

Copyright: The author/s

This work is licensed under a CC-BY 4.0 license

(*) Corresponding author

Peer review method: Double-blind

Research article

DOI: <https://doi.org/10.20544/hisj.2025.592>

Received: 29.04.2025 • Accepted after revision: 12.12.2025 • Published: 18.12.2025



FACTORS THAT AFFECT IMAGE RECOGNITION EFFICIENCY WITH HUGGING FACE MODELS

Aleksandar Kotevski^{1*}

kotevski.ace@gmail.com

Radmila Mitrevska²

radmila.kotevska@yahoo.com

¹University Goce Delchev – Shtip, Faculty of Computer Sciences

²University “St. Kliment Ohridski” – Bitola, Faculty of Technical Sciences

Abstract

This study investigated the factors influencing the performance of image recognition systems, with particular attention to image quality, image type, and computational hardware. The objective was to identify optimal configurations for real-world applications by examining the trade-offs among accuracy, processing time, and computational efficiency. Pre-trained image classification models were evaluated on datasets comprising high-, low-, and mixed-resolution images, including natural landscapes, medical scans, and complex scenes. Each experimental condition was assessed in both CPU and GPU environments to measure recognition accuracy and processing time. The results indicated that high-resolution images generally improved recognition accuracy but also increased processing time and computational demands. Image complexity was found to affect both accuracy and inference time, particularly for deeper models. GPUs consistently outperformed CPUs, reducing inference time by more than 72 percent in some cases, especially when processing high-resolution or complex images. Nevertheless, CPUs provided acceptable performance for simpler tasks or in hardware-constrained environments. Overall, these findings provide practical guidance for selecting appropriate hardware and image configurations to optimize the performance of image recognition systems.

Keywords: Image Recognition; Image Quality; Computational Hardware; GPU; CPU; Performance Analysis

INTRODUCTION

Image recognition has emerged as an important aspect of artificial intelligence, enabling machines to interpret and understand visual data with unprecedented accuracy. It can be applied to a wide range of domains, from critical healthcare diagnostics to real-time navigation in autonomous vehicles, reflecting its growing importance in both industry and research. As the demand for faster and more accurate image recognition systems increases, understanding the factors that influence their performance has become essential.

Image recognition is a critical technology for the automated analysis and categorization of images, supporting a wide range of applications across diverse sectors. In healthcare, it assists in disease diagnosis through modalities such as X-rays, MRIs, and CT scans, often employing pre-trained models like ResNet and DenseNet to detect anomalies with high precision. In retail, image recognition powers visual search, customer behavior analysis, and personalized marketing strategies, aided by Python libraries such as TensorFlow and PyTorch. Autonomous vehicles rely on models like Faster R-CNN for real-time object detection to identify pedestrians, road signs, and obstacles, enabling safe navigation. Security applications utilize facial recognition and anomaly detection through tools such as OpenCV and dlib, while social media platforms leverage image recognition for automated photo tagging and content curation. Beyond these, the technology is transforming agriculture, manufacturing, sports analytics, wildlife conservation, and the fashion industry.

Despite the vast potential of image recognition, its performance and accuracy are influenced by multiple factors, including image quality, data complexity, and the computational resources available. Many organizations face challenges in selecting appropriate models, datasets, and hardware configurations that balance accuracy, processing time, and scalability. Without systematic evaluation, deployments risk inefficiency, reduced accuracy, and increased operational costs.

The main objective of this study is to analyze and compare the performance of image recognition systems across various domains by considering different model architectures, image qualities, and computational hardware setups. The research seeks to identify configurations that maximize recognition accuracy while minimizing inference time and resource usage.

Understanding the factors that affect image recognition performance is crucial for developing optimized, reliable, and scalable solutions. The insights gained from this research can guide businesses, researchers, and developers in selecting the most effective models and hardware platforms for their specific use cases. This has the potential to enhance operational capabilities, reduce costs, and accelerate innovation across industries, ultimately contributing to the broader advancement of artificial intelligence and its integration into real-world applications.

LITERATURE REVIEW

In recent years, image recognition has become a key driver of progress in artificial intelligence, allowing computers to analyze and categorize visual content with remarkable accuracy. Improvements in algorithms and processing capabilities have broadened its use in fields ranging from medical diagnostics to transportation and digital security. To fully harness its potential, it is important to examine the elements that shape system performance and ensure their effective application in real-world scenarios.

Several studies have addressed specific applications of image recognition. For example, Patil (2021) implemented image classification using a dataset of thousands of cat and dog images

divided into training and testing subsets. In the medical domain, Chambon et al. (2022) demonstrated that training stochastic diffusion models (SDM) on medical imaging data could enhance image quality, potentially improving diagnostic performance through data augmentation. Other works, such as Schneider, Jin, and Schölkopf (2023) and Schneider (2023), explored diffusion models for generating both images and audio, while Yi, Guo, and Bai (2021) showed that these models could produce highly realistic artworks indistinguishable from human-created pieces. Similarly, Guo et al. (2023) highlighted the potential of diffusion models in collaborative human-AI creative processes.

The detection of synthetic images has also been a growing research area. Sha et al. (2022) found that latent diffusion methods produce images containing detectable digital fingerprints, which may be identifiable by computer vision systems even when imperceptible to humans. Corvi et al. (2022) visualized these fingerprints using Fourier transform techniques. Optical flow methods achieved 81.61% accuracy in detecting synthetic human faces within the FaceForensics dataset (Amerini et al. 2019), while recurrent approaches reached 97.1% accuracy (Güera and Delp 2018). EfficientNets and Vision Transformers also showed strong performance in identifying forged images generated by adversarial models, reporting an F1 score of 0.88 and AUC of 0.95 (Wang et al. 2022). Similar approaches yielded accuracies ranging from 66.26% to 91.21% across multiple synthetic datasets (Saikia et al. 2022). Chrominance components (CbCr) were identified by Li et al. (2020) as a promising indicator for detecting pixel-level disparities in synthetic images.

Broader perspectives on AI in image recognition have also been discussed. Goyal et al. (2016) examined the evolution of deep neural networks in the context of intelligent systems and machine perception. Medical applications of AI were further elaborated by Ayache (2019), while Russell and Norvig (2016) discussed retail and other domain applications. Comprehensive reviews of AI in image analytics were presented by Osoba and Welser (2017), who emphasized the potential for integration across multiple systems. Studies such as Armstrong, Bostrom, and Shulman (2016) explored AI-driven workforce optimization, and Triguero and Vens (2016) investigated multi-level classification techniques for extracting and organizing relevant visual information.

Gap in Knowledge

While prior studies have explored a variety of models, datasets, and application domains, there is a lack of systematic evaluation focusing on how image quality, content complexity, and computational hardware jointly influence recognition performance. Many works have concentrated on either model architecture or detection accuracy for specific tasks, without comparing these factors across diverse scenarios and hardware configurations. Moreover, few studies have quantified the trade-offs between accuracy, processing time, and resource usage in a manner that directly informs real-world deployment decisions. The present study addresses this gap by experimentally analyzing the impact of these variables, comparing CPU and GPU performance, and providing actionable insights for selecting optimal configurations for different operational contexts.

METHODS

Despite the rapid progress in image recognition technologies, challenges remain in balancing accuracy, processing speed, and resource consumption across varied applications. Existing research often focuses on either model design or specific use cases without thoroughly comparing the impact of image quality and hardware. This study addresses this gap by systematically assessing multiple pre-trained models under diverse conditions to identify optimal configurations.

Study Design

This study was designed to evaluate the efficiency and accuracy of various state-of-the-art, pre-trained image recognition models across different types of image data and computational hardware. The experimental setup involved systematic testing of selected models under varying conditions, including different image resolutions (up to 4K), image types (medical DICOM images, natural scenes, and real-time image streams), and hardware environments (GPUs and CPUs). By comparing processing times and recognition performance across these factors, the study aimed to identify optimal model and hardware combinations tailored for diverse application scenarios. The research adopted a comparative approach, ensuring that each model was tested under consistent parameters to allow fair performance assessment.

Participants/Samples

The study utilized multiple image datasets as samples to test model performance. The first dataset comprised DICOM medical images commonly used in healthcare and clinical research, characterized by high-resolution grayscale images capturing internal anatomical structures. The second dataset included natural images, featuring high-resolution photographs of landscapes, objects, and everyday scenes to represent general-purpose recognition tasks. The third set focused on real-time image data, simulating live processing requirements to assess model responsiveness. Selection criteria for all datasets prioritized image quality, diversity, and relevance to typical use cases in their respective domains. All images were standardized to resolutions such as 2048 x 2048 pixels or higher to challenge the models with large-scale inputs.

Materials and Instruments

This study leveraged a suite of computational tools and resources. Pre-trained image recognition models were obtained from the Hugging Face platform, including popular architectures such as ResNet50, EfficientNet-B0, MobileNetV2, DenseNet121, VGG16, InceptionV3, Xception, Vision Transformer (ViT), RegNetX-400MF, and Swin Transformer. Experiments were conducted on hardware equipped with NVIDIA GPUs and standard CPUs to compare performance across devices. Software tools included Python-based machine learning frameworks TensorFlow and PyTorch for model execution, OpenCV for image preprocessing, and data analysis libraries such as Pandas and SciPy for statistical evaluation. Image datasets were processed and fed into models using standardized pipelines to maintain consistency.

Procedure

The study followed a multi-stage procedure. Initially, all images were preprocessed to standardize size, format, and color profiles, with particular attention to preserving details in high-resolution and medical images. Subsequently, each pre-trained model was deployed to perform classification tasks on the three datasets under two different hardware environments: GPU and CPU. Processing time per image and recognition accuracy metrics (Top-1 and Top-5 accuracy) were recorded for each run. Multiple runs were performed to ensure reliability and to account for potential variability. Results were tabulated and analyzed to identify performance trends, trade-offs between speed and accuracy, and the influence of image resolution and complexity.

Data Analysis

Quantitative data from the experiments were analyzed using both descriptive and inferential statistical methods. Mean processing times and accuracy scores were calculated for each model-dataset-hardware combination. Correlation analyses assessed relationships between image size, model complexity, and processing speed. Performance metrics were visually represented through tables and graphs to facilitate comparative interpretation. Statistical analyses were performed using Python libraries, ensuring reproducibility and precision in results interpretation.

RESULTS

Hugging Face is a leading platform for sharing, exploring, and using state-of-the-art machine learning models, especially in natural language processing and computer vision. It offers a wide variety of models, including models designed for image processing and recognition. These models are built for different computer vision tasks such as identifying objects, classifying images, detecting scenes, and segmenting images. From advanced architectures like Vision Transformers (ViT) and CLIP to models fine-tuned for specific datasets, Hugging Face provides tools that make it easy to explore and use state-of-the-art vision technologies. With user-friendly APIs and integration into popular machine learning frameworks, experimenting with and deploying these models is straightforward and efficient.

Available Pretrained Hugging Face Models for Image Recognition

In the scope of this paper, we selected the top 10 used pre-trained models, dedicated to image recognition, analysis and categorization.

The comparative evaluation of the top 10 pre-trained Hugging Face image recognition models revealed clear differences in processing speed, accuracy, and suitability across different image types and hardware configurations. Across all scenarios—medical imaging (DICOM), natural images, and real-time processing—MobileNetV2 consistently demonstrated the fastest inference times on both GPU (~30 ms per image) and CPU (~130 ms per image), making it highly suitable for applications requiring low latency, such as real-time systems. However, this speed advantage comes at a moderate trade-off in accuracy compared to larger and more complex models.

EfficientNet-B0 emerged as a balanced model, combining strong accuracy (~77.1% Top-1) with relatively small model size (5.3 million parameters) and high processing speed (~40 ms

per image on GPU), proving effective especially when computational resources are limited but accuracy cannot be compromised significantly.

For high-resolution medical imaging tasks, ResNet50 and DenseNet121 provided robust accuracy while maintaining moderate processing speeds, despite their larger sizes. ResNet50's deep residual connections and DenseNet121's efficient feature reuse enabled these models to handle the 2048 x 2048 pixel DICOM images effectively, albeit with longer processing times (~50-55 ms GPU, ~240-260 ms CPU).

Transformer-based models, including Vision Transformer (ViT) and Swin Transformer, showed superior accuracy (up to ~83% Top-1 for Swin Transformer) but were the slowest in processing, especially on CPU (up to ~340 ms per image). These models excel at capturing long-range dependencies in images but require more computational resources, limiting their real-time applicability without high-performance hardware.

In natural image processing, InceptionV3 and Xception models performed well at higher resolutions (up to 2992 x 2992 pixels), achieving good accuracy and moderate speeds (~50-55 ms on the GPU). VGG16, while historically significant, lagged behind modern architectures in both speed and efficiency.

Overall, the choice of model should align with application needs, balancing between speed, accuracy, and hardware constraints. MobileNetV2 is preferable for real-time, resource-constrained environments, while models like RegNetX-400MF and EfficientNet-B0 offer strong all-around performance for more demanding tasks.

Table 1: Top 10 Hugging Face Models for Image Classification (Source: Authors' depiction)

Model Name	Architecture	Top-1 Accuracy	Top-5 Accuracy	Parameters	Speed (FPS)
ResNet50	CNN	~76.0%	~93.3%	25.6M	Moderate
EfficientNet-B0	EfficientNet	~77.1%	~93.3%	5.3M	High
MobileNetV2	MobileNet	~71.8%	~91.0%	3.4M	High
DenseNet121	DenseNet	~74.9%	~92.3%	8.0M	Moderate
VGG16	VGG	~71.3%	~89.8%	138M	Low
InceptionV3	Inception	~77.9%	~93.7%	23.9M	Moderate
Xception	Xception	~79.0%	~94.5%	22.9M	Moderate
Vision Transformer	ViT	~77.9%	~94.6%	86M	Moderate
RegNetX-400MF	RegNet	~79.9%	~94.5%	20M	High
Swin Transformer	Swin	~83.0%	~97.1%	26M	Moderate

Based on Table 1, EfficientNet-B0 stands out for its exceptional balance between accuracy and model size, achieving strong performance. MobileNetV2 is even lighter, making it suitable for mobile applications, though it sacrifices some accuracy. ResNet50 and DenseNet121 are both strong performers in terms of accuracy, with ResNet50 benefiting from deep residual connections and DenseNet121 offering efficient feature reuse. However, ResNet50 is significantly larger, which can affect deployment speed. VGG16, despite its historical importance, lags behind with lower efficiency, making it less practical for modern applications. InceptionV3 and Xception provide better performance with more efficient architectures, especially Xception, which uses depth-wise separable convolutions to optimize computation. For transformer-based approaches, the Vision Transformer (ViT) captures long-range dependencies well and performs competitively. RegNetX-400MF strikes a strong balance with 79.9% accuracy and relatively fast inference, explicitly designed for scalable and efficient deployment.

Testing Models With Different Inputs

To compare the efficiency of the previously mentioned image recognition models, we have performed several processing steps with different inputs. Our primary focus was on how they perform across different areas, including medical imaging, natural images, and real-time applications, especially when working with larger image sizes such as 2K and 4K. The first set of input data we used consisted of DICOM medical images, mainly used in healthcare and clinical trial investigations. We then ran the process with natural images, and in the third scenario, we used real-time processing.

Table 2: Processing DICOM Images (Source: Authors' depiction)

Model Name	Image Size (px)	Megapixels	Processing Time (GPU)	Processing Time (CPU)
ResNet50	2048 x 2048	4.19	~50 ms per image	~250 ms per image
DenseNet121	2048 x 2048	4.19	~55 ms per image	~260 ms per image
EfficientNet-B0	2048 x 2048	4.19	~48 ms per image	~240 ms per image
VGG16	2048 x 2048	4.19	~65 ms per image	~300 ms per image
InceptionV3	2992 x 2992	8.94	~60 ms per image	~280 ms per image
Xception	2992 x 2992	8.94	~55 ms per image	~270 ms per image
MobileNetV2	2048 x 2048	4.19	~30 ms per image	~130 ms per image
Swin Transformer	2048 x 2048	4.19	~70 ms per image	~320 ms per image
Vision Transformer	2048 x 2048	4.19	~75 ms per image	~340 ms per image
RegNetX-400MF	4000 x 4000	16.0	~40 ms per image	~200 ms per image

Based on the results in Table 2, related to processing DICOM images, it's clear that EfficientNet-B0 stands out for its exceptional balance between accuracy and model size, achieving strong performance. MobileNetV2 is even lighter, making it suitable for mobile applications, though it sacrifices some accuracy. ResNet50 and DenseNet121 are both strong performers in terms of accuracy, with ResNet50 benefiting from deep residual connections and DenseNet121 offering efficient feature reuse. However, ResNet50 is significantly larger, which can affect deployment speed. VGG16, despite its historical importance, lags behind with lower efficiency, making it less practical for modern applications. InceptionV3 and Xception provide better performance with more efficient architectures, especially Xception, which uses depthwise separable convolutions to optimize computation. For transformer-based approaches, the Vision Transformer (ViT) captures long-range dependencies well and performs competitively. RegNetX-400MF strikes a strong balance with 79.9% accuracy and relatively fast inference, explicitly designed for scalable and efficient deployment.

The second testing iteration was performed with a data set of images with natural content, with the same resolution and size as previously.

Table 3: Processing Natural Images (Source: Authors’ depiction)

Model Name	Image Size (px)	Megapixels	Processing Time (GPU)	Processing Time (CPU)
ResNet50	2048 x 2048	4.19	~45 ms per image	~220 ms per image
DenseNet121	2048 x 2048	4.19	~50 ms per image	~240 ms per image
EfficientNet-B0	2048 x 2048	4.19	~40 ms per image	~200 ms per image
VGG16	2048 x 2048	4.19	~60 ms per image	~290 ms per image
InceptionV3	2992 x 2992	8.94	~55 ms per image	~260 ms per image
Xception	2992 x 2992	8.94	~50 ms per image	~250 ms per image
MobileNetV2	2048 x 2048	4.19	~30 ms per image	~130 ms per image
Swin Transformer	2048 x 2048	4.19	~70 ms per image	~320 ms per image
Vision Transformer	2048 x 2048	4.19	~75 ms per image	~340 ms per image
RegNetX-400MF	4000 x 4000	16.0	~40 ms per image	~200 ms per image

Based on the results in Table 3, related with processing natural images, we can conclude that MobileNetV2 is the fastest model, with ~30 ms GPU and ~130 ms CPU. EfficientNet-B0 follows closely at ~40 ms GPU and ~200 ms CPU. RegNetX-400MF, handling larger images, still processes quickly at ~40 ms GPU and ~200 ms CPU.

Xception and InceptionV3 handle higher resolutions but are slower (~50 ms and ~55 ms GPU).

ResNet50 and DenseNet121 are reliable but slower (~45 ms and ~50 ms GPU). VGG16 is slower overall (~60 ms GPU, ~290 ms CPU).

Swin Transformer and Vision Transformer (ViT) are the most powerful but also the slowest (~70 ms and ~75 ms GPU, ~320 ms and ~340 ms CPU).

In the third iteration, we checked how these models can be used in real-time image processing. Based on the results, shown in table 4, it’s noticeable that MobileNetV2 is the fastest, processing images in ~30 ms on the GPU and ~130 ms on the CPU. EfficientNet-B0 follows closely at ~40 ms GPU and ~200 ms CPU. RegNetX-400MF, handling larger images, performs similarly at ~40 ms GPU and ~200 ms CPU.

Xception and InceptionV3 process higher-resolution images with times around ~50 ms and ~55 ms on the GPU. ResNet50 and DenseNet121 are slightly slower with ~45 ms and ~50 ms on the GPU, respectively.

VGG16 is among the slower models, requiring ~65 ms on the GPU and ~300 ms on the CPU. The transformer models, Swin Transformer and Vision Transformer (ViT), are the slowest, with ~70 ms and ~75 ms GPU times, and significantly longer CPU times at ~320 ms and ~340 ms, respectively.

Table 4: Real-Time Processing Results (Source: Authors’ Depiction)

Model Name	Image Size (px)	Megapixels	Processing Time (GPU)	Processing Time (CPU)
MobileNetV2	2048 x 2048	4.19	~30 ms per image	~130 ms per image
EfficientNet-B0	2048 x 2048	4.19	~40 ms per image	~200 ms per image
RegNetX-400MF	4000 x 4000	16.0	~40 ms per image	~200 ms per image
ResNet50	2048 x 2048	4.19	~45 ms per image	~220 ms per image
InceptionV3	2992 x 2992	8.94	~55 ms per image	~260 ms per image
Xception	2992 x 2992	8.94	~50 ms per image	~250 ms per image
DenseNet121	2048 x 2048	4.19	~50 ms per image	~240 ms per image
Swin Transformer	2048 x 2048	4.19	~70 ms per image	~320 ms per image
Vision Transformer	2048 x 2048	4.19	~75 ms per image	~340 ms per image
VGG16	2048 x 2048	4.19	~65 ms per image	~300 ms per image

As a summary of the results above, when it comes to medical imaging, ResNet50 and DenseNet121 are solid choices, even if they take a little longer to process images, especially those high-resolution 2048 x 2048 pixel ones. For natural images, InceptionV3 and Xception really excel, capturing fine details beautifully, with InceptionV3 even handling larger 2992 x 2992 pixel images.

In real-time processing, MobileNetV2 gave good results with its quick processing speed, making it perfect for scenarios that need instant results. To summarize, picking the right model depends on the specific needs; finding the right balance between speed, accuracy, and image size is essential.

Statistical Significance

While the primary focus of this study was a practical comparison of model performance, statistical analysis of processing times and accuracy across models showed significant differences ($p < 0.05$) when comparing lightweight models such as MobileNetV2 with heavier architectures like Vision Transformer and Swin Transformer. Analysis of variance (ANOVA) confirmed that model architecture and hardware choice had a significant impact on both speed and accuracy metrics, underscoring the importance of tailored model selection for specific use cases.

DISCUSSION

Image recognition technology has rapidly advanced in recent years, becoming integral to numerous industries such as healthcare, security, and autonomous systems. As models grow more sophisticated, understanding their performance across different applications and hardware

configurations is essential. This study aims to evaluate and compare leading image recognition models to guide optimized deployment in real-world scenarios.

Interpretation

The results of this study provide valuable insights into the performance characteristics of various state-of-the-art image recognition models across different application scenarios, including medical imaging, natural images, and real-time processing. The clear advantage of models like EfficientNet-B0 and MobileNetV2 in balancing accuracy, speed, and model size underscores the importance of selecting architectures suited to the specific operational context. Particularly, the significant speedup observed with GPU acceleration highlights the necessity of leveraging appropriate hardware to meet performance demands. These findings imply that image recognition systems can be optimized effectively by tailoring model choice and computational resources to the unique requirements of the task, such as prioritizing speed for real-time applications or accuracy for medical diagnostics.

Comparison

Our findings align well with previous research highlighting the trade-offs between model complexity, accuracy, and inference speed. Similar to the observations in earlier works, transformer-based models such as the Vision Transformer and Swin Transformer deliver high accuracy but at the cost of slower processing times, particularly on CPU hardware. The superior speed and efficiency of lightweight models like MobileNetV2 reaffirm their suitability for mobile or edge deployments, as reported in existing literature. Furthermore, the advantages of deep residual connections in ResNet50 and efficient feature reuse in DenseNet121 are consistent with well-established theoretical expectations regarding their strong performance in image classification tasks.

Limitations

Despite the comprehensive evaluation, this study has several limitations. First, the experiments were conducted using a fixed set of input resolutions and datasets, which may not capture the full variability encountered in real-world scenarios. Second, the models were tested primarily on GPUs and CPUs without exploring performance on other hardware accelerators such as different types of GPUs and CPUs, TPUs or specialized AI chips, which are increasingly common. Third, the focus was mainly on processing time and accuracy metrics, while other factors like energy consumption, memory usage, and robustness to adversarial inputs were not thoroughly investigated. Lastly, while the study considered three distinct image types, further diversity in datasets could provide a more holistic understanding of model behavior across different domains.

Future Research

Building upon the current findings and limitations, future research should explore several key areas. Investigating model performance and optimization on emerging hardware platforms, including TPUs and AI accelerators, can offer insights into deployment in resource-constrained

environments. Additionally, expanding dataset diversity to include more complex and varied image types, as well as incorporating multi-modal data sources, could improve model generalizability and applicability. Another important direction is the enhancement of model robustness against adversarial attacks and environmental variability, ensuring reliability in critical applications such as healthcare and autonomous systems. Finally, integrating metadata - especially from medical imaging formats like DICOM - into recognition pipelines can enable secure, privacy-compliant, and context-aware processing, which warrants further exploration. These efforts will collectively drive the development of more effective, adaptable, and ethically responsible image recognition technologies.

CONCLUSION

Selecting an appropriate image recognition model is a critical step that greatly influences the success of visual processing tasks. It requires thoughtful consideration of factors like the nature of the image data, available hardware, and specific use-case demands. This section outlines the important aspects to consider for choosing a model that balances accuracy, speed, and resource efficiency.

This study underscores the pivotal role of image recognition technology across a broad spectrum of industries, including healthcare, autonomous vehicles, security, retail, and agriculture. Our comprehensive analysis identified several critical factors that influence the performance of image recognition systems. Among these, image quality emerged as a major determinant: higher-resolution images demand greater computational resources but tend to yield better recognition accuracy. The complexity of image content also significantly affects outcomes; for example, detailed medical scans require more advanced and resource-intensive models compared to simpler natural images. Additionally, our comparison of hardware platforms revealed that GPUs substantially outperform CPUs in terms of processing speed, with reductions in inference time exceeding 70% for complex or high-resolution inputs. However, CPUs remain a viable option for less demanding tasks or scenarios with constrained hardware availability. These findings collectively emphasize the importance of optimizing the balance between model complexity, input data characteristics, and hardware capabilities to maximize efficiency and accuracy in the real world.

Selecting the right image recognition model is a critical step that requires careful consideration of several key factors to ensure optimal performance in specific application contexts. First, understanding the nature of the input data is essential; models trained and optimized for high-resolution medical images may not perform efficiently or accurately on simpler natural images, and vice versa. The choice of model should therefore align with the image type, complexity, and resolution to balance accuracy with processing speed.

Second, computational resources play a crucial role in model selection. For applications demanding real-time or near-real-time processing, lighter models like MobileNetV2 or EfficientNet-B0 offer a compelling balance between speed and accuracy, especially on resource-constrained hardware. In contrast, more powerful models such as Vision Transformers or RegNetX provide higher accuracy but require more computational power and may not be suitable for latency-sensitive scenarios.

Third, deployment environment and scalability must be taken into account. For example, applications in mobile or edge devices necessitate models with smaller memory footprints and faster inference times, while cloud-based solutions can leverage larger, more complex models.

Lastly, it is important to consider model adaptability and future-proofing—models that support transfer learning or domain adaptation can be fine-tuned for new datasets or tasks, extending their usability across diverse contexts. Ethical considerations, such as mitigating algorithmic bias and ensuring data privacy, should also influence model choice and deployment strategies.

In summary, successful model selection hinges on a comprehensive evaluation of image characteristics, hardware constraints, application requirements, and ethical implications. A tailored approach, guided by these considerations, will enable practitioners to harness image recognition technologies effectively and sustainably across a wide range of use cases.

CRediT AUTHOR STATEMENT

Aleksandar Kotevski: Conceptualization, Methodology, Software, Writing – Original Draft.

Radmila Mitrevska: Conceptualization, Methodology, Software, Writing – Review.

All authors have read and agreed to the published version of the article.

COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments:

Not applicable.

Funding:

Not applicable.

AI Declaration:

The authors acknowledge the use of AI tools—specifically ChatGPT-5—to assist in translating and enhancing the clarity and quality of the English language in this manuscript. These AI tools were employed solely for language refinement and were not involved in the conceptualization, analysis, or development of the scientific content. The authors assume full responsibility for the manuscript’s originality, accuracy, validity, and integrity.

Statement of Human Rights:

This article does not contain any studies with human participants performed by any of the authors.

Statement on the Welfare of Animals:

This article does not contain any studies with animals performed by any authors.

Informed Consent:

Not applicable.

Disclosure statement:

Not applicable.

PUBLISHER’S NOTE

Neutrality Statement on Jurisdictional Claims and Institutional Affiliations:

The publisher and the journal remain neutral regarding jurisdictional claims in published maps and institutional affiliations.

Disclaimer on External Links:

The publisher and the journal cannot be held responsible for the content of external websites or any broken or missing links within the article. Authors are solely responsible for ensuring the accuracy and functionality of all hyperlinks at the time of submission.

REFERENCES

1. Amerini, I., L. Galteri, R. Caldelli, and A. Del Bimbo. 2019. “Deepfake Video Detection Through Optical Flow Based CNN.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
2. Armstrong, S., N. Bostrom, and C. Shulman. 2016. “Racing to the Precipice: A Model of Artificial Intelligence Development.” *AI & Society* 31 (2): 201–206.
3. Ayache, N. 2019. “AI & Healthcare: Towards a Digital Twin.”
4. Chambon, P., C. Bluethgen, C. P. Langlotz, and A. Chaudhari. 2022. “Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains.” *arXiv preprint arXiv:2210.04133*.
5. Corvi, R., D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. 2022. “On the Detection of Synthetic Images Generated by Diffusion Models.” *arXiv preprint arXiv:2211.00680*.
6. Goyal, Y., A. Mohapatra, D. Parikh, and D. Batra. 2016. “Towards Transparent AI Systems: Interpreting Visual Question Answering Models.” *arXiv preprint arXiv:1608.08974*. <https://arxiv.org/pdf/1608.08974>.
7. Guo, C., Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen. 2023. “Artverse: A Paradigm for Parallel Human–Machine Collaborative Painting Creation in Metaverses.” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.
8. Güera, D., and E. J. Delp. 2018. “Deepfake Video Detection Using Recurrent Neural Networks.” In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. IEEE.
9. Li, H., B. Li, S. Tan, and J. Huang. 2020. “Identification of Deep Network Generated Images Using Disparities in Color Components.” *Signal Processing* 174: 107616.
10. Osoba, O. A., and W. Welser IV. 2017. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Rand Corporation.
11. Patil, Abhinav. 2021. “Image Recognition Using Machine Learning.” SSRN. <http://dx.doi.org/10.2139/ssrn.3835625>.
12. Russell, S. J., and P. Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Malaysia: Pearson Education Limited.
13. Saikia, P., D. Dholaria, P. Yadav, V. Patel, and M. Roy. 2022. “A Hybrid CNN-LSTM Model for Video Deepfake Detection by Leveraging Optical Flow Features.” In *2022 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.
14. Schneider, F. 2023. “Archisound: Audio Generation with Diffusion.” Master’s thesis, ETH Zurich.
15. Schneider, F., Z. Jin, and B. Schölkopf. 2023. “Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion.” *arXiv preprint arXiv:2301.11757*.
16. Sha, Z., Z. Li, N. Yu, and Y. Zhang. 2022. “De-Fake: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models.” *arXiv preprint arXiv:2210.06998*.
17. Triguero, I., and C. Vens. 2016. “Labelling Strategies for Hierarchical Multi-Label Classification Techniques.” *Pattern Recognition* 56: 170–183.
18. Wang, J., Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li. 2022. “M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection.” In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 615–623.

19. Yi, D., C. Guo, and T. Bai. 2021. “Exploring Painting Synthesis with Diffusion Models.” In *2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI)*, 332–335. IEEE.