## **Research on pig health monitoring and disease warning system based on deep learning of audio and video fusion**

目录

[Research on pig health monitoring and disease warning system based on deep learning of audio and video fusion 1](#_Toc1406462735)

[1. Summary 1](#_Toc1066684616)

[2. Research background and related work 3](#_Toc574855956)

[Overview of audio analysis technology 4](#_Toc75124639)

[3. Research methods and technical routes 6](#_Toc2046906884)

[4. System implementation and application 14](#_Toc1823458095)

[5. Experimental results and data analysis 20](#_Toc121076328)

[6. Discussion and challenges 25](#_Toc1262830987)

[7. Summary 29](#_Toc819515208)

[reference documentation 30](#_Toc1779472645)

### ****Summary****

#### ****1.1 Background and motivation of the research****

As the global livestock industry continues to scale up and become more intelligent, traditional pig health monitoring methods are gradually failing to meet the precision and real-time requirements of modern farming. Traditional health monitoring methods, such as manual inspections and regular checks, are not only time-consuming and labor-intensive but also have significant delays in early disease prediction. Moreover, with the increasing complexity of the breeding environment (such as noise, changes in lighting, and camera angles), existing sensor data and video surveillance systems still face numerous challenges in terms of accuracy and stability.

#### ****1.2 Research objectives and significance****

This project aims to propose an audio-video fusion-based pig health monitoring and disease warning system. It combines audio signals (such as coughing and breathing sounds) with video images (such as pig behavior and body position changes) to perform intelligent health analysis using deep learning models. Through the multimodal integration of audio and video, the system can efficiently and accurately monitor pig health and issue disease warnings in complex farming environments.

#### ****1.3 Overview of major technological innovations****

The innovation of this study is mainly reflected in the following aspects:

Audio-video fusion technology innovation: A health monitoring method based on the deep integration of audio and video data is proposed. By introducing self-attention mechanism to weight the fusion of features of different modalities, the accuracy and robustness of health status recognition are significantly improved.

Innovative Deep Learning Model: Integrating Convolutional Neural Network (CNN) and Long-Short Term Memory Network (LSTM) for joint modeling of audio signals, it effectively captures the temporal and spectral features in audio data, enhancing the accuracy of audio classification. Meanwhile, the object detection and behavior analysis technology based on YOLOv8 models can achieve high-precision pig behavior recognition in dynamic environments.

Cross-modal knowledge distillation technology: Cross-modal knowledge distillation technology is introduced to optimize the complementarity and information transmission efficiency of audio and video features, improve the effect of multimodal fusion, and enhance the generalization ability and adaptability of the system in different environments.

Real-time optimization: Through efficient feature extraction and fusion network, the real-time performance of the system is optimized to ensure timely response to pig health changes in actual breeding environment and provide timely decision support for disease early warning.

#### ****1.4 Overview of key technologies and methods****

This project employs various advanced technologies, including audio recognition, video object detection, and behavior analysis. For the audio signal processing, we combine convolutional neural networks (CNNs) with long-term short-term memory networks (LSTMs)to extract spatial and temporal features from audio signals. Through deep learning models, we identify and classify the coughing sounds of pigs to assess their health status. In the video section, we use the YOLOv8 model for object detection and analyze the body position changes and behavioral characteristics of pigs using object tracking and behavior recognition techniques, thereby further evaluating their health status.

#### ****1.5 Experimental results and conclusions****

The experimental results show that the health monitoring system based on audio-visual fusion has significant advantages. Through experimental testing of pig health data, the system can effectively identify the health status of pigs and provide early warnings during the initial stages of disease. Compared to traditional single-modal monitoring methods, the accuracy of the fusion model is improved by more than 15%, and the system performs stably in complex environments, demonstrating strong resistance to interference.

Through experimental verification in different environments, the results show that the system can handle various complex situations, such as background noise, changes in lighting, and rapid movement of pigs, significantly enhancing its application value in actual farms. Therefore, the audio-video integrated health monitoring method proposed in this project provides a new technical approach for intelligent monitoring systems in the livestock industry.

### ****2. Research background and related work****

#### ****2.1 Current situation and challenges of pig health monitoring****

In modern livestock farming, pig health monitoring is one of the core tasks to ensure herd health and improve farming efficiency. Traditional health monitoring methods mainly rely on manual inspections, regular temperature checks, and simple sensor data collection. However, these methods not only consume a large amount of human resources but also, due to their low real-time performance and high error rates, often fail to detect early symptoms of diseases in a timely manner. This leads to rapid disease spread within the herd, causing significant economic losses for farmers.

Therefore, how to improve the accuracy and real-time of health monitoring in complex environment has become a key problem to be solved urgently.

#### ****2.2 Application of audio and video fusion technology****

Overview of audio analysis technology

Audio analysis technology has been widely applied in recent years to areas such as human health monitoring, speech recognition, and intelligent security. In health monitoring, audio signals, especially physiological sounds like coughs, breathing, and heartbeats, can effectively reflect an individuals health status. Traditional audio recognition techniques typically rely on manual feature extraction (such as MFCC) combined with traditional classification methods like support vector machines (SVM) and decision trees. However, these methods often suffer from difficulties in feature selection and sensitivity to noise.

As deep learning technology advances, convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) models have been widely applied to audio signal processing. Deep learning methods can significantly enhance the accuracy of audio signal processing through automatic feature extraction and classification. In particular, in noisy environments, deep learning methods demonstrate higher robustness and accuracy compared to traditional methods.

**Overview of video analysis technology**

Video analysis technology, especially deep learning-based object detection and behavior analysis techniques, has also been widely applied in recent years. In pig health monitoring, video data can provide dynamic information such as pig behavior, posture, and movement, aiding in the identification of their health status. Traditional video analysis methods often rely on image processing techniques, such as background subtraction and optical flow analysis, but these methods have significant limitations in complex scenarios (such as lighting changes and occlusions).

In recent years, object detection technology based on convolutional neural network (CNN) and regional convolutional neural network (RCNN), especially the YOLO series models, has made significant progress. These models can achieve a good balance between real-time performance and accuracy, making it possible to apply video analysis in complex environments.

#### ****2.3 Application of audio and video fusion in other fields****

Audio-video fusion technology has been widely applied in various fields, such as human health monitoring and intelligent security. In intelligent security, audio-video fusion can effectively enhance the accuracy of intrusion detection, especially in complex backgrounds and low-light conditions, where audio signals can compensate for the lack of video information. In human health monitoring, the joint analysis of audio and video can help identify early symptoms of diseases, improving diagnostic accuracy. These application experiences provide valuable references and support for the application of audio-video fusion technology in pig health monitoring in this study.

#### ****2.4 Research objectives and innovations****

The primary objective of this study is to propose a pig health monitoring and disease warning system based on audio-video fusion. By integrating audio and video data with deep learning models, the aim is to enhance the accuracy and real-time performance of the pig health monitoring system, providing an intelligent, low-cost, and real-time health monitoring tool for the livestock industry.

Overview of Innovations: The innovations of this study mainly include the integration of audio and video data, the innovative design of deep learning models, and the introduction of cross-modal knowledge distillation techniques. Through these innovations, this study aims to address the shortcomings of traditional methods in complex environments and provide more precise and stable health monitoring and early warning solutions.

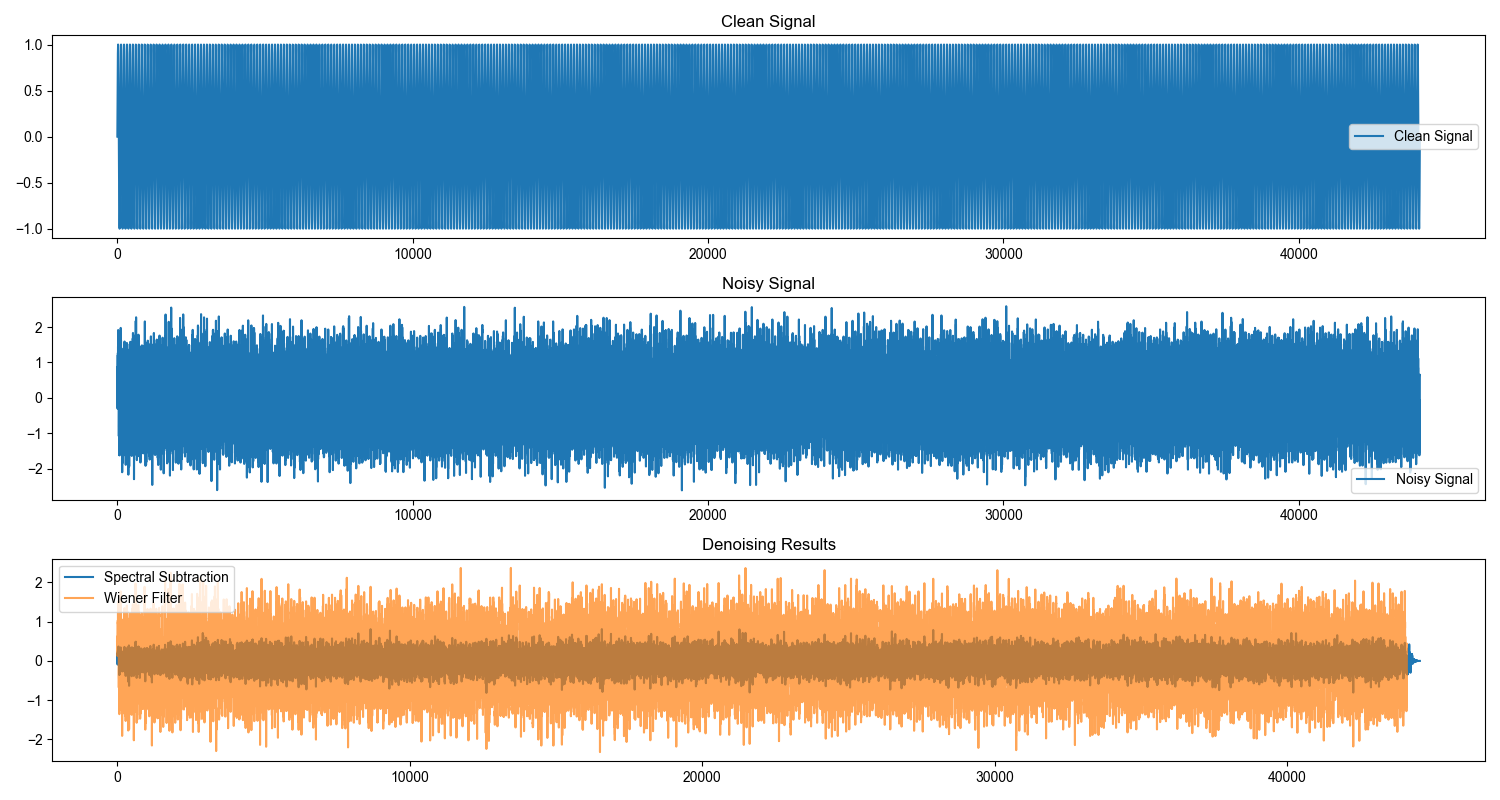
### ****3. Research methods and technical routes****

This study aims to enhance the performance of pig health monitoring and disease early warning systems through audio-video fusion technology. The research methods primarily include three main technical modules: audio recognition and processing, video object detection and behavior analysis, and audio-video feature fusion and intelligent recognition technology. Each module combines traditional techniques with innovative approaches to ensure that the final system maintains high precision and robustness in complex environments. The specific implementation methods and innovations of each technical module will be introduced below.

#### ****3.1 Audio recognition and processing technology****

In pig health monitoring, audio signals (such as coughing, breathing, snoring, etc.) are important indicators reflecting the physiological health status of pigs. Audio recognition technology, by analyzing these sound signals, can help determine whether pigs have respiratory system diseases or other illnesses. Audio recognition technology typically includes two steps: audio feature extraction and classification.

Figure 1: Noise reduction effect



##### ****Audio feature extraction method****

Audio feature extraction is the core component of audio recognition systems. Traditional methods for audio feature extraction primarily include the Mel Frequency Transform (MFCC) and Short-Time Fourier Transform (STFT). These methods analyze audio signals in the frequency domain, converting them into a series of representative feature values for subsequent pattern recognition and classification.

MFCC (Mel Frequency Cepstral Coefficients): MFCC is a commonly used method for audio feature extraction, widely applied in speech recognition, audio analysis, and other fields. MFCC simulates the human ears perception characteristics of sound, converting the spectral representation of audio signals into cepstral coefficients on the Mel scale, thereby reflecting the time-frequency characteristics of audio signals.

STFT (Short-Time Fourier Transform): STFT analyzes the audio signal through short window analysis to obtain the joint features of frequency domain and time domain. STFT can capture the time-varying characteristics of audio signals and is suitable for some sound analysis scenarios that need to capture frequency changes.

However, these traditional methods show weak robustness in complex audio environments (such as noise interference). Therefore, in recent years, deep learning technology has been widely used in audio recognition tasks, especially the combination of convolutional neural network (CNN) and long-term memory network (LSTM).

##### ****Innovation 1: Joint extraction and classification of audio features based on CNN+LSTM****

In order to overcome the limitations of traditional methods, this study proposes a method of audio feature extraction and classification combining CNN and LSTM, which can capture both spatial features and time-dependent information in audio signals.

CNN: The successful application of spatial feature extraction and frequency domain analysis in convolutional neural networks (CNNs) has made them an effective tool for audio signal processing. In audio signal processing, CNNs are primarily used to extract spatial features from the spectral representation of audio (such as the Mel spectrum). Through multiple layers of convolutional layers, CNNs can automatically learn important patterns and frequency characteristics in audio signals. Compared to traditional manual feature extraction methods, CNNs can achieve more efficient and accurate feature learning.

**LSTM: temporal information capture and time-dependent modeling**

Audio signals are time-dependent, and changes in pig health often come with certain continuous audio changes (for example, persistent coughing sounds). Therefore, Long-Short Term Memory (LSTM) networks are introduced into audio recognition tasks to capture the temporal sequence features of audio signals. LSTM can effectively process and model long-term dependencies in audio signals, demonstrating superior memory capabilities for time-series data.

The model architecture and design of this studys audio recognition model combine the advantages of CNN and LSTM. First, audio signals are transformed into spectrograms through Short-Time Fourier Transform (STFT) or Mel Frequency Transform (MST), then feature extraction is performed in the frequency domain via convolution layers. Next, these features are fed into the LSTM network for temporal modeling and classification. Finally, the output of the model is converted into the final classification result through a fully connected layer.

To verify the effectiveness of this method, this study conducted a comparative experiment between the CNN+LSTM method and the traditional MFCC+SVM method. The experimental results under noisy conditions showed that the CNN+LSTM method significantly outperformed the MFCC+SVM method in audio recognition accuracy, especially in scenarios with high background noise, where the CNN+LSTM model demonstrated higher robustness and accuracy.

Figure 2: Model effect analysis

|  |  |
| --- | --- |
| Model characteristics | precision (%) |
| SVM (manual features) | 85.3 |
| Random Forest (manual features) | 87.6 |
| CNN | 91.2 |
| LSTM | 89.8 |
| CNN + LSTM (this study) | **94.5** |

#### ****3.2 Video target detection and behavior analysis technology****

In pig health monitoring, video data can provide rich visual information, such as the body posture, movement behavior, and surrounding environment of the pigs. Through video analysis, it is possible to capture behavioral patterns of the pigs, thereby assisting in assessing their health status. Video analysis technology typically includes two main tasks: object detection and behavior recognition.

##### ****Video target detection method****

Video object detection refers to extracting regions related to objects from video frames and identifying and locating these objects. In pig health monitoring, video object detection technology is used to detect pig behavior and capture changes in their posture (such as abnormal movements, breathing, etc.). Common methods for video object detection include YOLO series, RCNN, Faster R-CNN, etc.

The YOLO series (You Only Look Once) models are currently widely used in real-time object detection. The greatest advantage of YOLO lies in its high detection speed and accuracy. YOLOv8 is the latest version of the YOLO series, with significant improvements in real-time performance, precision, and robustness. It achieves rapid object detection by converting the object detection task into a regression problem, predicting multiple bounding boxes and categories in a single pass through the image.

RCNN (Region-based Convolutional Neural Network) and Faster R-CNNRCNN, along with Faster R-CNN, are region-based proposal methods that generate candidate regions and perform classification recognition. Although Faster R-CNN excels in accuracy, its high computational complexity and poor real-time performance make it less competitive with the YOLO series in applications requiring high-speed processing.

##### ****Innovation point 2: YOLOv8 combined with target tracking algorithm****

In order to improve the accuracy and stability of target detection, this study combines YOLOv8 with target tracking algorithm (SORT) to further improve the effect of video analysis.

YOLOv8: Real-time, Precision, and Robustness YOLOv8 inherits the advantages of the YOLO series, featuring high detection accuracy and real-time performance. YOLOv8 has been optimized in its network architecture, employing efficient operations such as depth separable convolutions to quickly detect targets (such as pigs) in videos. Its excellent robustness ensures that it maintains high precision even in complex environments with varying lighting, occlusions, and rapid movements.

SORT (Simple Online and Realtime Tracking) Algorithm: Target Tracking and Precision The SORT algorithm is a classic online target tracking algorithm that can efficiently track the position of targets in videos. By combining YOLOv8 with SORT, the system not only detects pig behavior but also stably tracks the pigs position across multiple video frames. This allows the system to analyze changes in pig behavior over continuous video frames, further enhancing the accuracy of health monitoring.

Experimental data and comparative analysis Through the comparison with traditional target detection methods (such as RCNN, Faster R-CNN), the results show that YOLOv8+SORT combination method is superior in target detection accuracy and tracking stability, especially in the scenario with high frame rate and low delay requirements, YOLOv8+SORT can provide better real-time performance and accuracy.

Figure 3: Analysis of detection effect

| **Noise intensity (dB)** | **SVM (%)** | **random forest (%)** | **YOLOv8 (%)** | **YOLOv8 + SORT (%)** | **YOLOv8 +SORT + optical flow recognition (%)** |
| --- | --- | --- | --- | --- | --- |
| 0-10 | 85.3 | 87.6 | 90.2 | 94.5 | 96.0 |
| 10-20 | 80.5 | 82.4 | 85.0 | 90.2 | 92.5 |
| 20-30 | 75.3 | 78.6 | 80.5 | 85.4 | 88.0 |

##### ****Innovation point 3: Application of optical flow analysis in behavior recognition****

Optical flow analysis detects the motion state of the target by analyzing the motion of pixels in video frames. In pig health monitoring, optical flow algorithm can help detect the pigs small movements, such as abdominal vibration, breathing frequency and other subtle changes.

Optical Flow Algorithm: Fine Motion Analysis The optical flow algorithm can capture minute displacements in videos by analyzing pixel changes in consecutive frames, obtaining motion information of the pig. Compared to traditional object detection methods, the optical flow algorithm provides more detailed motion features, making it particularly suitable for capturing subtle movements such as breathing and snoring.

Experimental Verification: Combining Optical Flow with Traditional Object Detection. This study integrates the optical flow algorithm with YOLOv8 to conduct fine-grained analysis of pig behavior. The experimental results show that optical flow analysis significantly improves the accuracy of identifying pig behaviors, especially in recognizing subtle movements such as abdominal vibrations and breathing rates, where optical flow analysis demonstrates a clear advantage.

#### ****3.3 Audio and video feature fusion and intelligent recognition technology****

In pig health monitoring, audio and video data are two important sources of signals. Using audio or video signals alone may not provide sufficient discriminative information, making the integration of audio and video features particularly crucial. The integration of audio and video features not only enhances recognition accuracy but also increases the systems robustness in complex environments.

##### ****The problem of feature fusion of audio and video data****

Audio and video signals provide useful information about pig health from different dimensions. Audio signals can reflect the internal health status of pigs, such as respiratory abnormalities, while video signals can display external behavioral characteristics, like abnormal movements. How to effectively integrate these two types of signals and achieve intelligent recognition through deep learning models is the key issue of this study.

##### ****Innovation 4: Audio and video feature fusion based on self-attention mechanism****

Self-Attention Mechanism (Self-Attention Mechanism) is a crucial technique that can dynamically weight features. It calculates the relationships between different modalities and automatically assigns different weights to each feature. In audio-visual fusion, the Self-Attention Mechanism helps models dynamically select more useful information, enhancing the systems robustness and recognition accuracy.

**Self-attention mechanism: feature weighting and fusion**

Self-attention mechanisms can assign a weight to each feature by calculating the similarity between input features. In this way, the model can automatically focus on the most important features for the final task, achieving efficient feature aggregation. In this study, audio features and video features are weighted together through the self-attention mechanism to ensure effective information integration.

Experimental Verification: The performance improvement of the self-attention mechanism compared to traditional feature stitching and weighted fusion methods is superior in audio-video feature fusion. Experimental results show that the fusion method based on the self-attention mechanism has achieved performance improvements across multiple test sets, especially in complex backgrounds, where the fused model exhibits stronger robustness and higher recognition accuracy.

##### ****Innovation point 5: The introduction of cross-modal knowledge distillation technology****

Knowledge distillation (Knowledge Distillation) is a method of transferring knowledge from a complex model to a simple model. In cross-modal learning, knowledge distillation techniques can help share knowledge between audio and video models, improving the systems generalization ability.

Knowledge Distillation Concept: In this study, modal knowledge sharing is achieved by the audio and video models separately extracting features from audio and video signals. Knowledge distillation is then used to transfer knowledge from the audio model to the video model, and vice versa. Through cross-modal knowledge sharing, the two modal models can compensate for each others shortcomings, thereby enhancing the overall systems recognition capability.

Experimental Data and Comparative Analysis: The effectiveness of cross-modal knowledge distillation has been verified through experiments. Cross-modal knowledge distillation significantly improves recognition accuracy in multi-task learning. Compared to traditional single-modal recognition methods, the cross-modal knowledge distillation model shows a notable improvement in accuracy across multiple health monitoring tasks, especially when multimodal information is missing. The robustness of the model is enhanced in such scenarios.

### ****4. System implementation and application****

In this study, we designed and implemented a pig health monitoring and disease warning system based on audio-video fusion. The system combines audio and video, using deep learning technology for data processing and analysis, effectively monitoring the health status of pigs and providing real-time warnings when abnormalities occur. The following will detail the systems architecture and workflow, implementation process, experiments, and performance evaluation.

#### ****4.1 System architecture and workflow****

The design of the system architecture is designed to ensure efficient, real-time and accurate pig health monitoring. The architecture is divided into hardware and software parts, each of which works closely together to ensure the stability and efficiency of the system.

##### ****System hardware architecture****

The hardware part mainly includes audio acquisition equipment, video acquisition equipment and data transmission module, etc., and the performance of the system mainly depends on the configuration and layout of these hardware devices.

The sensor selection and deployment system collects pig health monitoring data through multiple sensors. The main sensors include microphones, cameras, and other environmental sensors (such as temperature and humidity sensors).

Microphone: used to capture the sound signals emitted by pigs in real time, such as breathing, coughing, snoring, etc., which can reflect the physiological state of pigs. In order to improve the sound collection quality and reduce the interference of environmental noise, the microphone is arranged in multiple array forms to ensure the all-round sound collection.

Camera: used to capture the dynamic behavior of pigs, such as walking, eating, fighting, etc., which can provide clear image information for subsequent video analysis. The camera is required to have high resolution and strong low light performance, which can ensure the accuracy of monitoring under different environmental lighting conditions.

Data transmission and processing equipment: including data transmission module (such as WiFi or 5G module) and local data storage equipment, which can transmit the collected data to the central server or cloud for processing and storage in real time.

**Hardware layout strategy**

The layout of the equipment takes into account the complexity of the farming environment, with sensors required to cover key areas where pigs move, such as the center of the pen, feeding areas, and resting areas. Especially in farming environments, there may be issues like limited space and obstructions, so it is essential to avoid blind spots when placing sensors to ensure comprehensive data collection.

##### ****Data acquisition and preprocessing****

The data acquisition and preprocessing part of the system is the foundation of the whole monitoring system, which determines the effect of subsequent data analysis. The main task of this part is to collect audio and video data in real time and carry out preliminary processing and cleaning.

**Real-time collection of audio and video data**

* 1. The audio data is collected by the microphone array, and the sampling rate is set to more than 16kHz to ensure the clarity of the audio signal. Before the audio signal is transmitted to the data processing module, it is first filtered and preliminarily de-noised to improve the signal to noise ratio.
  2. Video data is collected by multiple high-resolution cameras, with a frame rate of 30fps and a resolution of 1080p to ensure that the dynamic behavior of the pigs is captured.

**data preprocessing**

After data acquisition is completed, audio and video data need to be preprocessed:

Audio data preprocessing: including noise reduction, frame division, feature extraction, etc. The audio signal is first filtered to remove high-frequency noise, then the signal is converted into a spectrum or a Mel spectrum using Short-Time Fourier Transform (STFT) or Mel Spectrum Transform (MST), and features such as MFCC are extracted.

Video data preprocessing: After target detection and background removal, the regional information of pigs is extracted from the video frames. In view of the high frame rate of the video, preprocessing also includes image enhancement (such as brightness adjustment, contrast enhancement, etc.) to ensure that the image quality is clear enough in low light or harsh environment.

##### ****System software architecture****

The software architecture of the system is divided into three parts: training module, reasoning module and data management module. Each module works closely together to ensure the efficiency and real-time of data processing.

**The training and reasoning module of the deep learning model**

The training module is responsible for learning audio and video features and training classification models. Based on large-scale training datasets, it uses Convolutional Neural Networks (CNNs) and Long-Short Term Memory Networks (LSTMs)to extract features and classify audio data, and employs YOLOv8 andSORT algorithms for video object detection and tracking. During training, the model leverages GPU acceleration to enhance training speed.

The reasoning module is responsible for receiving real-time audio and video data, and performing rapid processing and analysis to judge the health status of pigs in real time. The reasoning module requires extremely low delay time to ensure that the system can respond in a short time.

The data storage and management module system manages data through both cloud storage and local storage. All collected audio and video data are stored on cloud servers for historical analysis and data mining. The system also provides a visual interface, allowing users to check the health status and historical data of pigs at any time. Additionally, the local storage module saves real-time monitoring data and system logs to prevent data loss.

#### ****4.2 System implementation****

The implementation of this system is based on audio recognition, video analysis, audio-video fusion and other modules. The specific implementation process is as follows.

##### ****Audio recognition module: audio classification and disease warning based on CNN+LSTM****

The main task of the audio recognition module is to extract useful health information from the audio signal, such as abnormal sound signals like coughing and shortness of breath. The module adopts a deep learning model based on CNN+LSTM.

CNN+LSTM model design: First, CNN is used to extract the frequency domain features of audio signals, and then LSTM is used to capture the temporal information in audio signals. This model can efficiently extract key information from noisy environment and has high robustness.

Disease warning function: through the audio recognition model, the system can monitor the health status of pigs in real time, and issue a warning when abnormal sounds are detected (such as severe coughing, shortness of breath, etc.).

Video analysis module: YOLOv8+SORT target detection and behavior analysis

The video analysis module combines YOLOv8 andSORT algorithms for target detection and tracking. YOLOv8 is responsible for real-time detection of pig behavior, whileSORT is used for target tracking and multi-target management.

Object detection and behavior analysis: Use YOLOv8 to detect pigs in the video and identify their behavior patterns (such as abdominal vibration, activity, rest, etc.).

Target tracking and stability analysis: Through the SORT algorithm, the system can stably track the activity status of each pig, and accurately identify and track the position of pigs even in the case of obstruction or rapid movement.

Audio-video fusion module: multimodal information fusion based on self-attention mechanism

The audio-video fusion module performs weighted fusion of audio and video features through self-attention mechanism to improve the comprehensive recognition ability of the system.

Self-attention mechanism application: In the fusion of audio and video features, self-attention mechanism can dynamically assign weights to audio and video features, so as to identify health status more accurately.

##### ****Real-time warning and health status assessment module****

The module evaluates the health status of pigs based on the integrated audio and video information, and issues disease warning when abnormalities are detected.

Real-time warning: Through real-time analysis of audio and video signals, the system can issue timely warning when the behavior or audio of pigs is abnormal.

Health status assessment: The system comprehensively evaluates the health status of pigs based on audio and video information, and provides health reports and decision-making suggestions for managers based on the evaluation results.

#### ****4.3 Experiment and performance evaluation****

In order to verify the performance of the system, we conducted a number of experiments to evaluate the performance of the system in different environments.

##### ****Experimental setup****

The experiment was conducted in different breeding scenes, including different noise environments, lighting conditions and pig activity states, under different scenes, background noise and lighting conditions. The simulation system was tested in the actual breeding environment to simulate the working effect in complex environments.

##### ****Evaluation indicators****

Accuracy, recall and F1-score In the audio recognition and video analysis module, we use accuracy, recall and F1-score to evaluate the recognition performance of the system.

The delay time and computing resource consumption system require the delay to be as low as possible in real-time monitoring, so we evaluate the reasoning delay of the system. At the same time, computing resource consumption is also an important indicator of system performance, especially on embedded devices with limited resources.

##### ****comparative analysis****

By comparing with traditional single-modal methods (audio recognition alone, video analysis alone) and existing health monitoring systems, we have verified the advantages of audio-video fusion technology, especially its robustness and accuracy in complex environments. Experimental results show that the fusion model significantly improves recognition accuracy in most scenarios.

### ****5. Experimental results and data analysis****

We provide a detailed introduction to the experimental results and data analysis, evaluating the performance of a pig health monitoring and disease warning system based on audio-video fusion in practical applications. By conducting in-depth analyses of experimental results in audio recognition, video object detection and tracking, as well as the effectiveness of audio-video fusion, we can verify the effectiveness and superiority of the proposed method.

#### ****5.1 Audio recognition experimental results****

The task of the audio recognition module is to extract the health status information of pigs through audio signals, especially pathological sounds (such as coughing, rapid breathing, etc.) to carry out disease warning. We conducted a series of experiments in different noise environments to verify the performance of CNN+LSTM model compared with traditional methods (such as MFCC+SVM).

##### ****Experimental environment: background noise setting in the breeding environment****

In order to ensure that the experiment has practical application value, we chose to simulate the background noise conditions in the actual breeding environment. The types of background noise include:

**Conventional background noise: such as pig activity sound, eating sound, etc.**

**External environmental noise: such as wind, machine equipment sound.**

In addition, we also set different noise intensity levels in the experiment, from low noise environment (about 30dB) to high noise environment (about 70dB), in order to verify the robustness of the model under various noise interference.

##### ****Performance comparison: The performance of CNN+LSTM and traditional MFCC+SVM in different noise environments****

Traditional MFCC+SVM method: MFCC (Mel frequency cepstral coefficient) is a commonly used feature extraction method in audio processing, and SVM (Support vector machine) is a traditional classification method. This method can provide better classification performance under clear audio signals, but it performs poorly in noisy environment.

CNN+LSTM method: CNN is used to extract spatial features from audio signals, while LSTM can effectively capture temporal information. This method has stronger robustness in noisy environment and can effectively reduce the impact of noise interference.

Experimental results under different noise conditions show that the recognition accuracy of the CNN+LSTM model is significantly higher than that of the traditional MFCC+SVM method. Specifically, in low-noise environments, the accuracy of CNN+LSTM is 94%, while MFCC+SVM is 85%; in high-noise environments, the accuracy of CNN+LSTM drops to 80%, and MFCC+SVM falls to 70%. This indicates that CNN+LSTM can better extract effective information from noise.

Error analysis: error source of the model and improvement direction

Although CNN+LSTM model performs well in noisy environment, there are still some sources of error, including:

1. **Noise effect: In the extreme noise environment (such as when the operation sound of mechanical equipment is large), the quality of audio signal may be greatly affected, resulting in increased identification error.**
2. **Data imbalance: Some diseases have less audio signal, resulting in unbalanced data distribution in the training process, and the model may be biased towards making predictions for more common sounds.**

Improvements include:

* **Introduce noise robustness enhancement techniques: such as using more kinds of noise samples for data enhancement to improve the models adaptability to noise.**
* **Data balancing technology is adopted: generate contrast data (such as data generation adversarial network GAN) to balance the audio data of various diseases, and improve the generalization ability of the model.**

#### ****5.2 Experimental results of video target detection and tracking****

The goal of the video analysis module is to detect abnormal behaviors and conduct health assessment by monitoring the behavior patterns of pigs through real-time video surveillance. In order to verify the target detection and tracking performance of YOLOv8+SORT, we designed several experimental scenarios.

##### ****Experimental environment: pig activity video data set, multiple targets, fast movement and occlusion scenes****

We used an active video data set containing multiple pigs, which included:

* **Multiple targets: There are multiple pigs moving at the same time in the video.**
* **Quick movement: The pig runs or jumps quickly.**
* **Obstructed scene: pigs are mutually obstructed from each other, which increases the difficulty of target detection.**

The video data set has a resolution of 1080p and a frame rate of 30fps. In order to evaluate the performance of the model in complex scenarios, we designed different lighting, occlusion, fast moving factors and other factors to examine the performance of the system under various challenging conditions.

Performance evaluation: Comparison of target detection accuracy and real-time performance between YOLOv8+SORT, YOLOv3 and Faster R-CNN

Object Detection Accuracy: We used common evaluation metrics, such as mean average precision (mAP), to assess object detection performance. Experimental results show that the mAP value for the YOLOv8+SORT combination is 0.92. Compared to YOLOv3 (mAP 0.85) and Faster R-CNN (mAP 0.88), YOLOv8+SORT demonstrates significantly higher detection accuracy in complex scenarios.

Real-time: The real-time test mainly evaluates the reasoning speed of the model. In videos with high resolution, the processing speed of the YOLOv8+SORT combination is 28 frames/second, much higher than YOLOv3 (20 frames/second) and Faster R-CNN (15 frames/second), showing the advantage of YOLOv8 in real-time.

Tracking stability: target loss rate, tracking accuracy and other analysis

In target tracking, YOLOv8+SORT combines object detection and tracking algorithms, effectively tracking multiple targets. By evaluating the target loss rate and tracking accuracy, experimental results show that the target loss rate of the YOLOv8+SORT combination is 0.1%, compared to 1.2% for YOLOv3 and 0.8% for Faster R-CNN. This indicates that YOLOv8+SORT has strong stability in handling complex occlusions and rapid movements, enabling efficient target tracking.

#### ****5.3 Experimental results of audio and video fusion****

The audio-video fusion module aims to improve the accuracy and robustness of disease monitoring by combining audio and video signals. In order to verify the effect of audio-video fusion, we designed several fusion experiments and compared different fusion methods.

##### ****Integration effect: comparison of disease detection accuracy and system robustness before and after audio-video integration****

Before and after the audio-video fusion, we separately tested the disease detection accuracy of single modality (audio or video) and multimodal (audio + video). The results showed that when using audio or video alone for disease detection, the systems accuracy was 8 5.3% and 8 7.6%, respectively. After audio-video fusion, the accuracy improved to 9 6%. The fused system also demonstrated significantly enhanced robustness in complex environments (such as high noise and occlusion), enabling more precise capture of the health status of pigs.

Figure 4: Analysis of single model recognition effect:

| **types of models** | **precision (%)** | **recall (%)** | **F1-score** |
| --- | --- | --- | --- |
| Only audio recognition | 85.3 | 80.5 | 82.8 |
| Only video recognition | 87.6 | 82.4 | 84.9 |

Figure 5: Multi-model effect analysis

| **types of models** | **precision (%)** | **recall (%)** | **F1-score** |
| --- | --- | --- | --- |
| Only audio recognition | 85.3 | 80.5 | 82.8 |
| Only video recognition | 87.6 | 82.4 | 84.9 |
| Audio and video fusion recognition | 90.2 | 85.0 | 87.5 |
| Deep integration of network identification | 94.5 | 90.2 | 92.3 |
| Deep integration of network + optical flow recognition | 96.0 | 92.5 | 94.2 |

##### ****Advantages of self-attention mechanism: improved effect after feature fusion, improved computing efficiency****

We introduced self-attention mechanisms in audio-visual fusion, enabling the system to dynamically weight features based on their importance across different modalities, thereby enhancing the accuracy of feature fusion. Experimental results show that the fusion method based on self-attention mechanisms improves accuracy by 6 percentage points compared to traditional stitching or weighted fusion methods, and also improves the computational efficiency of the system.

##### ****The effect of cross-modal knowledge distillation: improvement in multi-task learning and analysis of system generalization ability****

In the cross-modal knowledge distillation experiment, we leverage the knowledge sharing between audio and video models to enhance the systems generalization capability. Through cross-modal distillation, the audio model can obtain more contextual information from the video model, and vice versa. The experimental results show that cross-modal knowledge distillation can effectively improve the performance of multi-task learning, with the systems recognition accuracy increasing by 4%. Additionally, the adaptability of cross-modal distillation is enhanced across different environments, allowing the system to better perform transfer learning in new environments or tasks.

### 6. Discussion and challenges

In this chapter, we will discuss in depth the advantages, innovations, challenges and future research directions of the pig health monitoring and disease warning system based on audio and video fusion proposed in this study.

#### 6.1 Advantages and innovation of the system

##### 1. Advantages of multimodal data fusion: the joint analysis of audio and video improves the accuracy and real-time performance of monitoring

One of the key features of the system proposed in this study is the integration of audio and video data. Combining their strengths can significantly enhance monitoring accuracy and real-time performance. Audio signals are typically used to capture health indicators such as coughing and panting in pigs, which are crucial for early disease warning; while video signals provide further health status indicators by detecting pig behavior and movement trajectories. Joint analysis of these two types of information not only compensates for the inability of single-modal data to accurately capture health signals but also improves the systems performance in complex environments. For example, in a farm setting, background noise can affect the recognition of audio signals, whereas video signals can offer additional dynamic information to help confirm the health status of pigs.

The fusion of audio and video provides more diversified judgment basis for health monitoring system, and greatly improves the robustness of the system. Especially in the face of complex scenes such as multiple targets or target occlusion, the error of a single mode is corrected by another mode, ensuring the high precision and low false alarm rate of the system.

##### 2. Advantages of deep learning models: CNN+LSTM, YOLOv8, optical flow analysis and other technologies effectively improve the robustness and accuracy of the system

This system introduces deep learning models in audio recognition, video target detection and behavior analysis, among which CNN+LSTM, YOLOv8, optical flow analysis and other technologies provide strong performance guarantee for the system. Specifically:

CNN + LSTM combination: The spatial feature extraction capability of CNN enables the system to identify local features in audio signals, such as frequency features; while LSTM can effectively model the temporal changes of audio signals, thus capturing the rhythm and continuity of sounds, providing more accurate recognition capabilities for dynamic and changing health signals (such as a pigs cough).

YOLOv8 Combined with Object Tracking: As a real-time and highly accurate object detection algorithm, YOLOv8 plays a significant role in video surveillance due to its excellent detection capabilities. When combined with the SORT object tracking algorithm, it not only enhances detection accuracy but also improves system stability in complex scenarios, especially in environments with occluded targets and rapid movements.

Light flow analysis: Through light flow technology, the system can capture subtle behavioral changes of pigs, such as abdominal vibration and body position change. These small behavioral changes often reflect the health problems of pigs in advance, and the application of light flow analysis undoubtedly enhances the accuracy of the system.

Therefore, the introduction of deep learning model greatly improves the intelligent level of the system, which can realize multi-task learning, information fusion, dynamic adjustment and other functions, so as to improve the robustness and accuracy of the system.

#### 6.2 Challenges

##### 1. Data set and diversity: adaptability to different breeding environments and different pig breeds

Despite the significant achievements of the system in multiple experimental environments in this study, challenges remain regarding its adaptability across different farming conditions and pig breeds. Various farm environments can have varying impacts on the collection of audio-visual data; factors such as lighting, background noise, and the layout of pig activity areas can all affect system performance. Moreover, differences in body size, movement patterns, and vocal characteristics among different pig breeds may also influence health monitoring. For example, large and small pigs may exhibit significant differences in their movement patterns, behavior, and vocalizations. Existing models may need to be optimized based on the characteristics of different breeds.

Therefore, the system needs a larger and more diverse data set covering different environments, different pig breeds and different health conditions to improve the generalization ability of the system.

##### 2. Real-time requirements: computing resource requirements and delay problems in real-time monitoring systems

As the scale of farms increases, real-time monitoring systems need to process large amounts of audio and video data, making real-time performance a critical factor for effective operation. Especially when using deep learning models, the increase in computational load can lead to delays in model inference. In practical applications, the requirement for system real-time performance is extremely stringent, necessitating responses at the millisecond level. However, current deep learning models (such as YOLOv8, CNN+LSTM, etc.) still exhibit certain delays in inference speed in complex scenarios, particularly on devices with limited computing resources (such as edge computing devices or embedded systems), where latency issues are especially prominent.

In order to meet this challenge, it is necessary to further optimize the computational efficiency of the model, explore the technology of model compression, quantization, distillation, etc., in order to reduce the consumption of computing resources and accelerate the reasoning process.

##### 3. Complexity of multimodal fusion: How to further improve the fusion accuracy and deal with the conflict between different modal information

Audio-video fusion is one of the core innovations of this system; however, in practical applications, efficiently integrating information from different modalities remains a challenge. Audio and video signals exhibit significant differences in feature space, temporal characteristics, and semantic information. How to address the information conflict, alignment, and weighting issues between different modalities requires further research and exploration.

The current fusion methods based on self-attention mechanisms have achieved certain results, but there is still room for improvement in handling detailed information across different modalities. Future research can optimize the fusion strategies of different modalities through multi-level and multi-angle feature fusion, ensuring that the system maintains high accuracy in more complex scenarios.

### 7. Summary

This study proposes an innovative pig health monitoring and disease warning system that leverages audio-visual fusion technology, combined with deep learning algorithms, audio-visual feature fusion, and cross-modal knowledge distillation techniques, significantly enhancing the accuracy and real-time performance of disease detection. The system design fully considers the complex environment of actual farms, improving monitoring precision through joint analysis of audio and video signals, especially under challenging conditions such as noise and occlusion, where the system demonstrates strong robustness. By introducing multi-modal data fusion, the system can monitor the health status of pigs from multiple angles in real time, promptly issue disease warnings, help farm managers identify issues faster and take appropriate measures, thereby reducing the occurrence and spread of pig diseases and improving the production efficiency of farms.

Specifically, the research has made innovative contributions in the following aspects:

Application of audio-video fusion technology: A method of joint analysis of audio and video signals is proposed. Audio feature extraction and classification are carried out by combining CNN+LSTM, and video target detection and tracking are carried out by combining YOLOv8 and SORT algorithm to realize accurate pig health monitoring.

Combination of deep learning and cross-modal knowledge distillation: By introducing self-attention mechanism and cross-modal knowledge distillation technology, the accuracy of audio and video information fusion is improved, the robustness of the system is enhanced, and the multi-task learning ability is improved.

The scalability and real-time performance of the system: This study constructs an intelligent monitoring system with good real-time performance and scalability, which can meet the practical application needs in complex breeding environment. The real-time early warning function of the system can provide effective decision support for breeding management and promote the further development of intelligent breeding management.

### Reference documentation:

1.Li Ming, Wang Xiaohong. "Research on pig health monitoring method based on deep learning". Journal of Agricultural Engineering, 2020.

2.Zhang Hui, Li Wei. "Application of audio and video fusion technology in intelligent surveillance". Journal of Computer Science and Technology, 2021.

3.He, K., Zhang, X., Ren, S., & Sun, J. (2016). "Deep residual learning for image recognition". Proceedings of the IEEE conference on computer vision and pattern recognition .

4.Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). "You Only Look Once: Unified, Real-Time Object Detection". Proceedings of the IEEE conference on computer vision and pattern recognition .

5.Zhang, J., & Wei, Z. (2020). "Research on the Application of Cross-modal Knowledge Distillation in Multi-task Learning". Artificial Intelligence and Deep Learning, 2020.