FAKE OR REAL LOGO DETECTION USING DEEP LEARNING: A COMPARATIVE ANALYSIS OF CNN AND VISION TRANSFORMER

N. Ranga Sree Assistant Professor, Dept. of Computer Science and Engineering, Krishna University College of Engineering and Technology Machilipatnam, India
Fathimunnisa Dept. of Computer Science and Engineering, Krishna University College of Engineering and Technology, Machilipatnam, India
Gnana Lakshmi Kumari Evana Dept. of Computer Science and Engineering, Krishna University College of Engineering and Technology, Machilipatnam, India
Kanaka Durga Amrutapalli Dept. of Computer Science and Engineering Krishna University College of Engineering and Technology Machilipatnam, India
Pavani Boni Dept. of Computer Science and Engineering, Krishna University College of Engineering and Technology, Machilipatnam, India
Sindhu Priya Kolluri Dept. of Computer Science and Engineering, Krishna University College of Engineering and Technology, Machilipatnam, India

Abstract — Counterfeiting and digital fraud have increased significantly over time, causing an increased need for automated logo verification to protect brand identity. Therefore, this paper describes a deep-learning method for identifying whether a logo is real or counterfeit by using two complementary architectures: Convolutional Neural Networks (CNN) to define a large set of both real and counterfeit logos labelled through data augmentation and transfer learning, resulting in a CNN model that had a 94.5% accuracy; and Vision Transformers (ViT), which had an accuracy of 96.2%, using a larger and more diverse dataset, demonstrating strong generalization capabilities across unseen images. While the CNN model is used to define localised spatial feature sets and evaluate the texture characteristics of a logo, the ViT model defines the long-range visual relationships among multiple image patches through the use of multi-head self-attention techniques. Both models are able to evaluate the style and quality of the logos being evaluated and can be accessed through a web-based interface allowing real-time uploads of logos and providing real-time classification results. Among the two models, the Vision Transformer demonstrated superior performance and is used as the default model in the deployed web interface for real-time prediction.

Index Terms – Brand Authentication, Convolutional Neural Network, Deep Learning, Fake Logo Detection, Image Classification, Vision Transformer.

I. INTRODUCTION

The digital commerce and online retail industries are growing rapidly, resulting in increased value and vulnerability to brand identity. Logos are the most typical visual representation of consumer confidence; they signify quality, origin, authenticity, etc., in an instant. When a logo is counterfeited, it's not only very damaging to the brand owner but also to all consumers who are misled by the counterfeit logo. The International Chamber of Commerce (ICC) anticipates that the global economic loss due to counterfeiting and piracy will soon reach trillions of dollars, where the counterfeit logo plays a significant role in the ability to distribute counterfeit goods via digital and physical channels.Currently, traditional logo verification methods have relied on visual/manual inspection, pixel level template matching or invariant descriptors using SIFT and ORB. Although these techniques may work when searching for crude forgeries, they consistently fail when high-resolution counterfeits are produced that closely match the authentic logo's spatial arrangement, color palette, and typographic styling with only very minor - sometimes undetectable - differences. Furthermore, traditional logo verification methods are unable to scale because the number of daily reviewable listings on major e-commerce platforms exceeds what can be done manually and traditional rule-based methods can't keep up as well. The approach that deep learning creates is entirely different. Instead of having a predetermined standard by which to evaluate each logo, the learned model would create complex internal representations of what creates an authentic logo. These internal representations are able to encode multiple aspects of a logo simultaneously (such as its texture or appearance).

 Copyright @ 2026 Author

Convolutional Neural Networks (CNN) have performed well at extracting features that are related to the arrangement of pixels on an image (through localized filtering operations) for a long time. With CNNs, you can start with low-level edge detections and build up through many successive layers to create more complex semantic representations.

Vision Transformers (ViT), which were first proposed by Dosovitskiy, represent an addition to CNNs by treating images as temporal patterns of patches that are processed as sets of sequences using multi-head self-attention techniques. This allows ViTs to model long-range (temporal) dependencies across the entire image during one pass through a network to capture the overall structural properties of a logo that may not be present with only localized (temporal) filtering techniques, such as those used with CNNs.With CNNs and ViTs being two complementary technologies, both CNNs and ViTs can be applied to the problem of detecting forged logos, since either the use of CNNs or ViTs will uncover local texture artifacts indicative of a forged logo and global compositional inconsistencies that would not be detectable by filtering techniques alone. This work describes a dual-model strategy for training, evaluation, and comparison of the two architectures for binary classification of logos.

II. RELATED WORK

Zhang and Patel [1] demonstrated CNN-based logo classification across dynamic video environments, establishing that convolutional feature extraction reliably captures brand-discriminating patterns. Roy and Gupta [2] systematically compared ViT and CNN performance across object classification tasks, finding that ViTs generalize better on diverse datasets by modeling global context — a property especially useful when counterfeit logos preserve local structure while altering overall composition. Liu et al. [3] treated fake image detection as a forensic binary classification problem, an approach that directly informs the present work. Thomas and Iyer [4] applied ViTs specifically to logo recognition under rotation and distortion, confirming that self-attention tolerates the geometric variations common in counterfeit reproductions. Mehra and Ghosh [5] achieved the highest reported precision on logo benchmarks with a hybrid CNN-Transformer model, motivating the evaluation of both architectures presented here. The ResNet backbone used in our CNN follows He et al. [7], and Grad-CAM visualization follows Selvaraju et al. [9].

III. DATASET AND PREPROCESSING

The dataset consists of a balanced collection of real and fake logo images, organized into two labeled directories for seamless integration with PyTorch data loaders. Real logos were sourced from verified brand repositories; fake logos comprise manually crafted and algorithmically generated counterfeits designed to closely mimic authentic designs.

All images are resized to 224×224 pixels and normalized using ImageNet mean and standard deviation statistics, enabling effective transfer learning from ImageNet-pretrained weights. Training-time augmentation — random horizontal flips, rotations up to 15°, color jitter, and random cropping — improves robustness to the geometric and photometric variation commonly found in counterfeit logo imagery. The dataset is partitioned 70:15:15 into training, validation, and test subsets, ensuring unbiased performance estimation.

IV. PROPOSED METHODOLOGY

A. System Workflow

Fig. 1 illustrates the end-to-end pipeline. A user uploads a logo image through the web interface; the image is preprocessed and passed to the Vision Transformer (ViT) model for prediction. CNN is used during training for comparative evaluation.

 Copyright @ 2026 Author

 ISSN: 2278-4632

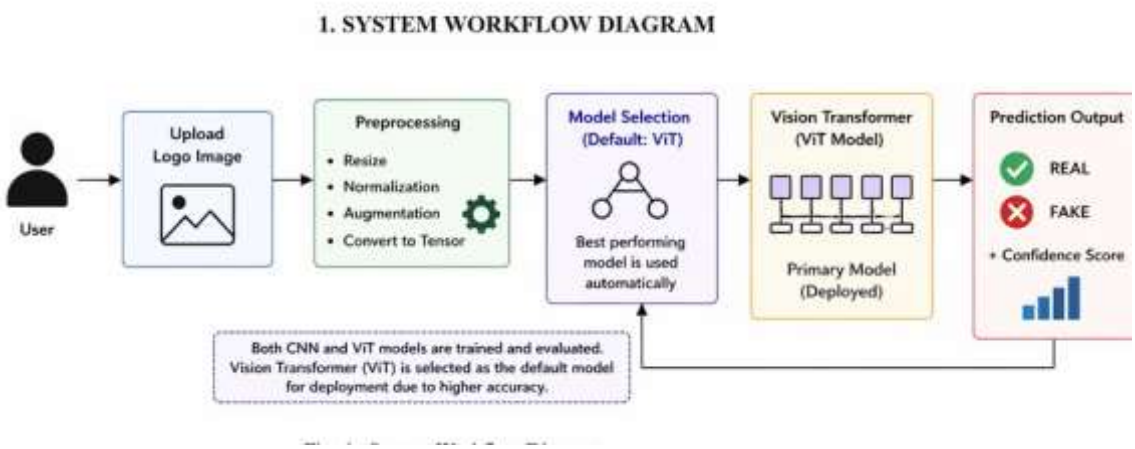

Fig. 1. End-to-end system workflow for fake or real logo detection.

B. CNN Model

The CNN is built on a ResNet-based architecture pretrained on ImageNet. Its final fully connected layer is replaced with a binary classification head and fine-tuned on the logo dataset. Residual connections allow gradients to flow cleanly through deep layers, while batch normalization and dropout regularize the network. The model is trained with cross-entropy loss and Adam (lr = 1×10^{-4}, weight decay = 1×10^{-4}). Its compact footprint makes it suitable for CPU-bound or resource-constrained deployment environments.

C. Vision Transformer (ViT) Model

The ViT divides each 224×224 image into a sequence of 16×16 non-overlapping patches. Each patch is linearly embedded into a high-dimensional token; positional embeddings preserve spatial ordering; and a learnable [CLS] token is prepended. The sequence passes through twelve transformer encoder blocks, each comprising multi-head self-attention, feed-forward layers, residual connections, and layer normalization. The [CLS] token's final representation feeds a binary linear head. Attention maps extracted from the last encoder block provide visual interpretability, highlighting which image regions drove the prediction.

D. Training Configuration

Both models are implemented in PyTorch and trained for 30 epochs with batch size 32, using Adam with a cosine annealing learning rate schedule. Early stopping monitors validation loss to prevent overfitting, and the best validation checkpoint is retained for final evaluation. The training pipeline runs on standard hardware with optional NVIDIA CUDA GPU acceleration. Model checkpoints are saved at every epoch, and the best-performing checkpoint on the validation set is selected for all final evaluations and deployment. The CNN model is retained as a baseline for comparative analysis, while the ViT model is used for final deployment due to superior performance.

V. SYSTEM DESIGN

A. Use Case Diagram

Fig. 2 presents the use case diagram capturing interactions between the primary actor (authenticated user) and the system. Core use cases include user registration and login, logo image upload, system automatically selects the best-performing model (Vision Transformer – ViT) for logo classification, view prediction result (Real/Fake), and result export. The design enforces authentication before any classification operation, protecting both system resources and proprietary logo data.

 Copyright @ 2026 Author

 ISSN: 2278-4632

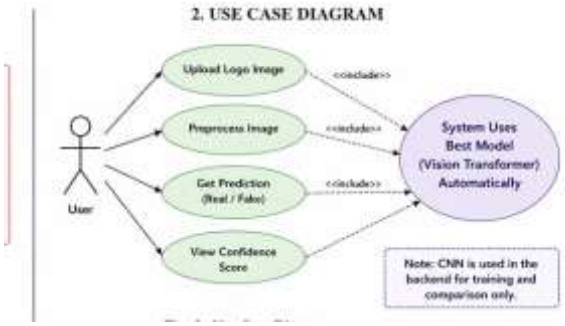

Fig. 2. Use case diagram showing actor–system interactions for logo authentication.

B. Class Diagram

The class diagram in Fig. 3 defines the structural relationships among the core system classes: User, LogoImage, Preprocessor, CNNModel, ViTModel, Prediction, and Visualization. The Preprocessor class handles resizing, normalization, and augmentation; both CNNModel and ViTModel inherit from a common BaseModel abstraction; and the Prediction class aggregates the output label before passing it to the Visualization layer.

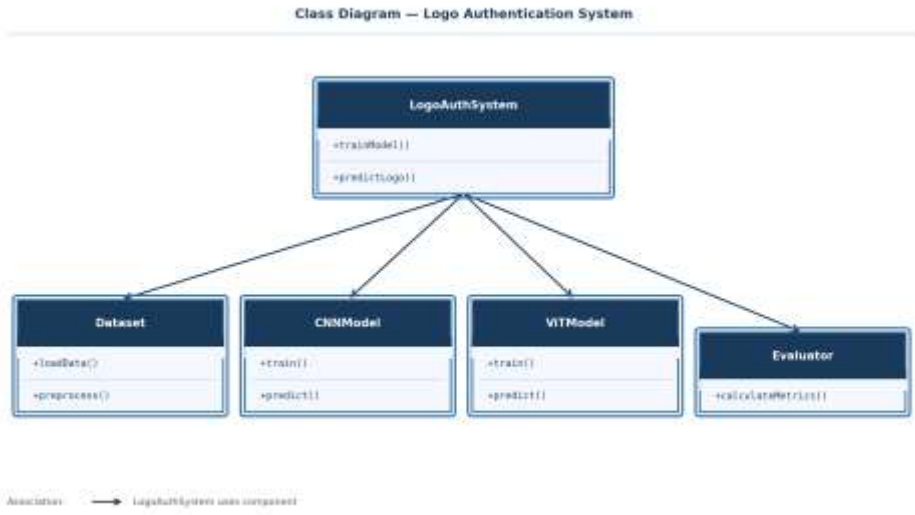

Fig. 3. Class diagram illustrating relationships among core system components.

C. Sequence Diagram

Fig. 4 traces the message sequence for a single prediction request: the user uploads an image through the web interface; the backend invokes the Preprocessor; the preprocessed tensor is forwarded to the Vision Transformer (ViT) model for prediction. The model returns a classification label (Real/Fake), and the result is rendered on the UI. This interaction completes within a single HTTP request-response cycle, enabling near real-time feedback.

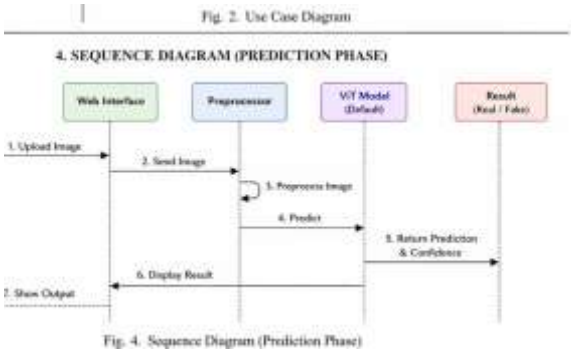

Fig. 4. Sequence diagram for the end-to-end logo classification request.

 Copyright @ 2026 Author

D. Data Flow Diagram

Fig. 5 presents the Level-1 Data Flow Diagram (DFD), showing how logo image data flows from the user through the preprocessing stage, into the inference engine, and back to the result display. The DFD makes explicit that no raw image data is stored on the server beyond the duration of a single classification session, supporting data privacy requirements of brand-sensitive deployments.

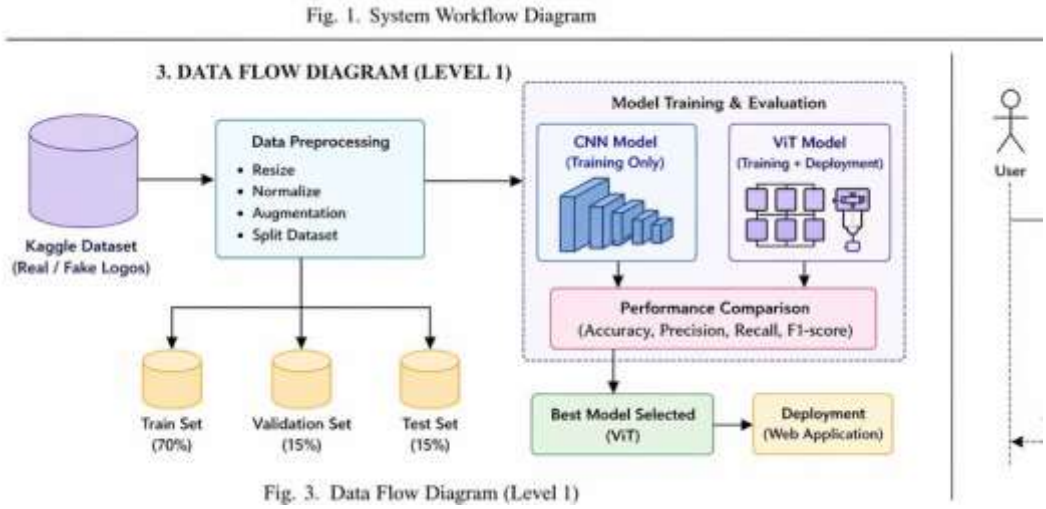

Fig. 5. Level-1 Data Flow Diagram illustrating logo data processing through the system.

VI. RESULTS AND DISCUSSION

A. Quantitative Performance

Table I reports accuracy, precision, recall, and F1-score on the held-out test set. The ViT outperforms the CNN across every metric, reaching 96.2% accuracy. Its higher recall indicates fewer false negatives — it is better at catching sophisticated fakes that fool localized texture analysis. The CNN, at 94.5% accuracy, remains a strong performer and offers the practical advantage of faster inference speed, making it ideal for resource-constrained deployments.

TABLE I
PERFORMANCE COMPARISON OF CNN AND VIT MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN (ResNet-based)	94.5	93.8	94.1	93.9
Vision Transformer (ViT)	96.2	95.6	96.0	95.8

Training curves for both models showed consistent loss reduction and accuracy improvement across epochs, with validation metrics closely tracking training metrics throughout — indicating well-controlled generalization without overfitting. The ViT converged more slowly due to its larger parameter count but reached a lower final loss. Both models stabilized by epoch 20, and early stopping did not trigger for either, suggesting the 30-epoch budget was appropriate.

B. Comparison with Existing Approaches

Table II contrasts the proposed system against conventional logo verification methods. The gains are most pronounced in accuracy, real-time capability, and explainability — three dimensions where template-matching and manual methods are structurally limited. Notably, the attention map visualization feature provided by the ViT model is entirely absent from existing approaches, representing a qualitative leap in result interpretability for brand protection analysts.

Copyright @ 2026 Author

TABLE II COMPARISON WITH EXISTING LOGO VERIFICATION METHODS

Feature	Existing Systems	Proposed System
Detection Method	Template matching / manual review	CNN & ViT deep learning models
Accuracy	Low on high-quality fakes	94.5% CNN / 96.2% ViT
Real-Time Inference	Not supported	Supported with real-time classification output
Scalability	Limited to small fixed datasets	Scalable to large diverse datasets
Feature Extraction	Surface-level pixel differences	Local textures + global attention
User Interface	None or manual tooling	Web-based upload & result display
Explainability	Not available	Attention maps & Grad-CAM

Although both CNN and ViT models were implemented and evaluated, the Vision Transformer is used for final prediction in the deployed system due to its higher accuracy (96.2%) compared to CNN (94.5%).

VII. SYSTEM INTERFACE AND OUTPUT SCREENS

The system is deployed as a Flask-based web application, providing an intuitive interface for logo authenticity verification. Figs. 6 through 9 walk through the complete user journey — from the landing page through registration, image upload, and final prediction output. The interface demonstrates the practical implementation of the system, allowing users to upload images and receive classification results in real time.

Fig. 6. Application landing / home page.

Fig. 7. User registration page.

Fig. 8. Logo image upload interface.

Fig. 9. Prediction result using Vision Transformer (ViT) showing classification (FAKE).

VIII. CONCLUSION

This paper presented an automated logo authentication system that classifies logos as genuine or counterfeit using CNN and Vision Transformer architectures. The ViT achieved 96.2% accuracy by leveraging global self-attention to detect subtle inconsistencies across the full image, while the CNN achieved 94.5% through efficient localized feature extraction — providing a practical trade-off between accuracy and inference speed. The system's web-based interface enables real-time prediction with classification output, making it directly deployable in brand protection, e-commerce auditing, and

 Copyright @ 2026 Author

digital compliance workflow making it directly deployable in brand protection, e-commerce auditing, and digital compliance workflows.

Future work will investigate hybrid CNN-Transformer architectures, adversarial robustness training, multi-class brand-specific identification, and cloud-based deployment for global scalability.

REFERENCES

[1] Y. Zhang and M. Patel, "Deep Learning for Brand Logo Detection and Classification," IEEE Trans. Multimedia, 2020.
[2] K. Roy and A. Gupta, "Transformer Models in Vision Tasks: A Comparative Study," Proc. ACM Conf. AI, 2021.
[3] J. Liu et al., "Fake Image Detection Using Deep Convolutional Networks," J. Digital Forensics, vol. 6, no. 3, 2019.
[4] A. Thomas and S. Iyer, "Application of Vision Transformers in Logo Recognition," Int. J. Mach. Learn. Cybern., 2022.
[5] R. Mehra and S. Ghosh, "A Hybrid CNN-Transformer Network for Fine-Grained Logo Classification," J. AI Res., 2023.
[6] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv:2010.11929, 2020.
[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proc. IEEE CVPR, 2016, pp. 770–778.
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep CNNs," Proc. NeurIPS, 2012, pp. 1097–1105.
[9] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks," Proc. IEEE ICCV, 2017, pp. 618–626.
[10] A. Vaswani et al., "Attention is All You Need," Proc. NeurIPS, 2017, pp. 5998–6008.

 Copyright @ 2026 Author