Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ27635

# *A Survey of Application of Machine Learning Algorithms in Signal Recognition*

**Hongmin Chen**

*School of Science, Beijing Forestry University, Beijing, China*
*CHM6342273@163.com*

***Abstract.*** Signal recognition is vital for sectors like medical diagnosis, security monitoring, intelligent transportation, and voice interaction. Traditional methods, however, relying on manually designed features, struggle with complex patterns and high-dimensional signals. While machine learning—especially deep learning—addresses these issues via end-to-end learning and automatic feature representation, it suffers from over-reliance on high-quality labeled data. Real-world challenges such as signal noise and high labeling costs lead to label noise, impeding practical application. This review explores signal recognition fundamentals, including feature extraction, selection and classification. It details deep learning applications in image recognition, like CNNs, RNNs and Transformers. It also discusses multimodal learning, using AV-ASR and BPO-AVASR as examples. Finally, it identifies challenges (data scarcity, high model complexity, privacy issues) and proposes future directions (lightweight models, few-shot learning). The review concludes that deep learning dominates signal recognition, with models achieving human-level performance on benchmarks. Multimodal learning, fusing speech and image data, is a key trend.

***Keywords:*** Signal Recognition, Deep Learning, CNN, Transformer, Multimodal Learning

## 1. Introduction

Signal recognition, an information processing task, has evolved from empirical methods to data-driven approaches, reflecting progress in signal processing. Traditional methods, while effective for specific tasks, have limitations: handcrafted features fail to capture complex patterns and nonlinearities, require expert knowledge for feature extraction, and generalize poorly with high-dimensional signals [1]. In contrast, machine learning algorithms facilitate an end-to-end method of learning, automatically representing complex intrinsic signal features and discovering patterns by itself without manual design. The significant advance on recognition accuracy and generalization ability is thus achieved which serves as a major improvement with respect to the problems existing in traditional ways when they process the high-dimensional unstructured signals such as the time-dynamic characteristics of a voice signal and spatial texture characteristics of an image signal, which are hard to fully express using traditional methods [1].

This paper reviews the application progress of machine learning algorithms, especially deep learning , in voice signal and image signal recognition. This paper seeks to guide and inspire further

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ27635

advancements in the application of deep learning to signal recognition, ultimately contributing to the development of more efficient and accurate recognition systems.

## 2. Overview of signal recognition

Signal recognition is the process of acquiring information carried by signals and classifying or understanding that information. For the field of signal processing, signal recognition mainly includes three key steps: signal feature extraction, signal feature selection, and signal classification decision-making. It primarily falls into two categories: speech recognition and image recognition.

Speech recognition, fully known as Automatic Speech Recognition (ASR), is a technology that converts speech signals into text or executes commands. It encompasses functions such as speech-to-text conversion, voice command recognition, and speaker recognition. However, in real-world scenarios, the application of speech recognition is hindered by various issues, including noise interference, speakers' accents, and differences between individual speakers [2].

Image recognition refers to the process of detecting, classifying, and understanding targets within images. Its core tasks mainly include image classification, object detection, semantic segmentation, and image retrieval. Compared with speech signals, image signals have more dimensions and a more complex structure. Therefore, enabling machines to gain a deeper understanding of image signals remains a highly challenging task [3].

## 3. Applications of machine learning algorithms in signal recognition

### 3.1. Application of deep learning in image signal recognition

The Convolutional Neural Network (CNN) is pivotal in image recognition. Since AlexNet's success at the 2012 ImageNet competition, CNNs have excelled in tasks like image classification, object detection, and semantic segmentation. They use a hierarchical structure of convolutional layers, pooling layers, and fully connected layers to automatically extract image features. Shallow layers capture low-level features such as edges and textures, while deeper layers learn high-level semantic features [4].

Not only has the original AlexNet performed well on large-scale ImageNet image datasets, but other advanced CNN architectures have also been successful, such as VGGNet, GoogLeNet, ResNet, and DenseNet, etc., and various small enhanced network structures exist. In order to solve the problem of gradient vanishing that occurs during the model training process of deeper networks, ResNet(Residual Network) brought up the concept of "skip connection" or "shortcut connection", enabling the gradients to be passed backward to the earlier layers faster, thus making it possible for the deep networks to be trained [5]. Currently, the widely used CNN-based image recognition technology is basically applied in fields including security surveillance, medical image analysis, autonomous driving and industrial inspection. In medicine, through CNN, the medical pictures such as X-ray, CT and MRI could be distinguished, helping to diagnose diseases or detect disease in advance; in the industry, the defect and quality judgement of products and improvement of quality can be done by CNN; in agriculture, the plants disease, pests recognition, and estimation of yield can be realized with CNN [6].

Transformer—a new type of deep learning architecture—has gradually entered this field and yielded excellent outcomes. The ViT (Vision Transformer) introduced Transformers to image recognition and divided images into fixed-sized patches, viewing patches as sequence tokens, then fed them straight into the Transformer encoder for direct processing; this model was a good solution

DOI: 10.54254/2755-2721/2026.TJ27635

as the major idea that mapped the image into a sequence, and extracted long-range dependency information via the attention mechanism of the Transformer itself. Contrary to the local convolutions in CNNs, the attention itself builds up relations between local or non-local parts of the image directly, thus making it more applicable to extract such large-scale connections [7].

In real-world scenarios, a typical application of the Transformer is in the analysis and processing of satellite images. Due to the high resolution of satellite images and the complex structural features of scenes within them, using CNNs for object detection on such images leads to the problem of an insufficiently large receptive field. However, the Transformer, by utilizing the self-attention mechanism, can capture the global context of the image. Meanwhile, it delivers better performance than CNNs in tasks such as object detection and ground object classification [8]. Beyond satellite images, another typical scenario is the use of the Transformer in medical image analysis. It can help extract spatial connections between multi-layer slices and compensate for the limitation of CNNs, which only focus on local connections within 2D slices [9].

### 3.2. Applications of deep learning in speech signal recognition

Deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized speech signal recognition. Initially used for image recognition, CNNs were adapted to speech by converting signals into spectrograms or mel-spectrograms, extracting acoustic features directly from raw data. This eliminates the need for manual feature engineering, as seen in traditional methods like Mel-Frequency Cepstral Coefficients (MFCC). CNNs have been successfully integrated into models like Microsoft's Deep Speech 2, a hybrid CNN-LSTM model that achieves human-level performance on the LibriSpeech dataset [10-11]. CNNs are also effective in environmental sound classification [12].

Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, excel at processing sequential data and are widely used in speech recognition. Unlike CNNs, RNNs can capture temporal relationships between speech signals, making them well-suited for sequence prediction tasks in speech recognition scenarios [13]. Compared with traditional RNNs, LSTMs effectively address the issues of gradient vanishing and gradient explosion, and can capture correlations in longer sequences. Therefore, in RNN-based speech recognition frameworks, LSTMs are typically used to construct encoders: input signals are fed into the encoder to generate long vectors, which are then input to a decoder to infer text sequences [14]. In practical scenarios, to handle the alignment problem between speech and text sequences, either the Connectionist Temporal Classification (CTC) loss function or the attention mechanism is generally adopted. As mentioned earlier, classic speech recognition systems using LSTMs include Google's Speech Recognition System, which employs a multi-layer LSTM network structure combined with a language model and achieves excellent performance across multiple languages [15].

The Transformer, a novel deep learning architecture initially designed for NLP (Natural Language Processing) tasks, has made significant progress in the field of ASR (Automatic Speech Recognition) in recent years. Unlike traditional RNNs and CNNs, Transformers rely entirely on the self-attention mechanism for learning. This enables them to better capture long-range information connections and possess strong parallel computing capabilities. In speech recognition tasks, Transformers can directly take raw speech signals or their spectrograms as input and learn global relationships within the speech signals [16]. In addition to the standard Transformer, there is also the Conformer—a hybrid architecture combining convolutions and self-attention—which also delivers excellent performance in ASR [17]. Tsinghua University's Dolphin model is a successful example of applying Transformers to speech recognition. Built on a CTC-Attention structure, Dolphin features an E-Branchformer encoder, a Transformer decoder, and a 4x downsampling layer. The entire model

DOI: 10.54254/2755-2721/2026.TJ27635

enables efficient operation of large-scale models and supports recognition for 40 Oriental languages as well as 22 Chinese dialects (including Mandarin) [18].

OpenAI's Whisper is one of the Transformer models utilized in speech recognition, which serves as a model for all speech recognition tasks, based on an unsupervised strategy and using general supervised data. Thereby the model utilizes different public datasets throughout the world to supervise its training, aggregating a total amount of 680,000 hours [19]. This supervising procedure benefited the model with training in various languages and works especially well with detecting the language and giving speech translations at the same time of recording a voice snippet. Whisper is suited for various scenarios such as for delivering speech for GPT-4 input or doing speech classification tasks for CLIP etc.

### 3.3. Application of multimodal deep learning in signal recognition

Multimodal learning involves jointly processing information from multiple modalities, such as speech, images, and text, to enhance model performance by leveraging complementary information. This approach has become a significant trend in signal recognition, particularly in the joint recognition of speech and image signals [20]. A prime example of multimodal deep learning is Audiovisual Speech Recognition (AV-ASR), which integrates video data with traditional ASR models. Unlike audio-only ASR systems, which can be limited by environmental noise, mispronunciations, and homophones, AV-ASR uses visual context to improve recognition accuracy. For instance, seeing a "Coca-Cola" bottle can help correctly recognize the word "cola" [21].

To enhance the robustness of multimodal speech recognition systems, researchers proposed BPO-AVASR (Bifocal Preference Optimization for Audiovisual Speech Recognition) [22]. This method optimizes input and output preferences simultaneously. Input-end preference optimization (Focal I) involves occluding audio or perturbing video during training, exposing the model to various interferences. Output-end preference optimization (Focal II) introduces artificial errors into the output text, such as homophone swaps or ambiguous word rewrites.

Experiments on BPO-AVASR have achieved State-of-the-Art (SOTA) performance across various benchmark datasets, particularly excelling in complex scenarios with noise, strong interference, and challenging video conditions. For example, its Word Error Rate (WER) on the Ego4D dataset reached 12.5%, outperforming MIAV-TST3—the previous SOTA AV-ASR model on Ego4D. Furthermore, BPO-AVASR trained on only 300 hours of data achieved performance comparable to that of AVFormer (the SOTA model on the He2 dataset), which required 131K hours of data. Additionally, experiments confirmed that BPO-AVASR has a stronger ability to utilize visual cues. In one example, when the video showed "cherry tomatoes", the model without BPO would incorrectly recognize it as "chair to mirror"; however, with BPO enabled, the recognition result was the correct "cherry tomatoes". This clearly demonstrates the superiority of multimodal deep learning models [22].

## 4. Challenges and future directions

Despite the outstanding achievements of machine learning models in signal detection tasks, several challenges remain. These problems can be put into four types: data issues, model issues, signals issue, and application issue.

Data scarcity is a significant problem, especially in fields like healthcare and finance, where high-quality labeled data is expensive and rare. Additionally, heavy bias exists in the data if it contains many classes, rare classes especially, because the model may not learn any distinction

DOI: 10.54254/2755-2721/2026.TJ27635

properly which will make it challenging to find rare-class samples. Models themselves face issues such as high computational complexity due to a large number of parameters, which requires substantial computing resources and is incompatible with resource-constrained devices like mobile terminals and embedded systems. Furthermore, most deep learning models are black boxes that do not have enough explainability and understandability; thus they cannot satisfy the interpretability requirements of some fields where understanding the model outputting reasons is important, such as healthcare and finance.

Beyond errors, another challenge present in signal processing is noise and interference. In real-life situations, signals are prone to be contaminated by the noise and other types of interferences, which will pollute the signals and therefore jeopardize the correctness of recognition results. Noise interference is another significant problem that arises besides errors when dealing with signals in practice. In fact, during actual usage signals will always be subject to noise and disturbances in many different ways so that the signals deteriorate and recognition quality gets ruined.

Finally, in terms of applications, the need for real-time performance is also a key consideration. In application scenarios such as real-time speech recognition and real-time video analysis, results are required within a short period of time, which demands high computational efficiency from the model and thus imposes high requirements on the model itself. Secondly, there is the issue of user privacy. Users are more concerned about privacy protection, and when processing sensitive data including voice recordings, images, etc., the question of how to better protect user privacy arises.

To address the aforementioned issues, future research should focus on developing efficient models through lightweight architectures and model compression techniques like pruning, quantization, and knowledge distillation. Efficient data learning can be enhanced by few-shot learning, meta-learning, and advanced data augmentation methods to improve model generalization. Multimodal fusion, including multimodal generative models like text-to-image, image-to-text, and speech-image conversion models, should also be explored for broader applications.

## 5. Conclusion

This paper reviews the application of machine learning in signal recognition, with a focus on expounding the applications of deep learning technologies in speech signal and image signal recognition. Based on the current research status, the following conclusions can be drawn:

Machine learning algorithms currently serve as the primary methods for signal recognition applications. In terms of image signal recognition, deep learning applications have reached a high level of maturity. For speech signal recognition, deep learning plays a core role. Models based on CNNs, RNNs and their variants, as well as Transformers, have all achieved or even surpassed human-level performance on multiple public speech recognition benchmark datasets. Additionally, multimodal deep learning stands out as one of the development trends in signal recognition. It can fuse information from two or more modalities simultaneously—for example, combining image and speech data. By fully leveraging the complementary information between different modalities, it effectively enhances recognition performance.

Machine learning algorithm development has made great achievements in recognizing signals but many challenges remain regarding how to enhance robustness, interpretability and generalization of models, learn efficiently even when resources are scarce, and fuse data with various forms. And yet these are problems which must be overcome as we move forward.

In conclusion, important advancements have occurred in signal recognition technologies based on machine learning methods, especially deep learning, meanwhile people think that more breakthroughs can be expected later. Signal recognition can greatly facilitate people's daily life,

DOI: 10.54254/2755-2721/2026.TJ27635

accelerate the popularization of AI and also can better serve the intelligent development of every aspect of our lives.

## References


[1] Sun, S., Peng, S., Yao, Y., & Yang, X. (2024). A survey on modulation recognition based on deep learning. Telecommunications Science, 37(5), 82–90.

[2] Xue, Y., He, H., & Yang, Y. (2023). Neural-network-based speech signal recognition and classification. Modern Electronics Technique, 46(24), 79-84. https: //doi.org/10.16652/j.issn.1004-373x.2023.24.014

[3] Li, Y., Shen, J., & Xu, Q. (2023). A review of learning algorithms for multi-layer spiking neural networks oriented toward image recognition. Journal of Image and Graphics, 28(2), 385–400.

[4] Purwono, P., Ma'arif, A., Rahmaniar, W., Fathurrahman, H. I. K., Frisky, A. Z. K., & ul Haq, Q. M. (2022). Understanding of convolutional neural network (cnn): A review. International Journal of Robotics and Control Systems, 2(4), 739-748.

[5] Ma, H. (2025). A Review of Image Segmentation Based on Deep Learning. Computer Knowledge and Technology, 21(19), 42-44. https: //doi.org/10.14004/j.cnki.ckt.2025.0938

[6] Gollangi, H. K., Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., & Reddy, M. S. (2020). Unveiling the hidden patterns: AI-driven innovations in image processing and acoustic signal detection. JOURNAL OF RECENT TRENDS IN COMPUTER SCIENCE AND ENGINEERING (JRTCSE), 8(1), 25-45.

[7] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z. H., ... & Yan, S. (2021). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 558-567).

[8] Jin, C., & Tong, C. (2023). Remote-sensing image classification method combining CNN and Transformer structures. Laser & Optoelectronics Progress, 60(20), 2028006-2028006.

[9] Tang, J. (2024). Research on Medical Image Registration Technology Based on Deep Learning (Master's Thesis, Jilin University). Master's thesis. https: //doi.org/10.27162/d.cnki.gjlin.2024.002589

[10] Liu, Y., Chen, A., Zhou, G., Yi, J., Xiang, J., & Wang, Y. (2024). Combined CNN LSTM with attention for speech emotion recognition based on feature-level fusion. Multimedia Tools and Applications, 83(21), 59839-59859.

[11] Zhu, H., Guo, J., & Ding, Z. (2023). Communication Signal Modulation Recognition Based on Inception-V3 Transfer Learning. Proceedings of the 2023 6th International Conference on Signal Processing and Machine Learning.

[12] Xu, S. (2024). Research on environmental sound classification methods based on neural networks. Electroacoustic Technology, 48(10), 54-56. https: //doi.org/10.16311/j.audioe.2024.10.015

[13] Ho, N. H., Yang, H. J., Kim, S. H., & Lee, G. (2020).Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention based Recurrent Neural Network. IEEE Access, 8, 1-1.

[14] Li, H. (2021). A Review of Natural Language Processing Research Based on RNN and Transformer Models. Information Recording Materials, 22(12), 7-10. https: //doi.org/10.16009/j.cnki.cn13-1295/tq.2021.12.081

[15] Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. Artificial intelligence review, 53(8), 5929-5955.

[16] Li, X. (2024). Research on Speech Recognition Technology Based on Deep Learning (Master's thesis, North China University of Technology). Master's dissertation. https: //doi.org/10.26926/d.cnki.gbfgu.2024.000086

[17] Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., ... & Zhang, Y. (2021, June). Recent developments on espnet toolkit boosted by conformer. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5874-5878). IEEE.

[18] Meng, Y., Li, J., Lin, G., Pu, Y., Wang, G., Du, H., ... & Zhang, W. Q. (2025). Dolphin: A Large-Scale Automatic Speech Recognition Model for Eastern Languages. arXiv preprint arXiv: 2503.20212.

[19] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR.

[20] Abeer A. Wafa, Mai M. Eldefrawi & Marwa S. Farhan.(2025). Advancing multimodal emotion recognition in big data through prompt engineering and deep adaptive learning.Journal of Big Data, 12(1), 210-210.

[21] Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., ... & Synnaeve, G. (2020). Rethinking evaluation in ASR: Are our models robust enough?. arXiv preprint arXiv: 2010.11745.

[22] Wu, Y., Lu, Y., Peng, Y., Wang, X., Song, R., & Watanabe, S. (2025, April). Enhancing Audiovisual Speech Recognition Through Bifocal Preference Optimization. In Proceedings of the AAAI Conference on Artificial

DOI: 10.54254/2755-2721/2026.TJ27635

Intelligence (Vol. 39, No. 24, pp. 25516-25524).