an understanding is crucial.

Chapter 3

Dataset Preparation

3.1 Dataset Creation

As highlighted earlier, a crucial aspect of detecting counterfeit coins is having a diverse dataset that accurately represents both genuine and fake coins. This is essential for training and evaluating machine learning models tasked with identifying subtle differences between the two categories.

To create such a dataset, we developed a collection that includes a significant number of both counterfeit and authentic coins. Specifically, we assembled a dataset comprising 732 Danish and 112 Chinese coins, which were meticulously scanned using a Keyence 3D scanner. This advanced scanner is capable of capturing full surface data across each coin with an exceptionally high resolution of 0.1 μm. The scanner's rotational scanning feature further enhances its capability by allowing it to capture the entire surface of the coins in great detail, ensuring that no features are missed. These coins were sourced from reliable and verified origins, including the Danish law enforcement agency and various coin exhibitions. The authenticity and provenance of the coins were rigorously verified to ensure that the dataset accurately reflects the characteristics of genuine and counterfeit coins. This step

is crucial, as the quality of the dataset directly influences the performance of the machine learning models that rely on it.

The scanning process itself was remarkably efficient. The Keyence 3D scanner required approximately 10 seconds to scan each coin, although larger coins naturally took slightly longer due to their size. The resulting scanned images boast a resolution of approximately 3550 pixels in both length and width, providing a highly detailed view of each coin's surface. This level of detail is vital for the detection of minute differences that may indicate counterfeiting. To ensure uniformity and maintain high image quality across the entire dataset, all scans were conducted under consistent lighting conditions, and each coin was handled with care throughout the scanning process. This meticulous attention to detail during the scanning process helps to minimize any variations that could introduce noise or bias into the dataset, thereby enhancing the reliability of the subsequent analysis.

The dataset created through this process serves as a robust foundation for developing and testing counterfeit coin detection models. Its diverse nature and high-quality imaging ensure that the models trained on it have access to the critical features necessary for accurate and reliable classification of coins as genuine or counterfeit.

Table 1: Number of Samples for Train and Test Data.

Datasets	Train		Test	
	Genuine	Fake	Genuine	Fake
Danish	31	32	367	302
Chinese	22	34	4	52

3.2 Dataset Preparation

After collecting the high-resolution images of both genuine and counterfeit coins using the Keyence 3D scanner, some preprocessing steps were undertaken to standardize the

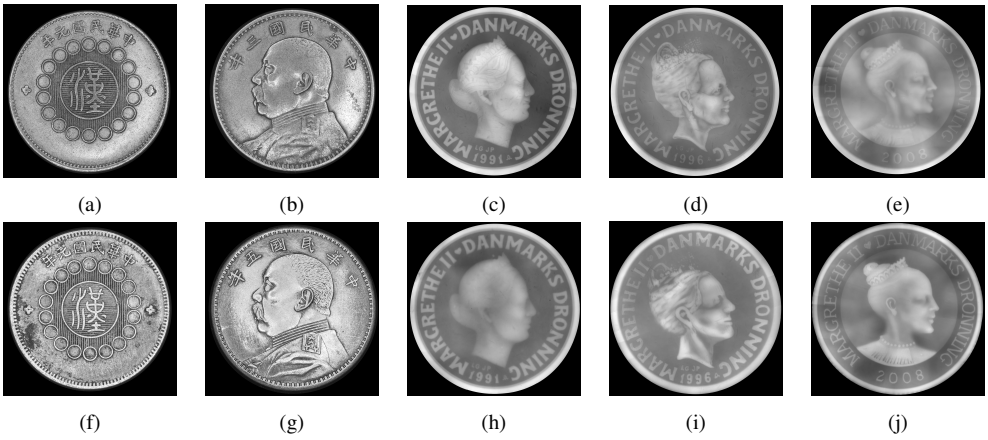

Figure 7: Comparative Display of Genuine and Counterfeit Danish and Chinese Coins. This collection features paired images illustrating genuine and counterfeit versions of specific coins: (a) & (f) a 1912 Chinese coin, (b) & (g) a 1921 Chinese coin, (c) & (h) a 1991 Danish coin, (d) & (i) a 1996 Danish coin, and (e) & (j) a 2008 Danish coin. Each pair highlights the nuances in detail as captured by high-resolution scanning.

dataset. Standardization is crucial to ensure consistency across all images, which, in turn, enhances the reliability and accuracy of the subsequent machine learning models. The first preprocessing step involved rotating all images in one consistent direction. Uniform orientation across the dataset ensures that the models do not mistakenly learn orientation-based features that are irrelevant to the task of counterfeit detection. Following this, all images were converted to grayscale. Grayscale conversion simplifies the image data by removing color information, which is often unnecessary for this task. This also reduces the computational load during model training, as the models now need to process only one channel of intensity values instead of three (RGB channels). Grayscale images focus the model's attention on structural and textural features, which are more pertinent to detecting subtle differences between genuine and counterfeit coins.

These standardized grayscale images were then used for training the Vision Transformer model. The ViT model requires high-quality, consistent input images to effectively learn and detect the nuances between genuine and counterfeit coins. It is important to note that the ViT training utilized only these individual images and not the triplet arrangements used later for the GPT-4 evaluation.

For the evaluation by the multimodal GPT-4 model, the dataset was further prepared to meet the specific requirements of this system. Given the unique nature of GPT-4, which integrates both visual and contextual data, the dataset needed to be tailored specifically for this model's input structure. To this end, triplets of coin images were created. Each triplet consisted of a test coin (anchor), a visually similar genuine coin (positive example), and a visually similar counterfeit coin (negative example) (Fig. 8). This arrangement is inspired by the concept of triplet loss in machine learning, where the goal is to make the model understand that the anchor is closer to the positive than the negative item. This method helps the model distinguish between similar and dissimilar items more effectively and enhances the model's ability to discriminate between subtle differences by leveraging the underlying logic of triplet loss.

Figure 8: Example of a Coin Image Triplet Used in Multimodal System Evaluations: From left to right: Genuine coin, counterfeit coin, and test coin.

The selection of these similar coins was based on feature vectors extracted from each image using a ResNet50 neural network [32]. ResNet50 is a powerful convolutional neural network widely used in computer vision tasks. It is particularly known for its residual connections, which help to alleviate the vanishing gradient problem, allowing for the training of very deep networks. ResNet50 has 50 layers and has been pre-trained on large datasets like ImageNet, making it adept at feature extraction. For each coin image, ResNet50 was employed to extract a feature vector, which is essentially a numerical representation of the image that encapsulates its most important features. These feature vectors were then

compared using cosine similarity to determine the visual closeness between different coin images. Cosine similarity is a measure of similarity between two non-zero vectors. It is calculated by taking the dot product of the two vectors and dividing it by the product of their magnitudes. Mathematically, cosine similarity is expressed as:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

where A and B are the feature vectors of the two images being compared. The result of this calculation is a value between -1 and 1, where 1 indicates that the two vectors are identical, 0 indicates that they are orthogonal (i.e., share no similarity), and -1 indicates that they are diametrically opposed.

Using this metric, the most visually similar genuine and counterfeit coins were identified for each test coin. These selected coins were then paired to form the triplets used in the GPT-4 evaluation process. The multimodal GPT-4 model was tasked with assessing the authenticity of the test coin by comparing it with the genuine and counterfeit coins in the triplet. The model leveraged both the visual similarities, as determined by cosine similarity, and contextual cues to make its determination.

Chapter 4

Multimodal Coin Authentication Using GPT-4

4.1 GPT-4 API: Functionality and Integration

The GPT-4 API [43], provided by OpenAI, serves as an interface through which developers and researchers can access the powerful capabilities of the GPT-4 multimodal model. The API is designed to be flexible and user-friendly, enabling easy integration into a wide range of applications that require processing both text and image inputs. It supports input in both text and image formats, allowing users to fully leverage GPT-4's multimodal capabilities.

Programming Environment and Language

The implementation of the GPT-4 API was developed using Python, which offers robust libraries for handling both image and text data, making it a suitable choice for integrating the API. Python's libraries like **requests**, **Pandas**, and **PIL (Pillow)** were utilized to manage HTTP requests, data processing, and image handling.

- **requests** was used to make HTTP-based API calls.

- **Pandas** handled data management, ensuring responses were logged and saved efficiently.
- **PIL** managed image processing, converting coin images into the required base64 encoding format for the GPT-4 API.

Workflow for Coin Authentication

The implementation involved a few key steps:

1. **API Key Access:** To use the GPT-4 API, developers must obtain an API key from OpenAI. This key is essential for authenticating requests and managing usage limits. In the implementation, the key was stored in the environment variables for security reasons.
2. **Image Preparation:** The input images of the coins (both genuine, counterfeit, and test coins) were stored in a local directory. Each image was encoded into a base64 format using the PIL library, which was necessary for sending the image data via the API.
3. **API Requests:** Requests to the GPT-4 API were structured using the requests library. The request payload included both image data (in base64 format) and a text prompt that provided context for the coin classification task. The specific prompt instructed GPT-4 to classify the unlabeled coin (on the right) based on a comparison of engravings, lettering, and other details between the genuine and counterfeit coins.
4. **Handling Responses:** Once the API responded, the output was captured in JSON format. The response included the GPT-4 model's prediction (genuine or counterfeit), a confidence score, and a detailed explanation of its decision (Table 2). These responses were parsed, and the relevant information was saved in a CSV file for further analysis.

5. **Rate Limiting and API Usage:** Given OpenAI's API rate limits, the implementation adhered to the restriction of one request per second. Additionally, error handling was implemented to manage situations where the API request failed, by retrying after a short delay.

Algorithm 1 GPT-4_Coin_Authentication

Require: Directory of coin images ($I = \{i_1, i_2, \ldots, i_n\}$), where each i represents a coin image.
Require: API Key for accessing GPT-4 services.
Ensure: A classification of each test coin image as genuine or counterfeit, along with a confidence score and explanation.

Initialize $F_o \leftarrow \emptyset$ (empty results dataframe)
if ExistingResults $\neq \emptyset$ **then**
    Load results R from CSV file and exclude processed images.
end if
for each image $i_j \in I$ **do**
    Encode the image i_j as base64.
    Prepare the API request with input parameters:
    - Include text prompt for coin authentication.
    - Add encoded image data.
    Send API request using HTTP POST method with API key.
    Receive API response:
    if response == success **then**
        Parse the classification, confidence score, and explanation.
        Append results to F_o.
    else if response == error **then**
        Log error and retry after waiting.
    end if
end for
Save results to CSV after each image is processed.
Return final result set F_o containing the classification, confidence, and explanation for each image.

Figure 9: The pseudo-code of the GPT-4 Coin Authentication Algorithm

4.2 Security Considerations for Using GPT-4 API

When deploying AI systems, particularly for sensitive tasks like counterfeit coin detection, data security is a critical concern. This section outlines the measures taken to ensure the secure handling of the dataset used in this research, which includes both genuine and counterfeit coin images.

1. **Data Encryption:** All data transmitted between the local system and the GPT-4 API is secured using HTTPS, which employs encryption protocols such as TLS (Transport Layer Security). This prevents unauthorized interception of the dataset during transmission [49].

2. **No Data Retention:** OpenAI, the provider of GPT-4, does not store or use submitted data for any purpose beyond the immediate interaction, unless explicitly permitted by the user. In this research, no such permission was granted, ensuring that images submitted to the API were processed securely and deleted immediately after use [50].

3. **Privacy and Compliance:** OpenAI adheres to industry standards such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). These regulations enforce strict protections on data privacy, ensuring that user data is handled with care and not exposed to unauthorized personnel or third parties [51].

Given these security measures, it is concluded that utilizing the GPT-4 API to process the coin dataset does not pose a significant security risk. The encryption protocols and non-retention policies ensure the confidentiality and integrity of the dataset are maintained throughout the process.

4.3 Application of Prompting Techniques in Coin Authentication

In our research, we explored and applied several prompting techniques to evaluate GPT-4 effectiveness in the task of counterfeit coin detection. This section details the specific prompting techniques used and how they were applied.

Table 2: Sample Output from GPT-4 API: Classification Results with Confidence Scores and Explanations.

Image Name	Label	Prediction	Confidence Score	Explanation
coin-262	Genuine	Counterfeit	50%	In comparing the three coins, several factors suggest that the third coin could be counterfeit ...
coin-263	Genuine	Genuine	80%	Upon examining the three coins, several factors contribute to the assessment of the authenticity of the coin on the right: 1. Quality and clarity of ...
coin-28	Genuine	Genuine	75%	Upon examining the three Danish coins, the focus has been ...
coin-379	Counterfeit	Counterfeit	90%	Upon careful inspection of the three coins, focusing on the elements you have outlined, it is possible to compare ...
coin-412	Counterfeit	Genuine	55%	Upon careful examination of the three coins, several observations can be made ...
coin-452	Counterfeit	Counterfeit	80%	Upon examining the three coins, it is noticeable that the right coin shares some similarities with the middle coin ...

Zero-Shot Prompting

In the Zero-Shot Prompting scenario, we tasked GPT-4 with predicting the authenticity of a coin using only the image of that coin, without providing any prior examples or contextual information about genuine or counterfeit coins. The model was required to make a prediction based solely on its pre-trained knowledge and understanding of visual features from the image. This approach simulates a scenario where the model encounters a completely new task or dataset and must rely on its inherent capabilities.

Few-Shot Prompting

For Few-Shot Prompting, we provided GPT-4 with triplet images as input. Each triplet consisted of a test coin image, a genuine coin image, and a counterfeit coin image. Alongside the images, we included a prompt that contained information about the authenticity

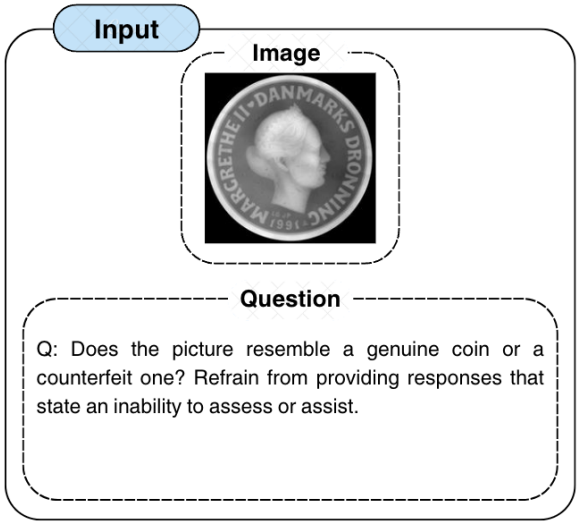

Figure 10: Example of Zero-Shot Prompting: The model is asked to determine whether the coin in the image is genuine or counterfeit without any prior examples.

of the two comparison coins. This method allowed GPT-4 to use these examples to inform its prediction about the test coin. By comparing the test coin to both the genuine and counterfeit examples, the model could make a more informed decision about the test coin's authenticity.

Chain-of-Thought Prompting

In our experiments with the Chain-of-Thought (CoT) Prompting technique, we encouraged GPT-4 to break down its reasoning process into a sequence of intermediate steps. This approach was particularly useful for complex reasoning tasks, such as distinguishing between genuine and counterfeit coins based on subtle visual differences. We combined CoT Prompting with Few-Shot Prompting to see if this approach would yield better results. In this setup, the model was prompted to explain its thought process step-by-step, helping it to systematically analyze the features of the test coin in comparison to the provided examples.

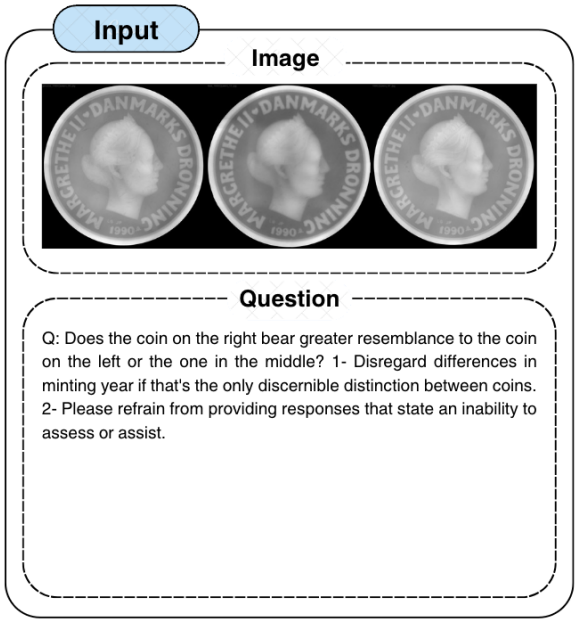

Figure 11: Example of Few-Shot Prompting: The model is provided with a set of three coins—one genuine, one counterfeit, and one test coin—and is asked to compare the test coin with the provided examples.

Generated Knowledge Prompting

Lastly, we explored the Generated Knowledge Prompting technique, which involves generating relevant information or context about the task before making a prediction. In this case, we provided GPT-4 with background knowledge about the characteristics of genuine and counterfeit coins. This additional context was then integrated into the prompt to guide the model's prediction. As illustrated in Fig. 13, this prompting technique includes a unique section labeled 'Knowledge' that is absent in other prompting methods. In this section, we provided the model with specific information to help it distinguish between a counterfeit coin and an authentic one. We instructed the model to focus on features such as the quality and clarity of the edge engravings, the sharpness and detail of the main design element, and the lettering and spacing of words and numbers, in order to evaluate the test coin.

In this section, we investigated the impact of various prompting techniques on the accuracy of GPT-4 in the task of counterfeit coin detection. The specific approaches—Zero-Shot, Few-Shot, Chain-of-Thought, and Generated Knowledge Prompting—were applied

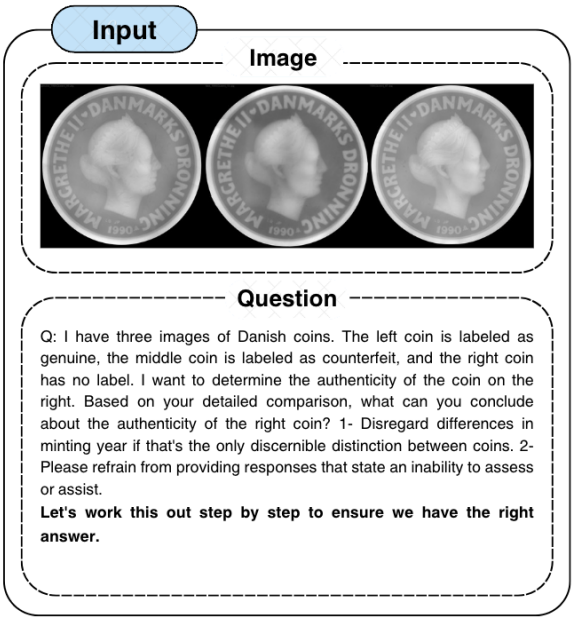

Figure 12: Example of Chain-of-Thought Prompting: Similar to Few-Shot Prompting, but with additional instructions to encourage step-by-step reasoning.

to assess how each technique influences the model's performance. The outcomes of these experiments are thoroughly analyzed and discussed in the subsequent results section.

4.4 Results and Discussion

In this section, we analyze the performance of the GPT-4 multimodal model for counterfeit coin detection using different prompting techniques, with the obtained results summarized in Table 3. The dataset used in these evaluations consisted of triplet images, designed specifically for this task. Each triplet included a test coin flanked by its most visually similar genuine and counterfeit counterparts, selected using cosine similarity.

The performance metrics obtained—precision, recall, F-Score, and accuracy—highlight the effectiveness and limitations of each prompting method. As illustrated in Table 3, Zero-Shot Prompting resulted in the highest recall (96.4%) but suffered from lower precision (55.0%) and overall accuracy (55.3%). This outcome is expected, as Zero-Shot Prompting does not leverage any prior examples or context, leading the model to make a prediction

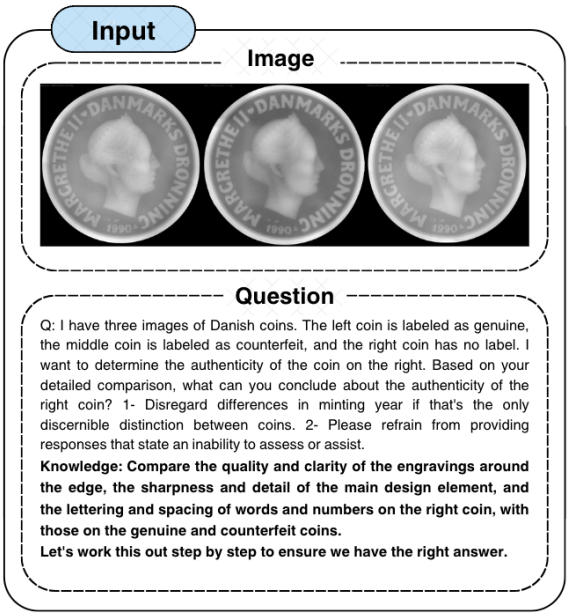

Figure 13: Example of Generated Knowledge Prompting: T

based solely on the single input image. The high recall indicates that the model is good at identifying counterfeit coins, but the lower precision suggests it also incorrectly identifies some genuine coins as counterfeit.

The Few-Shot Prompting approach provided a slight improvement in accuracy (55.7%) over Zero-Shot Prompting, with better precision (69.8%) but a drop in recall (57.8%). This method allowed the model to use examples from the prompt to inform its predictions, but the trade-off between precision and recall indicates that while the model became better at identifying genuine coins, it was less effective at detecting counterfeit ones.

Chain-of-Thought Prompting, which encourages step-by-step reasoning, resulted in moderate improvements, achieving an accuracy of 56.6%. This technique helped the model to break down the problem into more manageable steps, leading to slightly better performance in precision and recall compared to Few-Shot Prompting.

Lastly, Generated Knowledge Prompting, which incorporates additional information to improve prediction accuracy, provided a modest improvement, achieving an overall accuracy of 56.7%. While the improvements were not substantial, this technique demonstrates the

potential of enhancing model performance by integrating relevant external knowledge into the prompts.

Table 3: Comparative analysis of Precision, Recall, F-Score, and Accuracy for different prompting methods.

Method	Precision	Recall	F-Score	Accuracy
Zero-Shot	55.0	96.4	70.1	55.3
Few-Shot	69.8	57.8	63.3	55.7
Chain-of-Thought	57.5	79.1	66.5	56.6
Generated Knowledge	57.7	78.0	66.3	56.7

Despite these efforts to improve GPT-4's performance through advanced prompting techniques, the overall accuracy (average of 56.13%) was significantly lower than that of the ViT model and other specialized methods to be discussed in Chapter 5. This disparity can be attributed to the fundamental differences between GPT-4, a general-purpose language model with some visual capabilities, and models like ViT, which are specifically designed and optimized for image classification tasks. The findings underscore the necessity for further enhancements in GPT-4's ability to process and analyze visual data if it is to compete with specialized models in tasks such as counterfeit coin detection.

Chapter 5

Vision Transformer-Based Coin Authentication

5.1 Implementation of Vision Transformer for Coin Authentication

As mentioned earlier, the Vision Transformer model was another approach investigated in this research for counterfeit coin detection. This section outlines the steps taken to implement and fine-tune the ViT model for the task of distinguishing between genuine and counterfeit coins. The process involved configuring the model, preparing the dataset, and executing a detailed training and evaluation process. As previously discussed, the dataset consisted of 732 Danish and 112 Chinese coins, including both genuine and counterfeit specimens. To prepare our training dataset, we selected up to five coins from each year and each model, ensuring a representative selection of different coin types. This approach demonstrated the model's ability to predict authenticity accurately with a limited number of samples. The remaining, larger portion of the dataset was then used as the testing set to evaluate the model's generalization capabilities. The images were preprocessed by rotating

them to a uniform orientation and converting them to grayscale, ensuring consistency across the dataset.

We employed the pre-trained Vision Transformer model, google/vit-huge-patch14-224-in21k, available in the Hugging Face library. The ViT model is particularly suited for image classification tasks due to its ability to capture intricate patterns and fine-grained details through its multiple transformer layers.

Key Features:

- **Patch Embedding:** The model processes input images by dividing them into non-overlapping patches of size 14x14 pixels, effectively transforming the image into a sequence of patches.
- **Self-Attention Mechanism:** These patches are projected into a high-dimensional space, where self-attention mechanisms are applied. This allows the model to learn relationships between different parts of the image, capturing both local and global context.
- **Transfer Learning:** Leveraging the pre-trained weights on the ImageNet-21k dataset, the model is fine-tuned to adapt to the specific task of counterfeit coin detection. This approach reduces the need for extensive training data while maintaining high accuracy.

The model configuration also included the use of a linear classifier layer that takes the output from the ViT model and maps it to the classification labels (genuine or counterfeit). The classifier was fine-tuned on our dataset, allowing the model to learn the distinguishing features of the coins.

Training Process:

The training of the ViT model was conducted using the Hugging Face Trainer API, which simplifies the process of fine-tuning transformer models. During the training of the Vision Transformer model, the system utilized up to 13.5 GB of GPU RAM, reflecting the computational demands of fine-tuning a model of this scale. The training process involved the following steps:

1. **Training Parameters:**

    - **Learning Rate:** A learning rate of **1e-4** was selected to ensure steady progress during training without making overly large updates to the model's weights, which can help in avoiding overshooting the optimal point.
    - **Batch Size:** The batch size was set to **4**, allowing for efficient use of GPU memory during training.
    - **Epochs:** Table 4 demonstrates the model's performance across different numbers of epochs. As shown, training for more than 4 epochs did not result in significant accuracy improvements across the datasets and only consumed additional computational resources and training time. Conversely, reducing the number of epochs below 4 led to lower accuracy, as seen with the 2-epoch results for certain years, where the model had less exposure to the training data. Therefore, training the model for **4** epochs provides an optimal balance, ensuring sufficient exposure to the data while achieving high accuracy and avoiding unnecessary computational costs and time.
    - **Precision:** Mixed precision (FP16) was used to accelerate training while maintaining model accuracy.

Table 4: Model accuracy across different epochs for various years in the Danish coin dataset.

Dataset	Epoch				
	2	4	8	10	12
20 Kroner 1990	100	100	100	100	100
20 Kroner 1991	95.26	96.1	96.1	96.1	96.1
20 Kroner 1996	98.19	100	100	100	100
20 Kroner 2008	91.97	99.6	99.6	99.6	99.6

2. **Model Training:**

   - As discussed, the dataset was split into training and testing sets, with a larger portion reserved for testing. The training set was used to fine-tune the pre-trained ViT model, allowing it to learn the specific features that distinguish genuine coins from counterfeit ones.
   - During training, the model's performance was monitored using various metrics, including accuracy, precision, recall, and F1-score.

3. **Loss and Accuracy Monitoring:**

   - The training process was accompanied by continuous monitoring of the loss function and accuracy on the training set. The loss function used was cross-entropy, which is well-suited for classification tasks.

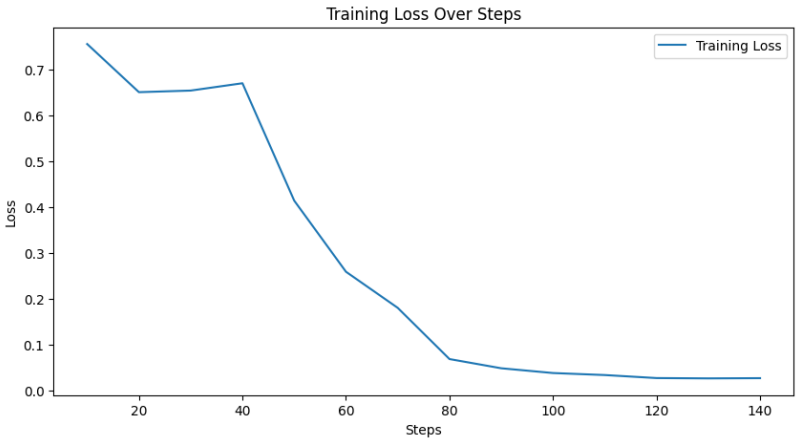

Figure 14: Training loss over steps curve for the Vision Transformer model

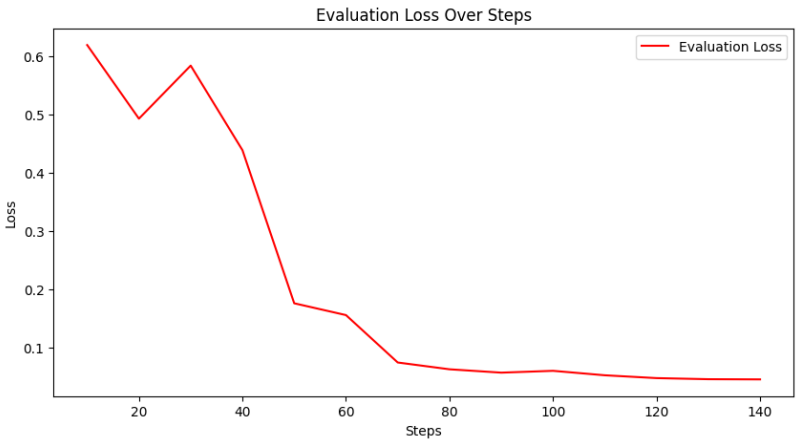

Figure 15: Evaluation loss over steps curve for the Vision Transformer model

- The accuracy on the training set was logged at each step, ensuring that the model was effectively learning the patterns in the data. The final testing accuracy was **99.31 %**, indicating that the model had successfully learned to distinguish between the two classes.

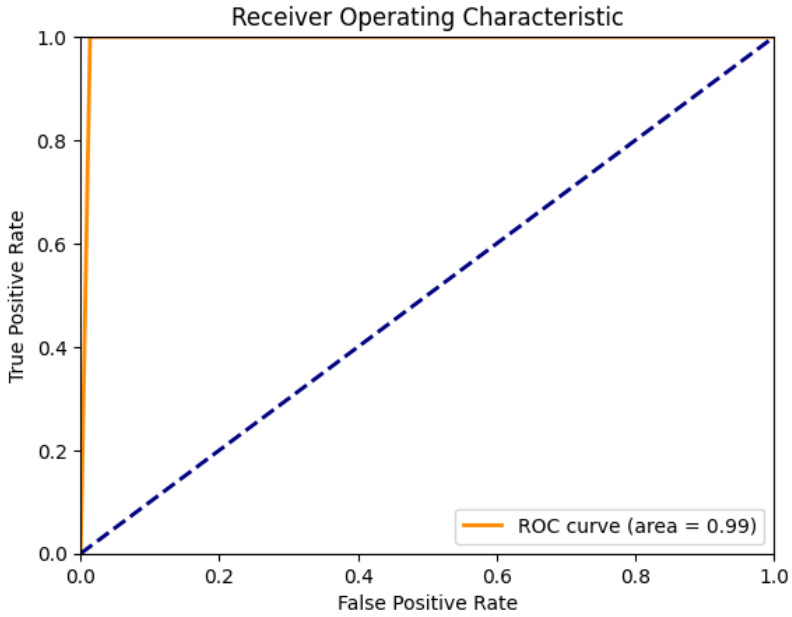

Figure 16: Receiver Operating Characteristic (ROC) curve.

Fig. 14 and Fig. 15, "Training Loss Over Steps" and "Evaluation Loss Over Steps," provide a comprehensive view of the model's learning dynamics during the training process. The "Training Loss Over Steps" diagram shows how the loss decreases as the training progresses, indicating that the model is effectively learning to minimize errors on the training data. The gradual decline of the loss curve suggests that the Vision Transformer model is becoming increasingly proficient at distinguishing between genuine and counterfeit coins as it processes more data.

The "Evaluation Loss Over Steps" diagram, on the other hand, reflects the model's performance on unseen data, which is crucial for assessing its generalization capabilities. The evaluation loss follows a similar downward trend as the training loss, signifying that the model is not only learning well from the training data but also generalizing effectively to new data points. The stability of the evaluation loss towards the end of the training process suggests that the model has reached an optimal point of learning, where further training is unlikely to yield significant improvements.

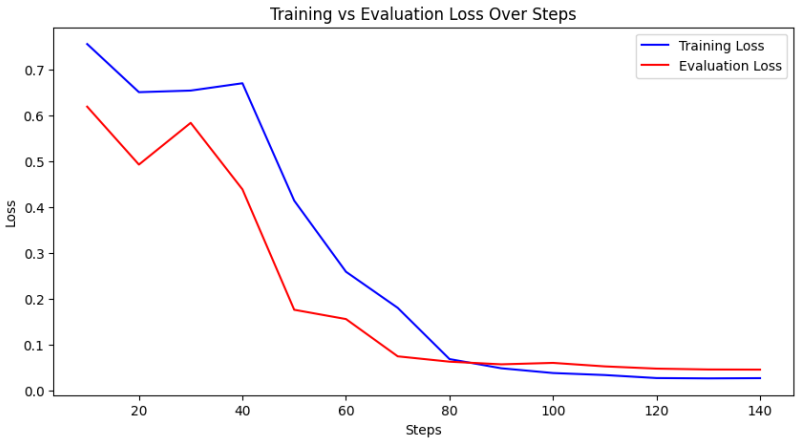

Figure 17: Training vs Evaluation loss over steps

The third diagram in Fig. 17, which compares the training and evaluation losses, provides further insights into the model's performance. Both losses decrease over time, indicating successful convergence. The close alignment between the training

and evaluation losses towards the end of the training process suggests that the model is not overfitting the training data, which is a positive outcome. Overfitting occurs when a model performs well on the training data but fails to generalize to new, unseen data. The convergence of the training and evaluation losses implies that the ViT model is well-generalized and capable of accurately predicting the authenticity of coins across different datasets.

Algorithm 2 ViT_Coin_Authentication

Require: Directory of coin images ($I = \{i_1, i_2, \ldots, i_n\}$), where each i represents a coin image.
Require: Pretrained ViT model
Ensure: Classification of coin images as genuine or counterfeit, along with performance metrics such as accuracy, precision, recall, and F1 score.

Initialize $F_o \leftarrow \emptyset$ (empty results dataframe)
Download datasets for genuine and fake coins from Google Drive.
Unzip the datasets into directories for genuine and counterfeit coins.
Data Preprocessing:
    Create Dataset:
        For each label $[r] \in \{genuine, fake\}$:
            Select **up to 5 images** from each coin type for the training set.
            Assign the remaining images from each coin type to the test set.
Model Initialization and Preprocessing:
    Load Pretrained Model and preprocess the images into a format suitable for the model.
Training Setup:
    Define Metrics:
        Specify metrics such as accuracy, precision, recall, F1 score.
    Set Training Arguments:
        Specify batch size, number of epochs, learning rate, etc.
    Initialize Trainer:
        Initialize the trainer with the model, datasets, and metric functions.
Training and Evaluation:
    Train Model:
        Train the model using the training set.
        Perform evaluation at defined intervals.
    Save Model and training state once training is completed.
Evaluation and Logging:
    Evaluate on Test Set:
        Use the trained model to evaluate the test set.
        Log accuracy, precision, recall, F1 score, and confusion matrix.
    Return final results F_o, containing evaluation metrics and confusion matrix details.

Figure 18: The pseudo-code of the ViT Coin Authentication Algorithm

5.2 Results and Discussion

In this section, we analyze and discuss the results obtained from the ViT model, focusing on its performance in detecting counterfeit coins.

As detailed in Table 5, the ViT model generally outperforms all other methods, particularly in challenging scenarios involving coins with subtle variations or worn-out features. This superior performance is primarily due to the Vision Transformer's self-attention mechanism, which captures long-range dependencies and contextual information across the entire image. Unlike the PrFA method, which relies on predefined rules and patterns and struggles to generalize to broader contexts, the ViT model's ability to learn hierarchical representations directly from raw pixel data enables it to generalize effectively across different coin types and conditions. Additionally, the ViT model's adaptability in handling various data types and its capacity to learn from a relatively small dataset make it a highly suitable choice for counterfeit detection tasks where data scarcity is a concern.

The PrFA method, on the other hand, leveraged augmentation techniques, which can effectively increase the size and variability of the training set. This is beneficial for methods like PrFA that rely on rule extraction, as the increased data diversity allows for more robust fuzzy association rules that can capture a wider range of feature combinations. However, this reliance on augmented data might have also introduced specific patterns or features that do not exist in the real world. Consequently, while PrFA's performance may be enhanced on the augmented dataset, it might not generalize as well to unaugmented or real-world data, where such artificial patterns are absent. This could explain instances where PrFA performs well on specific datasets, such as the 20 Kroner 1991 set, but may not maintain the same level of accuracy across other datasets.

Further demonstrating its robustness, Table 7 presents the ViT model's performance metrics on Chinese coins, including accuracy, precision, recall, and F1-score. The results underscore the model's effectiveness in handling diverse datasets, including coins from different

countries and historical periods.

Table 5: Performance accuracies obtained for methods by Sharifi Rad et al. (PrFA) [23], Bavandsavadkouh et al. [27], Hmood and Suen [28], Liu et al. [22], and the proposed Vision Transformer method.

Method	[22]	[28]	[27]	[23]	Proposed
20 Kroner 1990	92.9	90.0	98.0	93.2	100
20 Kroner 1991	96.6	95.6	97.1	97.5	96.1
20 Kroner 1996	98.4	99.5	99.7	99.9	100
20 Kroner 2008	99.6	93.4	99.6	99.8	99.6

Table 6: Detailed assessment of Precision, Recall, and F-Score for the ViT and the PrFA method across different coin types and years.

	PrFA			ViT		
Datasets	Prec.	Rec.	F-Score.	Prec.	Rec.	F-Score.
20 Kroner 1990	0.843	1.000	0.915	1.000	1.000	1.000
20 Kroner 1991	0.970	0.995	0.978	0.961	0.951	0.953
20 Kroner 1996	1.000	1.000	1.000	1.000	1.000	1.000
20 Kroner 2008	0.995	1.000	0.998	0.996	0.996	0.995

Table 7: Performance Metrics of ViT for Counterfeit Detection on Chinese Coins (Aggregate Data for Years 1912, 1921, 1923, 1927, and 1934)

Dataset	Precision	Recall	F-Score	Accuracy
Chinese	98.0	96.0	96.6	96.0

Table 8 illustrates the performance metrics for the Chinese coin dataset across different years using the ViT model. These performance values were obtained with varying numbers of training and testing samples, ranging from 2 to 31 for training and 4 to 33 for testing, depending on the year of the coin. It is important to note that a lower number of training samples, particularly when fewer than 5 samples were used, consistently resulted in poorer performance, as evidenced by lower accuracy and F1-scores. This is most evident in the results for the Yuan 1912 and Yuan 1934 coins, where limited training data led to an F1-score as low as 0.333. In contrast, higher-performing years such as 1921 and 1927 benefited from a more substantial number of training samples. These findings underscore the importance of an adequate number of training samples to achieve reliable performance in counterfeit coin detection.

Table 8: Performance metrics for the Chinese dataset across different years.

Dataset	Metric			
	Accuracy	Precision	Recall	F-Score
Yuan 1912	0.50	0.80	0.50	0.457
Yuan 1921	0.909	0.826	0.909	0.865
Yuan 1923	0.857	0.734	0.857	0.791
Yuan 1927	0.833	0.875	0.833	0.828
Yuan 1934	0.50	0.25	0.50	0.333

In summary, the Vision Transformer-based approach for counterfeit coin detection has demonstrated superior performance compared to existing state-of-the-art methods across the entire Danish dataset. The results highlight the model's ability to effectively differentiate between genuine and counterfeit coins, even with a limited dataset. This robustness, coupled with the model's ability to capture both local and global features, underscores its potential as a powerful tool in combating counterfeit currency. These findings suggest that ViTs could be widely adopted in practical settings where accurate and reliable counterfeit detection is critical.

Chapter 6

Conclusion and Future Works

This thesis focused on addressing the increasingly significant challenge of counterfeit coin detection, a task with serious implications for the integrity of monetary systems and the stability of financial markets. The core objective was to develop a robust system capable of accurately discerning genuine coins from counterfeit ones, while addressing the challenges posed by limited data availability and the need for precise authentication of intricate details.

6.1 Conclusion

We have presented an in-depth exploration of counterfeit coin detection using two advanced methodologies: the multimodal GPT-4 model and the Vision Transformer (ViT) model.

Multimodal Models (GPT-4):

We investigated the potential of the multimodal GPT-4 model, which integrates both visual and textual modalities, to discern genuine coins from counterfeit ones. This exploration included the application of various prompting techniques—Zero-Shot, Few-Shot,

Chain-of-Thought, and Generated Knowledge—each offering unique advantages in different contexts. Despite the innovative approach, our findings revealed that while GPT-4 shows promise in multimodal tasks, it faces inherent challenges in processing and analyzing specialized visual data like that required for counterfeit detection.

The overall accuracy of 56.13% across all prompting methods highlights limitations in GPT-4's ability to discern intricate visual details, which are critical for distinguishing between authentic and counterfeit coins. One of the key reasons for this lower performance is the fact that GPT-4 is a general-purpose model. It was not explicitly designed or optimized for high-precision visual tasks like counterfeit coin detection, which demands the ability to identify fine details such as subtle variations in minting or wear patterns. GPT-4's visual capabilities, while valuable in certain multimodal contexts, lack the specialized training needed to focus on these fine-grained features.

Additionally, GPT-4 is not specifically pre-trained for tasks like coin authentication, where domain-specific knowledge is crucial for accurate classification. The visual tasks it can handle are more general, meaning that it struggles when applied to domains requiring highly detailed analysis. These limitations underscore the necessity for further refinement of multimodal models in highly specialized visual domains, particularly where precise visual distinctions are critical.

Vision Transformer Approach:

In contrast, the Vision Transformer model, pre-trained on a large-scale dataset and fine-tuned for coin authentication, achieved a remarkable accuracy of 99.31%. The success of the ViT model can be attributed to its sophisticated architecture, which excels in capturing intricate patterns and details through self-attention mechanisms. This architecture enabled the model to focus on essential features such as minting details and wear patterns, crucial for distinguishing between genuine and counterfeit coins. Additionally, the

model's ability to transfer learned representations from the pre-training phase significantly enhanced its generalization capacity, leading to superior performance in detecting counterfeit coins. This performance notably surpasses previous methods, including the state-of-the-art Pruned Fuzzy Associative Classifier (PrFA) by Sharifi Rad et al. [23], in terms of precision, recall, and F1-score, across multiple coin types and years.

6.2 Future Works

Building on the findings of this research, several avenues for future work can be explored to further advance the field of counterfeit coin detection. One promising direction is the refinement and extension of the GPT-4 multimodal model. Given the observed limitations in its current form, future efforts could focus on enhancing the model's training with more comprehensive and up-to-date datasets. Moreover, OpenAI recently introduced the ability to fine-tune GPT-4, allowing developers to tailor the model more closely to specific tasks by training it on domain-specific datasets. Fine-tuning GPT-4 on our coin authentication dataset would enable the model to better recognize the intricate patterns and features critical for counterfeit coin detection. This process could significantly improve its ability to analyze visual data with higher precision and accuracy.

In addition, the exploration of additional prompting techniques could yield further improvements in performance. By combining different prompting strategies or integrating feedback loops, where initial predictions are iteratively refined, the accuracy of the GPT-4 model in discerning counterfeit coins could be substantially improved.

Another potential area for future research is the development of hybrid models that combine the strengths of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) [44, 45, 48]. In the standard implementation of ViTs, images are divided into non overlapping patches, which are then processed as input tokens. This method allows ViTs to capture global context effectively but may miss finer local details due to the non-overlapping nature

of the patches. On the other hand, CNNs, with their overlapping convolutional filters, are adept at capturing local spatial details and fine-grained features. By integrating overlapping patches or using CNNs in conjunction with ViTs, a hybrid model could leverage the best of both worlds—combining the global context understanding of ViTs with the local feature extraction capabilities of CNNs. Such a hybrid approach could improve the robustness and accuracy of counterfeit coin detection by ensuring that both global patterns and subtle local details are considered, making it particularly effective for tasks where both aspects are critical.

Finally, expanding the dataset to include a wider variety of coins from different regions and time periods could help generalize the models further, making them more applicable to a global context. Additionally, incorporating real-world challenges such as varying lighting conditions, different angles of coin placement, wear and tear, and other environmental factors into the dataset could provide more realistic testing scenarios. This, in turn, would enhance the practical applicability of the proposed methods, ensuring that they perform well in diverse and challenging real-world situations.

By pursuing these future research directions, it is hoped that the advancements in counterfeit coin detection presented in this thesis can be built upon, leading to more accurate, efficient, and globally applicable solutions for safeguarding the integrity of monetary systems worldwide.

References

[1] Global News, "Fake Toonies Are Circulating in Quebec and Ontario. Here's How to Spot Them," Nov. 29, 2023, Available: https://globalnews.ca/news/10108612/fake-toonie-quebec-ontario-how-to-spot-it/

[2] "A Newbie-Friendly Guide to Transfer Learning," Oct. 12, 2021, Available: https://www.v7labs.com/blog/transfer-learning-guide

[3] P. Bhavsar. "An ultimate guide to transfer learning in nlp," 2021, Available: https://www.topbots.com/transfer-learning-in-nlp/

[4] OpenAI, "GPT-4 Vision Guide," Available: https://platform.openai.com/docs

[5] "Keyence, 3D Scanner VR-6000," Available: https://www.keyence.ca/products/3d-measure/3d-scanner/vr-6000/

[6] "Prompt Engineering Guide, Prompting Techniques," Available: https://www.promptingguide.ai/techniques

[7] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. S. Albahri, B. S. N. Al-dabbagh, M. A. Fadhel, M. Manoufali, J. Zhang, A. H. Al-Timemy, et al. "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions,

tips, and applications," Journal of Big Data, Springer, vol. 10, pp. 46, 2023, doi: 10.1186/s40537-023-00727-2.

[8] "Deutsche Bundesbank, Considerably more counterfeits in circulation," Jan. 29, 2024, Available: https://www.bundesbank.de/en/press/press-releases/considerably-more-counterfeits-in-circulation

[9] Y. Furuya, T. Ishida, I. Fukuda, and G. Yoshizawa. "Method and apparatus for sorting coins utilizing coin-derived signals containing different harmonic components," Google Patents, US4971187, Nov.1990.

[10] M. Xu, S. Yoon, A. Fuentes, and D. S. Park. "A Comprehensive Survey of Image Augmentation Techniques for Deep Learning," Pattern Recognition, Elsevier, vol. 137, pp. 109347, 2023, doi: 10.1016/j.patcog.2023.109347.

[11] M. Xu, S. Yoon, A. Fuentes, and D. S. Park. "A survey on image data augmentation for deep learning," Journal of Big Data, Springer, vol. 6, pp. 1–48, 2019, doi: 10.1186/s40537-019-0197-0.

[12] Y. Song, T. Wang, P. Cai, S. k. Mondal, and J. P. Sahoo. "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," ACM Computing Surveys, ACM New York, vol. 55, pp. 1–40, no. 13, 2023, doi: 10.1145/3582688.

[13] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. "Generalizing from a few examples: A survey on few-shot learning," ACM Computing Surveys (csur), ACM New York, vol. 53, pp. 1–34, no. 3, 2020, doi: 10.1145/3386252.

[14] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C.

Xu, Y. Xu, et al. “A survey on vision transformer,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, pp. 87–110, no. 1, 2022, doi: 10.1109/TPAMI.2022.3152247.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020, doi: 10.48550/arXiv.2010.11929.

[16] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. “Language models are few-shot learners,” Advances in Neural Information Processing Systems, Curran Associates, Inc., vol. 33, pp. 1877–1901, 2022.

[17] J. Howard and S. Ruder. “Universal language model fine-tuning for text classification,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, vol. 1, pp. 328–339, 2018, https://aclanthology.org/P18-1031

[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova. “Bert: Pre-training of deep bidirectional transformers for language understanding,” North American Chapter of the Association for Computational Linguistics, 2019, https://api.semanticscholar.org/CorpusID:52967399

[19] B. Dong, P. Zhou, S. Yan, and W. Zuo. “Self-promoted supervision for few-shot transformer,” European Conference on Computer Vision, Springer Nature Switzerland, vol. 13680, pp. 329–347, Oct. 2022, doi: 10.1007/978-3-031-20044.

[20] M. Tresanchez, T. Pallejà, M. Teixidó, and J. Palacin. “Using the Optical Mouse

Sensor as a Two-Euro Counterfeit Coin Detector," Sensors, vol. 9, no. 9, pp. 7083–7096, 2009, https://www.mdpi.com/1424-8220/9/9/7083

[21] A. K. Hmood and C. Y. Suen. "Statistical edge-based feature selection for counterfeit coin detection," Multimedia Tools and Applications, Springer, vol. 79, pp. 28621–28642, 2020, doi: 10.1007/s11042-020-09447-8.

[22] L. Liu, Y. Lu, and C. Y. Suen. "An image-based approach to detection of fake coins," IEEE Transactions on Information Forensics and Security, vol. 12, no.5, pp. 1227–1239, 2017, doi: 10.1109/TIFS.2017.2656478.

[23] MS. Rad, S. Khazaee, and C. Y. Suen. "A framework for image-based counterfeit coin detection using pruned fuzzy associative classifier," Expert Systems with Applications, Elsevier, vol. 249, pp. 123577, 2024, doi: 10.1016/j.eswa.2024.123577.

[24] MS. Rad, S. Khazaee, L. Liu, and C. Y. Suen. "A Blob Detector Images-Based Method for Counterfeit Coin Detection by Fuzzy Association Rules Mining," International Conference on Pattern Recognition and Artificial Intelligence, Springer, pp. 669–684, 2020, doi: 10.1007/978-3-030-59830-3.

[25] S. Khazaee, MS. Rad, and C. Y. Suen. "Detection of counterfeit coins based on 3D height-map image analysis," Expert Systems with Applications, Elsevier, vol. 174, pp. 114801, 2021, doi: 10.1016/j.eswa.2021.114801.

[26] S. Khazaee, MS. Rad, and C. Y. Suen. "Detection of counterfeit coins based on modeling and restoration of 3D images," Computational Modeling of Objects Presented in Images. Fundamentals, Methods, and Applications: 5th International Symposium, Niagara Falls, NY, USA, Springer, pp. 178–193, 2017, doi: 10.1007/978-3-319-54609-4.

[27] I. Bavandsavadkouhi, S. Khazaee, and C. Y. Suen. "An Autoencoding Method for Detecting Counterfeit Coins," Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, pp. 292–301, 2022, doi: 10.1007/978-3-031-23028-8.

[28] A. K. Hmood and C. Y. Suen. "An ensemble of character features and fine-tuned convolutional neural network for spurious coin detection," Frontiers in Pattern Recognition and Artificial Intelligence, World Scientific, pp. 169–187, 2019, doi: 10.1142/9789811203527.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems, Curran Associates, Inc., vol. 25, 2012.

[30] J. Maurício, I. Domingues, and J. Bernardino. "Comparing vision transformers and convolutional neural networks for image classification: A literature review," Applied Sciences, MDPI, vol. 13, no. 9: 5521, 2023, doi: 10.3390/app13095521.

[31] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy. "Do vision transformers see like convolutional neural networks?," Advances in Neural Information Processing Systems, vol. 34, pp. 12116–12128, 2021.

[32] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[33] K. He, J. Liu, A. Liu, X. Lu, S. Welleck, P. West, R. l. Bras, Y. Choi, and H. Hajishirzi. "Generated knowledge prompting for commonsense reasoning," Association for Computational Linguistics, pp. 3154–3169, 2021, https://aclanthology.org/2022.acl-long.225

[34] Z. Zhang, A. Zhang, M. Li, and A. Smola. "Automatic chain of thought prompting in large language models," The Eleventh International Conference on Learning Representations, 2022, doi: 10.48550/arXiv.2210.03493.

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. v. Le, D. Zhou, et al. "Chain-of-thought prompting elicits reasoning in large language models," Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc., vol. 35, no. 14: 1800, pp. 24824–24837, 2024.

[36] T. Kojima, S. S. Gu, M. Reid, A. Smola, Y. Matsuo, and Y. Iwasawa. "Large language models are zero-shot reasoners," Advances in Neural Information Processing Systems, vol. 35, pp. 22199–22213, 2022, doi: 10.48550/arXiv.2205.11916.

[37] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. "A Comprehensive Survey on Transfer Learning," Proceedings of the IEEE, vol. 109, no. 1, pp. 43–76, 2021, doi: 10.1109/JPROC.2020.3004555.

[38] K. Weiss, T. M. Khoshgoftaar, and D. Wang. "A survey of transfer learning," Journal of Big data, Springer, vol. 3, pp. 1–40, 2016, doi: 10.1186/s40537-016-0043-6.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention is all you need," Advances in Neural Information Processing Systems, Curran Associates, Inc., vol. 30, 2017, doi: 10.48550/arXiv.1706.03762.

[40] OpenAI, "GPT-4 Research," Mar. 14, 2023, Available: https://openai.com/index/gpt-4-research/

[41] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[42] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. "Minigpt-4: Enhancing vision-language understanding with advanced large language models," The Twelfth International Conference on Learning Representations, 2024.

[43] OpenAI, "Models Overview," Available: https://platform.openai.com/docs/models

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, pp. 10012–10022, 2021, doi: 10.1109/ICCV48922.2021.00986.

[45] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu. "CMT: Convolutional Neural Networks Meet Vision Transformers," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12175–12185, 2022, doi: 10.1109/CVPR52688.2022.01186.

[46] Y. Guo and Z. Wan. "Performance Evaluation of Multimodal Large Language Models (LLaVA and GPT-4-based ChatGPT) in Medical Image Classification Tasks," 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), IEEE Computer Society, pp. 541–543, 2024, doi: 10.1109/ICHI61247.2024.00080.

[47] D. Ono, D. W. Dickson, and S. Koga. "Evaluating the efficacy of few-shot learning for GPT-4Vision in neurodegenerative disease histopathology: A comparative analysis with convolutional neural network model," Neuropathology and Applied Neurobiology, Wiley Online Library, vol. 50, no. 4, pp. e12997, 2024.

[48] 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.

[49] OpenAI, "Enterprise privacy," 2024, Available: https://openai.com/enterprise-privacy/

[50] OpenAI, "Security and privacy," 2024, Available: https://openai.com/security-and-privacy/

[51] OpenAI, "Privacy policy," 2023, Available: https://openai.com/policies/privacy-policy/