iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

LEVERAGING HYBRID DEEP ENSEMBLES FOR CUTTING-EDGE IMAGE FORGERY DETECTION WITH ATTENTION-CNN, EFFICIENTNETB0, AND VISION TRANSFORMER

Purna Chandu Challagulla,
Saga Bhuvan Sai
B. Tech Students, Dept. of Computer Science and Engineering,
R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

Dr. Malempati Sreelatha
Professor & Head of Department, Dept. of Computer Science and Engineering,
R.V.R & J.C College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India

ABSTRACT
With the increasing realism of AI-generated images, distinguishing them from real ones has become a growing challenge. Previous studies have used Convolutional Neural Networks (CNNs) for classifying synthetic and real images, but CNNs present several limitations. They often fail to focus on critical regions, depend heavily on increasing depth—raising computational costs—and lack global context understanding due to localized receptive fields. To overcome these issues, this work introduces a hybrid deep learning approach combining three advanced architectures: Attention-CNN with CBAM, EfficientNetB0, and Vision Transformer (ViT). CBAM enhances CNNs with spatial and channel attention, improving focus on key image features like textures and artifacts. EfficientNetB0 applies compound scaling to optimize network depth, width, and resolution for better performance with fewer resources. ViT captures global dependencies by treating images as patch sequences, enabling recognition of subtle, long-range patterns. The proposed ensemble was trained and evaluated on the CIFAKE dataset, which contains both real and AI-generated images. It achieved an accuracy of 96.81%, significantly outperforming standard CNNs. This study highlights the advantages of combining attention mechanisms and transformer-based models for more accurate and efficient synthetic image detection.

Keywords:
Convolutional Neural Networks (CNN), Convolutional Block Attention Module (CBAM), EfficientNetB0, Vision Transformer (ViT).

INTRODUCTION

The rapid advancement of artificial intelligence (AI) in generating synthetic images has introduced both impressive capabilities and serious challenges. Where early AI-generated visuals were often marred by obvious flaws, modern generative models now produce images with such high fidelity that they are virtually indistinguishable from real photographs. This evolution has profound implications for the authenticity of visual data, particularly as these images begin to appear in artistic competitions, media, and online platforms.

One of the most powerful tools enabling this progress is the Latent Diffusion Model (LDM)[1], which allows for the creation of detailed, high-resolution images in seconds. While this innovation has expanded creative possibilities, it also poses ethical and societal risks. The ease with which realistic images can be fabricated raises concerns about privacy, digital manipulation, and the spread of misinformation[2][3].

These developments prompt deep philosophical questions. If human observers cannot tell the difference between real and AI-generated images, what does that say about our perception of reality? Furthermore, the inability to distinguish truth from fabrication in visual content challenges the foundations of knowledge and trust. As such, there is an increasing need for reliable detection methods to identify AI-generated imagery and preserve the integrity of digital content.

BACKGROUND

Convolutional Neural Networks (CNNs)[4][5] have played a pivotal role in the advancement of computer vision, especially in tasks like image classification and object recognition. Their multi-layered architecture is designed to automatically extract hierarchical

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

features from visual data, allowing them to perform well on standard datasets. However, their reliability tends to decrease in more challenging applications, such as identifying AI-generated imagery within complex and large-scale datasets.

Although CNNs are effective in many scenarios, they exhibit several key limitations that restrict their generalization and scalability:

a) **Inability to Prioritize Key Image Regions:** Standard CNNs treat all regions of an image equally, lacking the ability to selectively emphasize the most critical or informative areas. This uniform processing can result in missed subtle patterns, which are often essential when trying to differentiate between real and synthetically generated images.

b) **Overdependence on Deep Architectures:** To enhance accuracy, CNNs typically add more layers, making the network deeper. While this can improve feature learning, it also significantly increases the demand for computational resources and memory. Beyond a certain depth, the performance gains tend to plateau, making the model less efficient and more challenging to train.

c) **Insufficient Global Context Awareness:** Due to their reliance on localized receptive fields, CNNs are mainly effective at capturing short-range spatial information. This design limits their ability to understand the global context of an image, which is crucial in tasks that require interpreting spatial relationships across the entire scene.

d) **Poor Performance on Complex, Multi-Class Datasets:** CNNs often struggle when applied to large datasets with numerous categories. In such cases, similarities between different classes and variability within a single class can confuse the model, resulting in lower accuracy and weaker generalization.

These drawbacks highlight the growing need for more advanced and efficient models that can overcome these limitations—models that not only reduce computational load but also integrate attention mechanisms and global understanding to better handle complex, high-dimensional visual data.

DATASET

The dataset employed in this research is CIFAKE, which includes a total of 120,000 images evenly split between real and AI-generated (fake) samples. There are ten distinct categories, with each class containing 12,000 images—6,000 real and 6,000 synthetic. The real images, labeled as "REAL", are obtained from the CIFAR-10 dataset[6], which features 60,000 color images of size 32×32 pixels. These images are distributed across ten classes: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship,* and *truck*, with 6,000 samples per class. From this, 50,000 images are used for training and 10,000 for testing purposes.

To construct the "FAKE" class, synthetic counterparts were generated using a Stable Diffusion Model (SDM)[7], producing 6,000 AI-generated images per class, aligning with the structure of CIFAR-10. Prompt modifiers were employed during image generation to increase intra-class variation. The fake dataset[8] maintains the same train-test split as the real dataset, with 50,000 synthetic images used for training and 10,000 for testing, each labeled to indicate its artificial origin.

Figure 1: Examples of Real images

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

Figure 2: Examples of AI-generated

METHODOLOGY

As highlighted in the background, traditional Convolutional Neural Networks (CNNs) have key limitations, including the inability to emphasize important regions of an image, reliance on increased depth for better accuracy, and a lack of global context due to localized receptive fields. To address these issues, we propose a methodology that combines three advanced models: Attention-CNN with CBAM**,** EfficientNetB0**,** and Vision Transformer (ViT)**.** By integrating these models into an ensemble framework, we aim to overcome the weaknesses of traditional CNNs while enhancing performance. This approach enables precise feature extraction, optimized computational efficiency, and better capture of long-range dependencies, resulting in higher classification accuracy and robustness. The ensemble strategy merges the strengths of each model, ensuring a more reliable and effective solution.

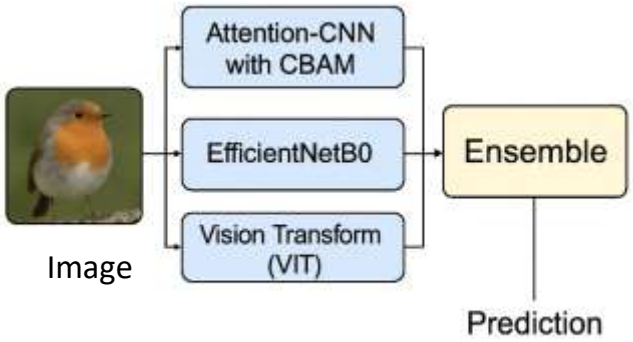

Figure 4: Complete Architecture

1.Attention-CNN with CBAM :

Convolutional Neural Networks (CNNs) are widely used for image classification due to their strong performance. However, they typically process all spatial regions and feature channels in a uniform way, which can limit their ability to detect subtle differences—especially in tasks like distinguishing real from AI-generated images.

To address this shortcoming, the CBAM-augmented CNN incorporates attention mechanisms[9] through the Convolutional Block Attention Module (CBAM)[10]. This module enhances feature representation by applying two types of attention:

a) **Channel Attention**: Identifies and emphasizes the most relevant feature maps (determining *what* to focus on).
b) **Spatial Attention**: Pinpoints and highlights important spatial locations in the image (deciding *where* to focus).

By integrating both attention types, the CBAM-equipped CNN can more precisely attend to significant visual cues. This improves its ability to capture subtle differences, leading to better classification performance—particularly on datasets like CIFAKE, where fine-grained details are crucial.

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

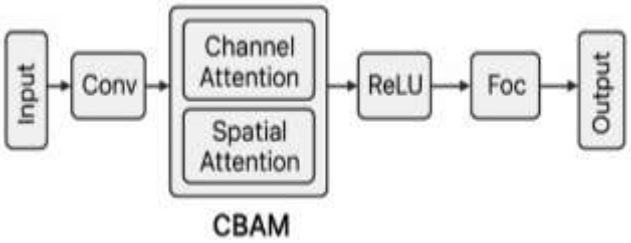

Figure 5: Attention-CNN with CBAM Architecture

Input Layer	(32, 32, 3)	—	(32, 32, 3)	0
Conv2D #1	(32, 32, 3)	Conv2D (128 filters, 3x3, same)	(32, 32, 128)	3,584
BatchNorm #1	(32, 32, 128)	Batch Normalization	(32, 32, 128)	512
ReLU Activation #1	(32, 32, 128)	ReLU	(32, 32, 128)	0
MaxPooling2D #1	(32, 32, 128)	2x2 Max Pooling	(16, 16, 128)	0
CBAM #1	(16, 16, 128)	CBAM: Channel + Spatial Attention	(16, 16, 128)	~34,049
Conv2D #2	(16, 16, 128)	Conv2D (128 filters, 3x3, same)	(16, 16, 128)	147,584
BatchNorm #2	(16, 16, 128)	Batch Normalization	(16, 16, 128)	512
ReLU Activation #2	(16, 16, 128)	ReLU	(16, 16, 128)	0
MaxPooling2D #2	(16, 16, 128)	2x2 Max Pooling	(8, 8, 128)	0
CBAM #2	(8, 8, 128)	CBAM: Channel + Spatial Attention	(8, 8, 128)	~34,049
Conv2D #3	(8, 8, 128)	Conv2D (128 filters, 3x3, same)	(8, 8, 128)	147,584
BatchNorm #3	(8, 8, 128)	Batch Normalization	(8, 8, 128)	512
ReLU Activation #3	(8, 8, 128)	ReLU	(8, 8, 128)	0
MaxPooling2D #3	(8, 8, 128)	2x2 Max Pooling	(4, 4, 128)	0
CBAM #3	(4, 4, 128)	CBAM: Channel + Spatial Attention	(4, 4, 128)	~34,049
Flatten	(4, 4, 128)	Flatten	(2048,)	0
Dense #1	(2048,)	Dense (128 units, ReLU)	(128,)	262,272
Dropout	(128,)	Dropout (rate=0.5)	(128,)	0
Output Dense	(128,)	Dense (1 unit, sigmoid)	(1,)	129

Table 1: Attention-CNN with CBAM model Summary

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

2. EfficientNetB0 :

Traditional CNNs often attempt to improve accuracy by increasing the number of layers, which can result in significant inefficiencies in terms of computation and memory. Without a well-balanced scaling approach, these models can become excessively large and less practical for applications that demand both high performance and efficiency—such as distinguishing subtle differences between authentic and AI-generated images. The EfficientNetB0-based CNN[11] overcomes these challenges by employing a compound scaling method that proportionally adjusts the network's depth, width, and input resolution. This coordinated scaling ensures a balanced model expansion, enhancing accuracy while keeping computational costs low.

Key architectural enhancements include:

a) **Compound Scaling:** EfficientNetB0 scales depth (number of layers), width (number of channels), and input resolution (image size) together, unlike traditional CNNs that typically increase only one of these components.

b) **MBConv Blocks with SE Modules:** The architecture incorporates mobile inverted bottleneck (MBConv) layers combined with Squeeze-and-Excitation (SE) modules, enabling more efficient channel-wise feature recalibration and improving representational strength with fewer parameters.

c) **Swish Activation and Global Average Pooling:** Instead of the traditional ReLU, EfficientNetB0 uses Swish activation, which enables smoother gradient flow. Global average pooling is also employed to condense features effectively, helping reduce the risk of overfitting.

By integrating these architectural strategies, EfficientNetB0 achieves strong performance on low-resolution datasets like CIFAKE while remaining computationally efficient. This makes it particularly suitable for deployment on edge devices and in real-time environments where lightweight models are essential.

Figure 6: EfficientNetB0 Architecture

Layer/Block	Input Shape	Operation	Output Shape	Trainable Params
Input Layer	(32, 32, 3)	Input	(32, 32, 3)	0
EfficientNetB0 Base	(32, 32, 3)	Convolutional Feature Extractor	(1, 1, 1280)	~2.5M* (partial)
GlobalAveragePooling2D	(1, 1, 1280)	Pooling	(1280,)	0
Dropout (0.5)	(1280,)	Dropout	(1280,)	0
Dense (ReLU, 128 units)	(1280,)	Fully Connected	(128,)	163,968
Dense (Sigmoid, 1 unit)	(128,)	Output Layer	(1,)	129

Table 2: EfficientNetB0 model Summary

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

3. Vision Transformer (ViT):

Traditional Convolutional Neural Networks (CNNs) excel at identifying local features in images, such as edges, textures, and corners. This is achieved through the use of small receptive fields, which allow the network to focus on localized areas of the image. However, one of the major limitations of CNNs is their difficulty in capturing global context. Understanding long-range relationships across distant regions of an image is essential for high-level vision tasks like scene understanding and object detection. CNNs, due to their inherently local nature, often struggle with this. To address this limitation, the Vision Transformer (ViT)[12] was introduced. Inspired by the success of Transformer models in natural language processing (NLP), ViT adapts the Transformer architecture to visual data. It treats an image as a sequence of patches, similar to how text is treated as a sequence of words, allowing the model to capture both local and global dependencies using self-attention mechanisms.

Key Components of Vision Transformer (ViT)

a) **Patch Embedding**: The input image is divided into fixed-size patches (e.g., 16×16 pixels). Each patch is flattened into a 1D vector and then passed through a linear projection (a fully connected layer) to generate patch embeddings.

b) **Positional Embedding**: Since Transformers do not have an inherent understanding of spatial relationships, positional embeddings are added to the patch embeddings. These embeddings encode the location of each patch within the original image, preserving spatial structure.

c) **Transformer Encoder**: The core of ViT consists of a stack of Transformer blocks. Each block includes multi-head self-attention and a feedforward neural network. These blocks process the entire sequence of patch embeddings simultaneously, enabling the model to learn complex interactions between patches.

d) **Multi-Head Self-Attention (MHSA)**: This mechanism allows each patch to attend to every other patch in the image. By doing so, the model can learn global patterns and long-range dependencies that are not easily captured by CNNs.

e) **Feedforward Neural Network (FFN)**: Following the attention mechanism, each patch embedding is further refined using a feedforward network. This enhances the model's capacity to learn and represent intricate features.

f) **Classification Head**: A special learnable token, known as the class token, is prepended to the sequence of patch embeddings before entering the Transformer. After encoding, this token aggregates information from all patches and is used for classification or other downstream tasks.

By leveraging the power of self-attention and treating images as sequences, the Vision Transformer effectively captures both local details and

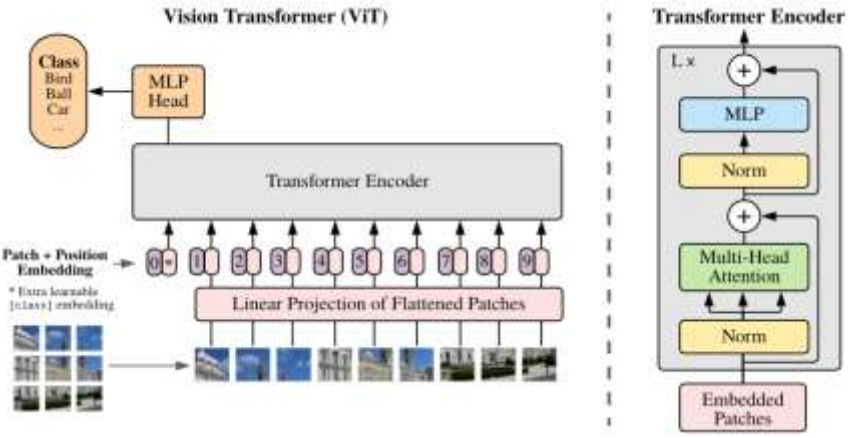

Figure 7: Vision Transformer (ViT) Architecture

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

Layer/Block	Input Shape	Operation	Output Shape	Trainable Params
Input Layer	(32, 32, 3)	-	(32, 32, 3)	0
Patch Embedding	(32, 32, 3)	Patch Extract (4x4), Dense(64)	(64, 64)	~12K
Positional Embedding	(64,)	Learnable Embedding	(64, 64)	~4K
Add	(64, 64)	Add Positional Embedding	(64, 64)	0
Transformer x3	(64, 64)	LayerNorm, MHA(4 heads), MLP(128, 64)	(64, 64)	~160K
LayerNorm	(64, 64)	Layer Normalization	(64, 64)	~128
Flatten	(64, 64)	Flatten	(4096,)	0
Dropout	(4096,)	Dropout (0.2)	(4096,)	0
Dense (ReLU)	(4096,)	Dense(128), ReLU	(128,)	~524K
Dropout	(128,)	Dropout (0.2)	(128,)	0
Output Dense (Sigmoid)	(128,)	Dense(1), Sigmoid	(1,)	129

Table 3: Vision Transformer (ViT) model Summary

4. Ensemble Strategy:

In this approach, an ensemble[13] of three distinct models is used to enhance the reliability and accuracy of predictions. Rather than relying on a single model, the outputs of all three models are combined to make the final decision. This ensemble method utilizes soft voting, a technique that takes into account the predicted probabilities from each model.

Soft Voting Mechanism:
Each individual model produces a probability score that represents its confidence in the input belonging to a particular class (for example, the "real" class or any class of interest). These scores are not hard classifications (i.e., not just 0 or 1), but continuous values between 0 and 1, representing the likelihood of the input belonging to a specific class.
The aggregation process follows these steps:

a) **Probability Averaging**: The predicted probability scores from all three models are averaged for each instance.
b) **Thresholding**:
The final prediction is made by comparing the averaged score against a threshold value—commonly set at **0.5**.
   i. If the average score ≥ 0.5, the instance is classified as **positive** (e.g., real class).
   ii. If the average score < 0.5, the instance is classified as **negative** (e.g., fake or other class).

RESULTS

The performance of the proposed ensemble model was rigorously evaluated on the CIFAKE dataset**,** which is designed for detecting deepfake images in a controlled and challenging setting. The model's evaluation metrics are presented in Figure 8 and reflect its effectiveness across multiple classification measures.

The ensemble model achieved an accuracy of 96.81%, demonstrating a high level of correctness in distinguishing real images from fake ones. The precision was 96.88%, indicating the model's strong ability to minimize false positives when predicting fake images. Similarly, the recall stood at 96.81%, reflecting the model's effectiveness in identifying nearly all true fake instances. The resulting F1-score of 96.77% confirms the model's balanced performance between precision and recall. Furthermore, the model reported a loss value of 0.3100, suggesting that it generalized well on the CIFAKE dataset without significant overfitting.

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

These findings affirm that the ensemble approach is particularly well-suited for deepfake detection tasks. Its high and consistent performance across all evaluation metrics makes it a promising candidate for real-world deployment in digital media integrity verification systems.

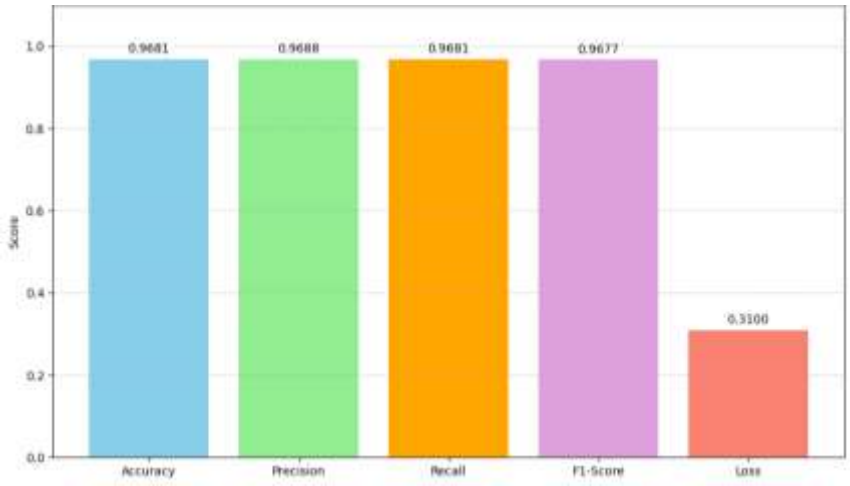

Figure 8: Ensemble Model Performance Metrics

CONCLUSION

In conclusion, the proposed hybrid deep learning model, which combines Attention-CNN with CBAM, EfficientNetB0, and Vision Transformer (ViT), significantly improves the detection of real versus AI-generated images. By overcoming the shortcomings of traditional CNNs, such as their limited ability to focus on important regions and the computational inefficiency that comes with increasing depth, this method offers a more optimized and precise solution. The use of CBAM enhances feature attention, EfficientNetB0 provides better resource efficiency through compound scaling, and ViT captures long-range dependencies by processing images as sequences of patches. Evaluations on the CIFAKE dataset showed an impressive accuracy of 96.81%, surpassing the performance of standard CNN models. This work highlights the advantages of merging attention-based mechanisms and transformer architectures to effectively address the growing challenges in distinguishing synthetic images, offering a more robust approach for AI-generated image classification.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘‘High-resolution image synthesis with latent diffusion models,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.

[2] G. Pennycook and D. G. Rand, ‘‘The psychology of fake news,’’ Trends Cogn. Sci., vol. 25, no. 5, pp. 388–402, May 2021.

[3] B. Singh and D. K. Sharma, ‘‘Predicting image credibility in fake news over social media using multi-modal approach,’

[4] C. Deng, G. Lu, H. Li, and Z. Li,"CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8756–8765.

[5] K. Simonyan and A. Zisserman,"Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations (ICLR), May 2015.

[6] Krizhevsky, A., & Hinton, G. E. (2009). *Learning multiple layers of features from tiny images*. Technical Report, University of Toronto.

iJETRM
International Journal of Engineering Technology Research & Management
(IJETRM)
https://ijetrm.com/

[7] R. Rombach, E. K. Lucic, S. Esser, B. Ommer, and L. Cremers,"High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.

[8] Deng, G. Lu, and H. Li,"CIFAKE: A large-scale synthetic dataset for fake image detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 1–8.

[9] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin,"Attention is all you need," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Dec. 2017, pp. 6000–6010.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon,"CBAM: Convolutional block attention module," in Proc. European Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 3–19.

[11] and Q. V. Le,"EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach

[12] Dosovitskiy, A., Beyer, L., & Uszkoreit, J. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[13] Z.-H. Zhou,"A survey of ensemble learning: Methods and applications," Neural Netw., vol. 18, no. 5–6, pp. 585–602, Aug. 2005.