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

Dina Omidvar Tehrani[1,2][0009−0007−2403−6058] and Ching Y. Suen[1,3][0000−0003−1209−7631]

[1] CENPARMI, Concordia University, 2155 Guy St, Suite ER1243, Montreal, H3H 2L9, Quebec, Canada
[2] d_omidvar@live.concordia.ca
[3] suen@encs.concordia.ca

Abstract. The proliferation of counterfeit coins poses a substantial threat to the integrity of monetary systems and the stability of financial markets. With advanced counterfeiting techniques, these fraudulent coins can closely mimic genuine ones, complicating the detection process. Addressing this challenge effectively requires robust methods capable of discerning minute differences between genuine and fake coins. In this paper, we tackle the problem of counterfeit coin detection by introducing a dataset comprising high-resolution images of both Danish and Chinese coins, categorized into genuine and counterfeit sets across multiple years. To address the detection task, we employ two approaches: a Vision Transformer (ViT) model and a multimodal GPT-4 model. The ViT model was selected for its ability to capture fine-grained visual patterns even with limited training data, while GPT-4 was incorporated to explore the performance and limitations of large multimodal models (LMMs) in visually complex tasks. Our results show that the ViT model outperforms previous methods and the state-of-the-art in terms of accuracy and robustness. Although the accuracy of GPT-4 was modest, its evaluation provides valuable insight into the current capabilities and constraints of such models in visual reasoning.

Keywords: Counterfeit coin detection · Vision Transformer · GPT-4 · Prompting techniques · Multimodal model

1 Introduction

In the realm of numismatics and the broader financial world, counterfeit coins pose a significant threat, undermining the trust and reliability of global currency systems. The advent of advanced technologies, including the use of intricate alloys and precision machinery, has enabled counterfeiters to produce fakes that are increasingly challenging to distinguish from genuine ones. Consequently, this rise in counterfeit coins affects not only collectors and investors but also has broader economic implications by destabilizing monetary values and damaging the reputation of minting institutions. Therefore, addressing the challenge of counterfeit coins is imperative not only for financial security but also for preserving our

heritage and ensuring the trustworthiness of monetary systems. To address the challenge of counterfeit coin detection, researchers across multiple disciplines, notably computer science, are vigorously exploring innovative solutions. Given the scarcity of data typical in this domain, the focus is on developing methods that maximize accuracy with minimal data [1]. Techniques range from image processing to deep learning, analyzing both 2D and 3D images of coins to train sophisticated models. These approaches enable the identification of counterfeit coins, leveraging advanced algorithms to discern even the subtlest discrepancies.

Given the previously highlighted data scarcity in the counterfeit coin detection task, this challenge has posed an obstacle to the accuracy of existing methods. This scarcity complicates the task of training deep learning models, which typically rely on varied and extensive datasets for optimal learning. Counterfeit detection's reliance on nuanced details further complicates dataset augmentation through common techniques like Generative Adversarial Networks (GANs) [2, 3], noise injection, cropping, and flipping, etc [4, 5]. These methods risk introducing features that could be mistaken for counterfeiting indicators or, conversely, obscure such indicators. Therefore, the focus shifts to employing strategies that promise high accuracy with minimal data. Among these, few-shot learning [6, 7] stands out as a promising approach. It offers a pathway to discerning counterfeit coins by leveraging the power of machine learning with significantly smaller datasets, thus circumventing the limitations posed by traditional dataset augmentation methods. Moreover, it's crucial to maintain image details throughout their processing. Our dataset comprises high-dimensional images, with resolutions about 3550 pixels in both length and width. This highlights the necessity for methods that can handle high-resolution data effectively without degrading image quality. Such attention to detail is critical in the field of counterfeit coin detection, as the ability to discern minor nuances is essential for the accurate identification of forgeries.

In this study, we address the challenges of limited data and high-resolution requirements by exploring two complementary paths. First, we fine-tune a Vision Transformer model [8, 9, 10] for counterfeit coin detection. Second, we conduct an exploratory analysis using the multimodal GPT-4 [4] model to benchmark the performance and limitations of current-generation large multimodal models (LMMs) in fine-grained visual classification tasks. To support learning from limited examples, we adopt few-shot learning techniques in both approaches [11, 12, 13]. The first method utilizes the transfer learning capabilities of a pretrained vision transformer model from Hugging Face [5], tailored to our specific dataset. On the other hand, the second approach leverages the capabilities of the multimodal model GPT-4 to assess coin authenticity. Additionally, this research was supported by a dataset we compiled, featuring high-resolution images of genuine and counterfeit Chinese and Danish coins, captured with a Keyence 3D scanner [6]. Our findings indicate that while ViT significantly outperforms ex-

[4] GPT-4 vision model: https://platform.openai.com/docs/guides/vision.
[5] https://huggingface.co/docs/models.
[6] Keyence 3D scanner: https://www.keyence.ca/3d-scanner/vr-6000/.

isting solutions in this domain, the GPT-4 evaluation provides insight into the current limitations of LMMs when applied to visually nuanced tasks. To guide the reader, the remainder of this paper is structured as follows. Section 2 reviews the methods proposed in the field of counterfeit coin detection so far, including the current state-of-the-art. Section 3 outlines the materials and methods, including data collection and visual analysis, the application of the Vision Transformer and GPT-4 models, and the experimental settings. Section 4 presents and discusses the results, with attention to practical deployment considerations. Finally, Section 5 concludes the paper by summarizing key findings.

2 Related work

The field of forgery analysis, particularly in counterfeit coin detection, has seen notable advancements recently. Considerable attention has been given to analyzing various image processing techniques, focusing specifically on edge detection and feature recognition [14, 15]. Several studies have highlighted the importance of data quality, whether it involves two-dimensional or three-dimensional images [16], the volume of samples, and the application of data augmentation methods [19]. Despite these developments, there are still inherent limitations in these approaches, including low performance, processing times, and sample bias, which continue to challenge the efficacy of these methods in practical applications. Nonetheless, other approaches leveraging machine learning and deep learning techniques, such as Pruned Fuzzy Associative Classifiers (PrFA) [15], Convolutional Neural Networks (CNN) [18], and Autoencoders [17], have demonstrated effectiveness in accurately detecting counterfeit coins.

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's [18] approach included the use of an ensemble technique that combined outcomes from three classifiers. In another method, Khazaee et al.'s [16] study utilized 3-D height-map imaging, which has shown potential by identifying more distinct features and helping to alleviate some of the issues inherent in 2-D image classification techniques.

Additionally, by decomposing grayscale images of coins into the SMG channel and augmenting the dataset with a specialized GAN, Khazaee et al. [19] enhanced the capability of fine-tuned CNNs for counterfeit coin detection. Their hybrid method could even be applied to those types of coins whose genuine or fake counterparts have never previously been seen by the model. Furthermore, the innovative counterfeit detection methodology proposed by Bavandsavadkouhi et al. [17] 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, Sharifi Rad et al. [15] 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.

The primary focus of the studies mentioned above has been on augmenting the number of data samples and subsequently training neural networks on these datasets. While the concept of learning from limited data is well-established in pattern recognition, to the best of our knowledge, few-shot learning techniques have not yet been applied to counterfeit coin detection—particularly using Vision Transformers and high-resolution coin imagery. This gap motivated our investigation and highlights the novelty of our approach.

3 Materials and Methods

This section outlines the materials used and the methodologies employed to develop and evaluate our counterfeit coin detection system. We begin by detailing the dataset composition and acquisition process, including the types of coins and scanning equipment utilized. Next, we describe the design and implementation of two complementary detection approaches — one based on a Vision Transformer (ViT) classifier and the other leveraging the multimodal capabilities of GPT-4 through various prompting techniques. Finally, we explain the experimental setup, including how the models were trained and tested, and how we structured the evaluations to assess performance under different prompting strategies.

3.1 Data collection

As highlighted earlier, a crucial aspect of detecting counterfeit coins is having access to a varied dataset. In this work, we developed a dataset that includes both fake and genuine coins, focusing on Danish and Chinese examples (Fig. 1). Specifically, we scanned a collection of 732 Danish and 112 Chinese coins using a Keyence 3D scanner (Table 2). This 3D scanner captures full surface data across the coins with a resolution of 0.1 µm, and its rotational scanning feature greatly expands the system's measurement capabilities. These coins were collected from reliable sources, including the Danish law enforcement agency and various coin exhibitions.

The scanning process was notably efficient, with the Keyence 3D scanner requiring approximately 10 seconds to scan each coin. The larger the coin, the longer it takes to scan. This resulted in scanned images with a resolution of about 3550 pixels in both length and width. To ensure uniformity and high quality of the images, we conducted all scans under consistent lighting conditions and handled each coin carefully. After collecting the data, we took further steps to standardize the dataset. Specifically, all images were rotated in one direction and converted to grayscale to ensure consistency.

Dataset Analysis and Visual Statistics To further assess the diversity and visual complexity of our dataset, we conducted a quantitative analysis of key

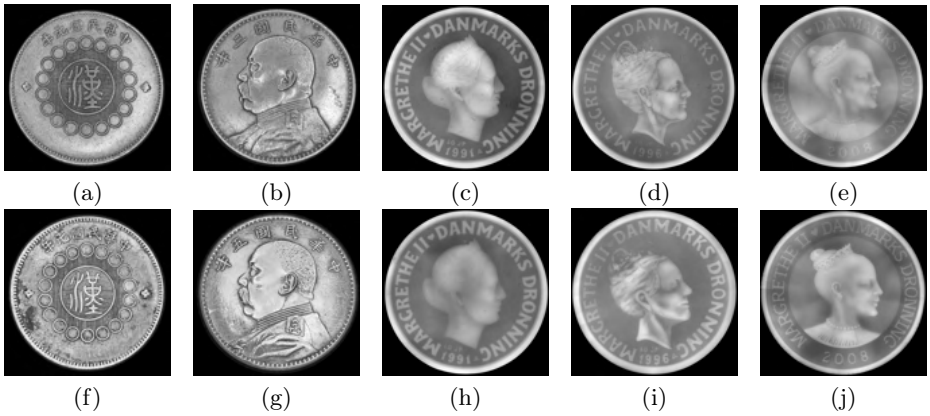


Fig. 1. Comparative display of genuine (top row) and counterfeit (bottom row) Danish and Chinese coins. This collection features paired images illustrating genuine and counterfeit versions of specific coins: (a) & (f) a 1912 Chinese coin, (b) & (g) a 1921 Chinese coin, (c) & (h) a 1991 Danish coin, (d) & (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.

image properties, including pixel intensity, contrast, texture entropy, and sharpness. These statistics help characterize the subtle differences between genuine and counterfeit coins, which are often not obvious to the naked eye.

Fig. 2 presents the distribution of mean pixel intensity across the dataset. As shown, genuine coins tend to cluster around a narrower brightness range, while fake coins exhibit wider variation, indicating inconsistent scanning or surface characteristics.

Fig. 3 compares the Laplacian sharpness of coin images by class. Genuine coins consistently appear sharper, while fake coins include both blurry and overly sharp outliers. This supports the hypothesis that counterfeit coins may have surface irregularities or lower minting fidelity.

Table 1. Summary statistics of image features by class. Values are reported as mean ± standard deviation.

Label	Mean Intensity	Std Dev	Entropy	Sharpness
Genuine	115.87 ± 7.82	74.48 ± 3.44	0.301 ± 0.018	51.67 ± 21.46
Fake	112.39 ± 14.49	69.98 ± 6.14	0.321 ± 0.031	41.35 ± 21.43

Table 1 summarizes the mean, standard deviation, entropy, and sharpness for both classes. These statistics demonstrate that counterfeit coins are not only more visually diverse but also less consistent — justifying the use of robust, high-capacity models like Vision Transformers. These findings provide strong

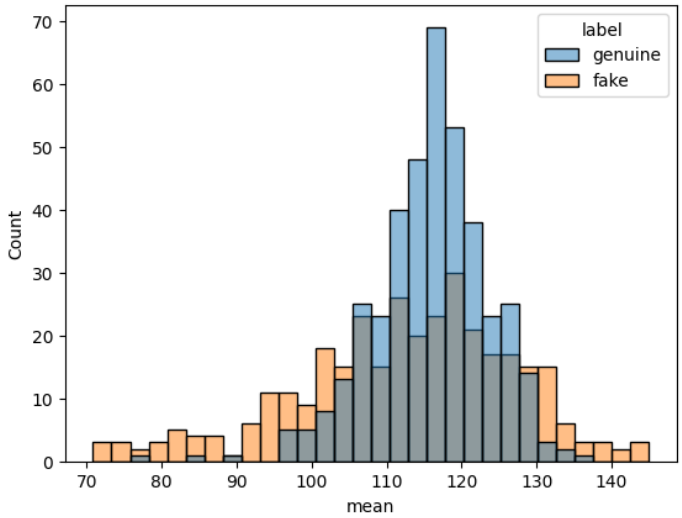


Fig. 2. Histogram of mean grayscale pixel intensity per image for genuine and counterfeit coins.

evidence that the dataset contains meaningful and non-trivial differences, further motivating the use of deep models capable of capturing such variations.

Table 2. 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

To avoid any potential data leakage, the dataset split was performed at the coin instance level. Specifically, for each coin type (combination of country, year, denomination, and authenticity), we randomly selected up to 5 coins for training, and all remaining coins of that type were assigned to the test set. This ensured that no images of the same physical coin appeared in both training and test sets. The selection process was applied independently to the Danish and Chinese datasets. Furthermore, the coins were scanned individually under varying conditions, and duplicate or near-duplicate images were excluded prior to model training.

The test set for the Chinese coins was originally highly imbalanced (Table 2), with 4 genuine and 52 counterfeit samples. To address this and ensure a fair evaluation across both classes, we incorporated several established imbalance-

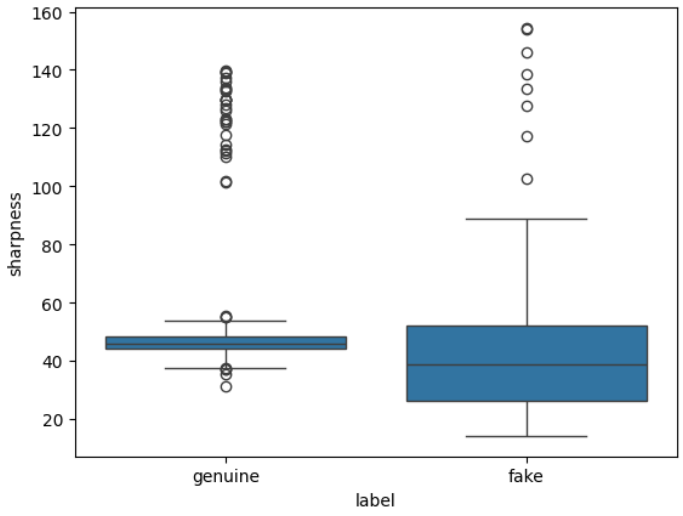


Fig. 3. Laplacian-based sharpness of genuine and counterfeit coins.

handling techniques directly into the model training process. Specifically, we calculated class weights based on the training data distribution and applied them in the loss function to penalize misclassifications of the minority "genuine" class more heavily. We further employed partial oversampling of genuine coins [28] in the training batches to increase their representation without replicating them to the point of overfitting. In addition, we used a focal loss formulation [27], which dynamically scales the loss to focus more on hard-to-classify samples while retaining stability for well-classified ones. These modifications were implemented in the same few-shot learning framework and hyperparameter setting used in our main experiments, ensuring comparability while improving robustness to class imbalance.

To evaluate the performance of the multimodal system GPT-4, we specifically designed the dataset to be suited for this task. For this purpose, we created triplets of coin images, which were used as input prompts during each evaluation session with the multimodal system, as illustrated in Fig. 4. Each triplet was strategically composed by placing the most similar genuine and counterfeit coins — determined by cosine similarity — next to a test coin. The system was then tasked with assessing the authenticity of the test coin based on the comparison with the other two coins in the triplet.

3.2 ViT-Based Method and Exploratory Benchmarking

In this study, we examine two independent and complementary approaches for counterfeit coin detection: (i) a Vision Transformer (ViT)-based classifier and (ii) a prompting-based evaluation using the multimodal capabilities of GPT-4.

Fig. 4. Example of a Coin Image Triplet Used in Multimodal System Evaluations: From left to right: Genuine coin, counterfeit coin, and test coin.

These models were used separately and not integrated as a combined pipeline. The ViT model was fine-tuned using supervised learning to classify coin images, while GPT-4 was used as an exploratory tool to assess visual reasoning capabilities using image triplets and prompting techniques. The goal was to compare traditional deep learning against emerging multimodal models in the context of few-shot learning.

Vision Transformer-based Classifier. Vision Transformers (ViT) have rapidly gained prominence in the field of computer vision, challenging the dominance of traditional Convolutional Neural Networks (CNNs) [20, 21]. Unlike CNNs, which rely on local receptive fields and weight sharing, Vision Transformers apply the principles of self-attention, allowing them to focus on the most relevant parts of an image regardless of their spatial location. This ability to model long-range interactions between pixels makes ViT particularly effective for tasks requiring an understanding of the global context within images [8, 9]. Originating from advances in natural language processing, ViT segments an image into fixed-size patches, treats these patches as tokens similar to words in a sentence, and processes them through multiple layers of self-attention. Pre-trained on extensive image datasets like ImageNet-21k, which comprises 14 million images across 21,000 classes, these models can be fine-tuned to achieve remarkable accuracy in specific vision tasks, making them highly versatile and powerful tools for image classification, object detection, and beyond. For this study, we fine-tuned the Hugging Face Vision Transformer model using a subset of the dataset. The model processes input images through a series of self-attention layers, ultimately producing a 1280-dimensional feature vector. This vector is then passed through a classification head to predict whether the input coin is genuine or counterfeit. We used a few-shot learning approach where a maximum of five images per coin type (year and authenticity) were used for training to assess the generalization capabilities under data-scarce conditions.

Multimodal Model for Coin Authentication (GPT-4). We also conducted an exploratory analysis using the multimodal capabilities of GPT-4 to assess the feasibility and limitations of current LMMs in fine-grained visual reasoning.

Rather than positioning GPT-4 as a competing method, our aim was to use it as a benchmark to understand how well a large generalist model can reason about visual similarity in specialized tasks like counterfeit detection. Using a triplet-based prompting framework, we probed whether GPT-4 could identify visual patterns and distinctions between genuine and counterfeit coins. This experiment highlights the boundaries of current multimodal model capabilities in handling detailed and domain-specific visual classification.

To evaluate GPT-4's ability in this task, we designed a prompt-based evaluation using triplets of coin images: a test coin and its two most visually similar counterparts — one genuine and one counterfeit. The selection of these coins was based on feature vectors obtained from each image using a ResNet50 [22] neural network, with cosine similarity metrics used to determine the visual closeness. The multimodal model was then tasked with assessing the authenticity of the test coin, drawing insights from its comparison with the genuine and counterfeit samples. Additionally, various prompting strategies [7] were explored to refine and measure the accuracy of the GPT-4 API [8], an application programming interface provided by OpenAI, in discerning the authenticity of coins. These experiments were designed not to maximize performance, but to explore how far GPT-4 can reason about visual similarity when given contextual cues, a capability not yet fully understood in general-purpose models.

3.3 Experimental setting

To explore the use of a multimodal model for counterfeit coin detection, we conducted experiments to examine a key factor influencing model accuracy: prompting techniques. In investigating the impact of prompting, we evaluated several methods, including Zero-Shot Prompting [26], Few-Shot Prompting [11], Chain-of-Thought Prompting [24, 25], and Generated Knowledge Prompting [23]. For each of the prompting techniques, we measured the accuracy of the model on the entire dataset.

With Zero-Shot Prompting, we asked the model to predict the authenticity of a coin using only the image of that coin. No examples of other coins or their classifications (genuine or counterfeit) were provided, and the model had to make a prediction based solely on the single coin image, as illustrated in Fig. 5.

In Few-Shot Prompting, we provided the model with triplet images along with a prompt that included information about the authenticity of two of these coins. This method allowed the model to use these examples to inform its predictions (Fig. 6). Next, we examined the Chain-of-Thought (CoT) Prompting technique, which claims to enable complex reasoning capabilities through intermediate reasoning steps. We combined CoT with Few-Shot Prompting to see if this approach would yield better results. In this experiment, the prompt is more detailed, encouraging a step-by-step reasoning process, which can help the model break down the problem and reach a conclusion more systematically (Fig. 7).

[7] www.promptingguide.ai/techniques.

[8] For more information on the GPT-4 API, visit https://platform.openai.com/docs/models.

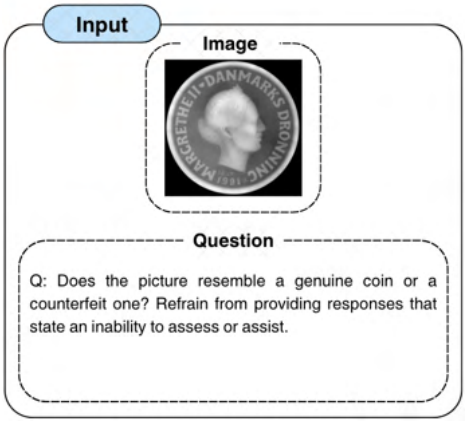


Fig. 5. Example of Zero-Shot Prompting: The model is asked to determine whether the coin in the image is genuine or counterfeit without any prior examples.

Lastly, we explored the Generated Knowledge Prompting technique, which incorporates additional knowledge and information to improve prediction accuracy. We integrated relevant information into the prompts to evaluate whether this would enhance the model's performance, as shown in Fig. 8.

In the first scenario, which utilized zero-shot prompting, the entire dataset was used to test the model. However, the techniques applied in the subsequent three experiments differed. Initially, two subsets of 20 images each were selected: one subset comprised genuine coins, and the other contained counterfeit coins. This dual selection provided a balanced representation of both genuine and counterfeit specimens. For the remaining coins in our dataset, a two-fold comparison was conducted. Each coin was compared against the 20-image subsets of both genuine and counterfeit coins to identify the most similar genuine coin and the most similar counterfeit coin for each test coin. These selected coins were then combined to form a unique set of triplet images. Each image in this new dataset included a test coin from the main set, flanked by its closest genuine and counterfeit counterparts. In addition to the GPT-4 multimodal technique, we employed a pre-trained Vision Transformer model from the Hugging Face library to enhance our counterfeit coin detection system. Specifically, we utilized the ViTForImageClassification model and fine-tuned it to effectively distinguish between genuine and counterfeit coins. To prepare our training dataset, we selected up to five coins from each year and each model, ensuring a generalized representation of different coin types. This approach demonstrated the model's ability to predict authenticity accurately with a limited number of samples. We used the google/vit-huge-patch14-224-in21k model, which features multiple transformer layers that capture intricate patterns and fine-grained details within the coin images. The model processes input images by dividing them into patches, projecting these patches into a high-dimensional space, and applying self-attention

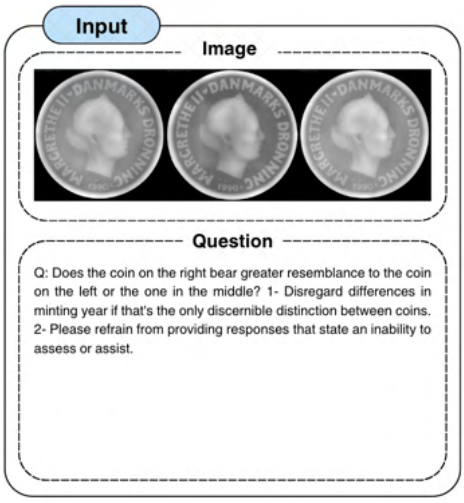


Fig. 6. Example of Few-Shot Prompting: The model is provided with a set of three coins — one genuine, one counterfeit, and one test coin — and is asked to compare the test coin with the provided examples.

mechanisms to learn the relationships between different parts of the image. The model was fine-tuned using a batch size of 4 for up to 10 epochs with the Adam optimizer and an initial learning rate of 1e-4. An early stopping criterion based on validation loss with a patience of 3 epochs was applied to prevent overfitting. Training was conducted on an NVIDIA RTX 3090 GPU. Each experiment was repeated three times with different random seeds, and the observed variance in performance was minimal (standard deviation $< 0.5\%$).

During training, we fine-tuned the model on our dataset using the Trainer API from Hugging Face. The dataset was split with a larger portion reserved for testing to evaluate the model's generalization capabilities. The transformer's self-attention layers effectively captured spatial dependencies, enhancing the model's ability to differentiate subtle variations between genuine and counterfeit coins. This self-attention mechanism allowed the model to focus on critical features such as minting details, wear patterns, and other distinguishing characteristics that indicate a coin's authenticity Fig. 9. Moreover, to evaluate the statistical significance of our results, we compared the accuracy obtained by the proposed approaches with that obtained from previous studies, demonstrating the superiority of our methods.

4 Results and Discussion

In this section, for each of the analyzed methods, we report and discuss the results obtained from the proposed approaches (GPT-4 and ViT) in relation to the current state-of-the-art.

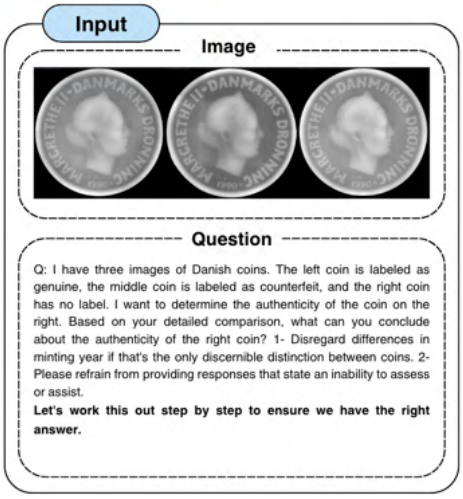


Fig. 7. Example of Chain-of-Thought Prompting: Similar to Few-Shot Prompting, but with additional instructions to encourage step-by-step reasoning.

Vision Transformer-based Classifier. We report the obtained results for the ViT model in Table 3 and Table 4. As is evident in the tables, the accuracy obtained by the state-of-the-art, PrFA, is generally lower than the accuracy of the proposed method, which has an overall accuracy of 99.39 %. This outcome is expected due to the fundamental differences in how these methods process and interpret the image data, as further evidenced by baseline comparisons in Table 6. The Vision Transformer model excels in capturing long-range dependencies and contextual information across the entire image due to its self-attention mechanism. This allows it to process images in a holistic manner, leading to superior performance in image classification tasks. By contrast, the PrFA method relies on detecting local patterns through blob detection and fuzzy association rules. While these techniques can be effective in identifying local features, they may not capture the broader context as comprehensively as the ViT model. Moreover, the ViT model benefits from its ability to learn hierarchical representations directly from raw pixel data, which enables it to adapt to the specific characteristics of the dataset. On the other hand, the PrFA framework uses predefined patterns and rules to classify images, which might not be as adaptable to the complexity and variability present in real-world datasets. As a consequence, the proposed approach significantly outperforms the state-of-the-art. To further support the superiority of the Vision Transformer over traditional methods, we evaluated five baseline classifiers on the same dataset: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), ResNet-18, DenseNet-121 [30], and EfficientNet-B0 [29]. The results, shown in Table 6, indicate that while all these baselines perform reasonably well, they fall short of the ViT's performance. KNN and SVM achieved accuracies of 94.8% and 96.0% respectively, with slightly lower F1-

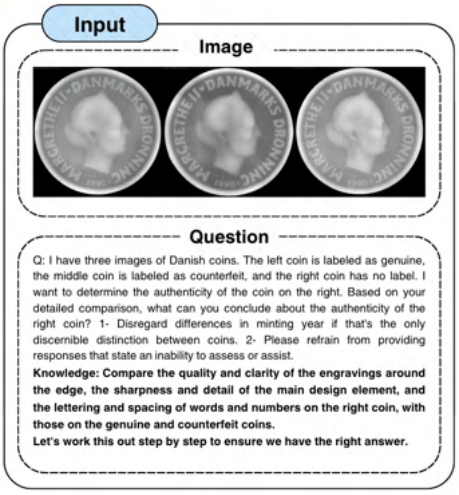


Fig. 8. Example of Generated Knowledge Prompting: The model is prompted to consider specific factors such as font type and size in its assessment of the coin's authenticity.

scores. ResNet-18 and DenseNet-121, two deep convolutional models, performed more competitively with accuracies of 97.0% and 99.0%. However, EfficientNet-B0 underperformed with an accuracy of 93.0%, possibly due to its sensitivity to small domain-specific datasets. All models were outperformed by ViT, which achieved the highest accuracy of 99.39%. These findings further reinforce the advantages of transformer-based models in capturing subtle variations within coin images, particularly under limited training data.

Table 3. Performance accuracies obtained for methods by Sharifi Rad et al. (PrFA) [15], Bavandsavadkouhi et al. [17], Hmood and Suen [18], Liu et al. [14], and the proposed Vision Transformer method.

Method	[14]	[18]	[17]	[15]	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

The results in Table 4, obtained after applying the imbalance-handling techniques, show consistently high performance for both classes in the Chinese dataset. The model achieved a precision of 94.0% and recall of 94.0% for counterfeit coins, and a precision of 90.0% and recall of 88.0% for genuine coins, resulting in macro-averaged precision, recall, and F1-scores of 92.0%, 91.0%, and 91.5%,

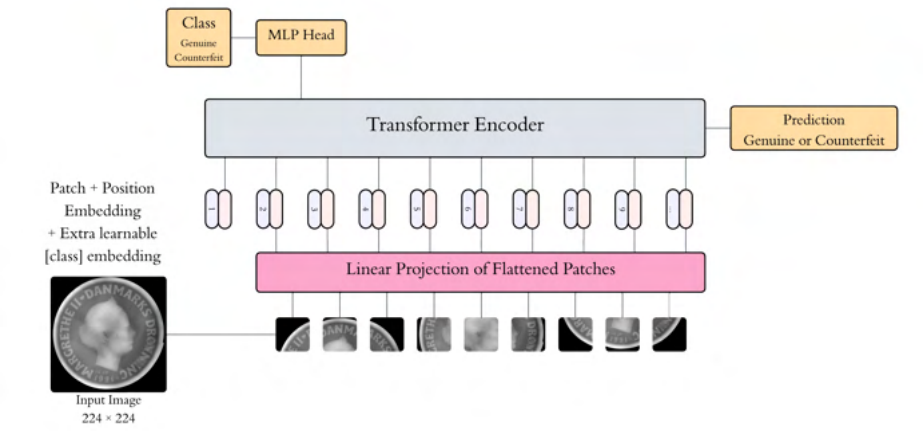


Fig. 9. Architecture of the Vision Transformer model used for coin classification. The model divides each input coin image into patches, embeds positional information, and passes the representations through transformer layers to predict whether the coin is genuine or counterfeit.

Table 4. Performance metrics of ViT for counterfeit detection on Chinese coins (aggregate data for years 1912, 1921, 1923, 1927, and 1934).

Class (Chinese dataset)	Precision (%)	Recall (%)	F1-Score (%)
Fake	94.0	94.0	94.0
Genuine	90.0	88.0	89.0
Macro Avg	92.0	91.0	91.5

respectively. The overall accuracy reached 92.0%, with a balanced accuracy of 91.0%, indicating that the model now predicts genuine and counterfeit coins with comparable reliability. These results demonstrate that the adopted strategies effectively mitigated the class imbalance and provided a fairer, more realistic assessment of model performance on this subset.

GPT-4 Based Method. Rather than being a competitive method, GPT-4 serves here as a diagnostic tool to better understand the gap between general-purpose models and task-specific architectures in visual classification. In this case, the data taken into consideration consists of triplet images obtained in the data processing stage. The performance metrics obtained by the GPT-4 method across different prompting techniques — Zero-Shot (55.32 %), Few-Shot (55.78 %), Chain-of-Thought (56.65 %), and Generated Knowledge (56.79 %) — are reported in Table 5. The overall average accuracy achieved by GPT-4 across prompting strategies was 56.13%, only marginally above chance. While slight variations were observed between the prompting methods, these differences — on the order of 1% — are not statistically significant. This is expected given

that GPT-4 was not fine-tuned on domain-specific coin imagery and relies on its general visual representations learned during pretraining. In such a setting, modifying the textual prompting strategy alone has limited influence because the bottleneck lies in the model's underlying visual encoder, which struggles to capture the subtle, pixel-level differences between genuine and counterfeit coins. Without domain adaptation or targeted visual feature learning, prompt engineering cannot substantially improve classification accuracy, resulting in performance fluctuations that fall within the range of random variability rather than reflecting true methodological gains. These results underscore the current limitations of LMMs like GPT-4 in handling fine-grained visual tasks that require domain-specific visual reasoning.

Table 5. 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

Table 6. Performance comparison of baseline models and the proposed Vision Transformer.

Model	Prec. (F/G)	Rec. (F/G)	F-Score (F/G)	Acc. (%)
KNN (PCA=30)	0.96 / 0.94	0.92 / 0.97	0.94 / 0.95	94.8
SVM (PCA=30)	0.96 / 0.96	0.95 / 0.97	0.96 / 0.96	96.0
ResNet-18	0.95 / 1.00	1.00 / 0.95	0.97 / 0.98	97.0
DenseNet-121	0.97 / 1.00	1.00 / 0.98	0.99 / 0.99	99.0
EfficientNet-B0	0.87 / 1.00	1.00 / 0.88	0.93 / 0.94	93.0
ViT (Ours)	0.997 / 0.992	0.990 / 0.997	0.993 / 0.995	99.4

As shown in Table 6, the ViT model achieves an overall accuracy of 99.39%, outperforming all baseline models. Specifically, it reaches a precision of 0.9966 and recall of 0.9898 for counterfeit coins, and 0.9917 precision and 0.9972 recall for genuine coins, resulting in F1-scores of 0.9932 and 0.9945 respectively. These results confirm the ViT's ability to capture subtle differences in coin texture and design with high reliability, particularly under few-shot training conditions.

The following table provides detailed evaluation metrics (precision, recall, and F-score) for the Vision Transformer model compared to the Pruned Fuzzy Associative Classifier (PrFA) across individual Danish coin types and years.

Table 7. 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

4.1 Practical Deployment Considerations

Beyond academic benchmarks, our proposed framework (ViT) has strong potential for real-world deployment in settings such as automated coin-sorting machines, museum curation systems, and law enforcement verification tools.

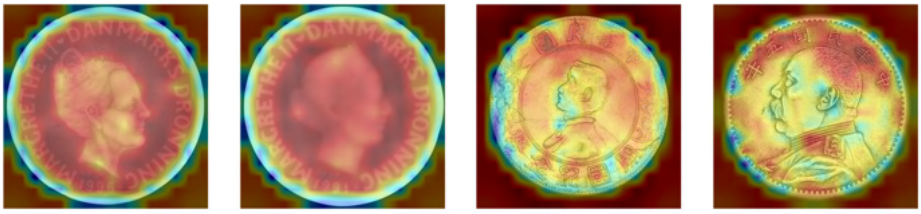
- **Inference Time and Efficiency.** The Vision Transformer model achieves near real-time inference speeds when deployed on a GPU, requiring approximately 18 seconds to process the full test set of 655 coin images (roughly 35 images per second). This throughput is sufficient for high-volume coin authentication scenarios such as automated vending machines or sorting lines.
- **Computational Load Comparison.** While the ViT model is relatively lightweight and can be deployed efficiently on consumer-grade GPUs or embedded systems with acceleration (e.g., Jetson Nano or Coral TPU), the GPT-4-based multimodal analysis introduces significant latency and cloud API dependency. Thus, ViT is more appropriate for edge deployment, while GPT-4 is better suited for offline or batch evaluation.
- **Real-Time Feasibility.** Due to its speed and accuracy, the ViT model is suitable for real-time use in devices with limited power and computing capability. For instance, integrating the model with a Raspberry Pi and camera module could support live authentication of coins inserted into machines, while maintaining over 99% accuracy.

In summary, ViT method is not only accurate, but also deployable across a wide range of practical environments depending on latency, hardware, and interpretability requirements.

4.2 Interpretability Analysis of ViT and GPT-4

To evaluate the interpretability of the proposed ViT-based classification model, we visualized the attention maps from the final transformer layer for selected test samples. Fig. 10 presents examples from both the Chinese and Danish datasets. For correct predictions, the model's attention was concentrated on informative coin regions such as inscriptions, fine engravings, and facial details. This behavior indicates that the model is leveraging relevant high-frequency features for decision-making.

For misclassified samples, attention maps reveal that the model focused on less informative regions or localized artifacts (e.g., lighting reflections, scratches, or border areas). These visualizations highlight the model's ability to capture discriminative patterns, while also exposing limitations where attention is drawn to misleading features. This analysis supports future improvements such as guiding the model's attention towards more meaningful regions using localization constraints or attention regularization.



(a) (b) (c) (d)

Fig. 10. Attention heatmaps from the ViT model highlighting regions most influential in classification decisions. (a) Genuine coin correctly predicted as genuine. (b) Counterfeit coin correctly predicted as counterfeit. (c) Genuine coin misclassified as counterfeit. (d) Counterfeit coin misclassified as genuine. Warmer colors indicate areas of higher model attention, while cooler colors correspond to regions of lower focus.

To evaluate interpretability for the GPT-4–based counterfeit detection Table 7, we analyzed GPT-4's text-based rationales for both correct and incorrect classifications. The model was prompted to describe the visual and textual cues influencing its decisions. For correctly classified samples, GPT-4 often referred to features such as inscriptions, portrait details, or engraving alignment. For example, in a correctly identified counterfeit Chinese coin, it cited inconsistencies in font spacing and character curvature.

In misclassified cases, the rationales frequently indicated reliance on superficial similarities or minor variations present in genuine coins. For instance, a genuine Danish coin was misclassified as counterfeit due to perceived border irregularities that were actually normal for that mint year. This analysis highlights specific reasoning patterns behind GPT-4's outputs and indicates where misinterpretations occur, providing insight into potential areas for prompt refinement.

5 Conclusions

In this paper, we focused on the task of counterfeit coin detection. We collected a dataset consisting of images from both counterfeit and genuine Danish and Chinese coins from various years. To address the challenge of forgery detection, we introduced and evaluated two benchmark approaches:

Table 8. Examples of GPT-4 outputs and rationales compared to ground truth labels.

Sample ID	Prompt to GPT-4	Output	GPT-4 Rationale	Ground Truth
CN_001	"This coin shows a portrait with Chinese inscriptions..."	**Fake**	"Inscriptions font and spacing inconsistent with genuine coins."	Fake
DK_015	"This coin features Queen Margrethe II, dated 1996..."	**Genuine**	"Engraving details and text alignment match genuine samples."	Genuine
CN_047	"This coin depicts a man in profile with surrounding Chinese characters..."	**Genuine**	"The wear pattern and portrait style appear consistent with authentic issues."	Fake
DK_023	"This coin shows a crowned monogram surrounded by intricate designs..."	**Fake**	"Lettering and crown detailing appear inconsistent with known genuine coins."	Genuine

1. Utilization of Multimodal Models (GPT-4): The inclusion of GPT-4 in this study was not intended to compete with or replace domain-specific models such as ViT. Rather, it served as an exploratory diagnostic benchmark to assess the current capabilities and limitations of large multimodal models (LMMs) when applied to fine-grained visual discrimination tasks like counterfeit coin detection. Despite its strong performance in general reasoning and multimodal understanding, GPT-4 achieved only 56.13% accuracy — highlighting a key limitation of current LMMs in handling subtle visual differences without explicit textual context. This outcome is not surprising given that GPT-4 was not fine-tuned on domain-specific coin imagery, and its training likely did not include similar visual patterns. Furthermore, the model's architecture and capabilities are optimized for tasks that involve textual reasoning or multimodal input combinations (e.g., image + question), rather than pure visual classification. Even when provided with few-shot visual examples in the prompt, GPT-4 failed to generalize meaningfully to unseen images. Qualitative observation suggests that the model tends to misclassify coins with minor wear or lighting differences, relying on superficial cues rather than semantically meaningful features. This suggests that GPT-4 currently lacks the ability to focus on high-resolution pixel-level features necessary for tasks like forgery detection, where subtle minting variations define authenticity.

The poor performance of GPT-4 on this task should be seen as an informative failure that underscores a crucial gap in current LMM capabilities. As such, this benchmark provides valuable insight for the research community, highlighting the limitations of deploying general-purpose multimodal models in highly specialized visual domains. Future work in this area may require fine-tuning LMMs on domain-specific datasets or integrating hybrid pipelines combining LMMs with specialized vision models for reliable performance.

2. Vision Transformer (ViT) Approach: We utilized a pre-trained Vision Transformer model from the Hugging Face library, fine-tuning it specifically for the task of coin authentication. Even with a maximum of only 5 coins of each type for training, our fine-tuned ViT model achieved a remarkable accuracy of 99.39 % in distinguishing between genuine and counterfeit coins. This high accuracy, despite the limited training data, can be attributed to several factors.

Firstly, the ViT model's architecture excels in capturing intricate patterns and details through its self-attention mechanisms, allowing it to focus on critical features such as the precise design elements created during the minting process and the specific patterns of wear and tear that coins naturally accumulate over time. Secondly, the model benefits from being pre-trained on a vast dataset, enabling it to leverage learned representations and transfer this knowledge effectively to the task of coin authentication. This transfer learning approach significantly enhances the model's ability to generalize from a small dataset. Consequently, the ViT model outperformed previous methods in most cases, demonstrating its superior performance in this domain.

Also, comparative analysis with the state-of-the-art method, the Pruned Fuzzy Associative Classifier (PrFA) by Sharifi Rad et al., highlighted the strengths of our proposed models. The ViT model, in particular, showed superior performance in terms of precision, recall, and F-measure across most coin types and years. A detailed breakdown of these metrics is provided in Appendix Table A1. Finally, as a suggestion for future work, it is recommended to explore additional prompting techniques for the multimodal GPT-4 model. By investigating and combining various prompting strategies, we can aim to further enhance the model's ability to accurately discern genuine coins from counterfeit ones. Moreover, integrating feedback loops, where the model's initial predictions are refined through iterative prompting, could potentially boost performance.

Bibliography

[1] Alzubaidi, L.; Bai, J.; Al-Sabaawi, A.; Santamaría, J.; Albahri, A.S.; Aldabbagh, B.S.N.; Fadhel, M.A.; Manoufali, M.; Zhang, J.; Al-Timemy, A.H.; et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J. Big Data* **2023**, *10*, 13–439. https://doi.org/10.1186/s40537-023-00727-2.

[2] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. https://doi.org/10.1145/3422622.

[3] Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A.A. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **2018**, *35*, 53–65. https://doi.org/10.1109/MSP.2017.2765202.

[4] Xu, M.; Yoon, S.; Fuentes, A.; Park, D.S. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognit.* **2023**, *137*, 109347. https://doi.org/10.1016/j.patcog.2023.109347.

[5] Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. https://doi.org/10.1186/s40537-019-0197-0.

[6] Song, Y.; Wang, T.; Cai, P.; Mondal, S.K.; Sahoo, J.P. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* **2023**, *55*, 1–40. https://doi.org/10.1145/3582688.

[7] Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv.* **2020**, *53*, 3, 1–34. https://doi.org/10.1145/3386252.

[8] Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. https://doi.org/10.1109/TPAMI.2022.3152247.

[9] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929.

[10] Jamil, S.; Jalil Piran, M.; Kwon, O.-J. A comprehensive survey of transformers for computer vision. *Drones* **2023**, *7*, 287. https://doi.org/10.3390/drones7050287.

[11] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Virtual Conference, 6–12 December 2020; pp. 1877–1901.

https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[12] Dong, B.; Zhou, P.; Yan, S.; Zuo, W. Self-promoted supervision for few-shot transformer. In *Proceedings of the European Conference on Computer Vision (ECCV 2022)*, Springer Nature Switzerland, 2022; pp. 329–347. https://doi.org/10.1007/978-3-031-20044-1_19.

[13] Ren, J.; Li, C.; An, Y.; Zhang, W.; Sun, C. Few-shot fine-grained image classification: A comprehensive review. *AI* **2024**, *5*, 405–425. https://doi.org/10.3390/ai5010020.

[14] Liu, L.; Lu, Y.; Suen, C.Y. An image-based approach to detection of fake coins. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1227–1239. https://doi.org/10.1109/TIFS.2017.2656478.

[15] Rad, M.S.; Khazaee, S.; Suen, C.Y. A framework for image-based counterfeit coin detection using pruned fuzzy associative classifier. *Expert Syst. Appl.* **2024**, *249*, 123577. https://doi.org/10.1016/j.eswa.2024.123577.

[16] Khazaee, S.; Rad, M.S.; Suen, C.Y. Detection of counterfeit coins based on 3D height-map image analysis. *Expert Syst. Appl.* **2021**, *174*, 114801. https://doi.org/10.1016/j.eswa.2021.114801.

[17] Bavandsavadkouhi, I.; Khazaee, S.; Suen, C.Y. An Autoencoding Method for Detecting Counterfeit Coins. In *Proceedings of the Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Springer, 2022; pp. 292–301. https://doi.org/10.1007/978-3-031-23028-8_30.

[18] Hmood, A.K.; Suen, C.Y. An ensemble of character features and fine-tuned convolutional neural network for spurious coin detection. In *Proceedings of Frontiers in Pattern Recognition and Artificial Intelligence*, World Scientific, 2019; pp. 169–187. https://doi.org/10.1142/9789811203527_0010.

[19] Khazaee, S.; Rad, M.S.; Suen, C.Y. Decomposing Relief Maps to Detect Counterfeit Coins Using a Hybrid Deep Learning Method. *Available online:* https://dx.doi.org/10.2139/ssrn.3990636.

[20] Maurício, J.; Domingues, I.; Bernardino, J. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Appl. Sci.* **2023**, *13*, 5521. https://doi.org/10.3390/app13095521.

[21] Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, Virtual Conference, 6–14 December 2021; pp. 12116–12128. https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf.

[22] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, Las Vegas, NV, USA, 27–30 June

2016; pp. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

[23] Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Bras, R.L.; Choi, Y.; Hajishirzi, H.; He, K. Generated knowledge prompting for commonsense reasoning. *arXiv* **2021**, arXiv:2110.08387. https://arxiv.org/pdf/2110.08387.

[24] Zhang, Z.; Zhang, A.; Li, M.; Smola, A. Automatic chain of thought prompting in large language models. *arXiv* **2022**, arXiv:2210.03493. https://doi.org/10.48550/arXiv.2210.03493.

[25] Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA, 28 November–9 December 2022; pp. 24824–24837. https://doi.org/10.48550/arXiv.2201.11903.

[26] Kojima, T.; Gu, S.S.; Reid, M.; Smola, A.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. In *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, LA, USA, 28 November–9 December 2022; pp. 22199–22213. https://doi.org/10.48550/arXiv.2205.11916.

[27] Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* **2017**, 2999–3007. https://doi.org/10.1109/ICCV.2017.324.

[28] Bowyer, K.W.; Chawla, N.V.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *CoRR* **2011**, *abs/1106.1813*. http://arxiv.org/abs/1106.1813.

[29] Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946. https://doi.org/10.48550/arXiv.1905.11946.

[30] Huang, G.; Liu, Z.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2016**, arXiv:1608.06993. https://doi.org/10.48550/arXiv.1608.06993.

Acknowledgements This work was supported by the Natural Sciences and Engineering Research Council of Canada.