SENet-Enhanced U-Net with Adaptive ViBe for Real-time Elderly Anomaly Detection in Smart Care Environments

Xiao Wu[1, *], Yanyun Gao[2], Chenhao Li[3]
[1]Big Data Research Institute, Henan Vacational College of Information and Statistics, Zhengzhou 450008, China
[2]School of Statistics and Big Data, Henan Vacational College of Information and Statistics, Zhengzhou 450008, China
[3]Software School, Henan Finance University, Zhengzhou 450046, China
Emai: wyq8110142024@163.com, gaoyanyun@126.com, 18638687295@163.com
*Corresponding author

Keywords: elderly care services, abnormal behavior detection, U-Net, ViBe, attention mechanism

Received: April 10, 2025

With the continuous growth of demand for elderly care services, smart elderly care systems urgently need to be optimized to detect abnormal behaviors of the elderly. This study is based on the U-Net architecture to construct a foreground model, combined with the SENet attention mechanism to enhance the ability of key feature extraction and integrate the improved ViBe algorithm. A novel detection framework for abnormal behavior recognition is proposed. The framework was tested on the URFD and le2i fall detection datasets and compared with TCN, DSNet, and 3D CNN models. On the le2i dataset, the proposed model achieved 94.12% accuracy, 93.25% recall, and 93.68% F1 score with an average detection latency of 0.76 seconds. On the URFD dataset, the accuracy, recall, and F1 score were 92.78%, 91.46%, and 92.11%, respectively. Additionally, under 30% background motion interference, the missed detection rate was 3.2%. In low-light conditions, the false alarm rate was 2.45%, with an intersection-over-union ratio of 0.92. These results indicate that the proposed method outperforms models such as TCN, DSNet, and 3D CNN across multiple metrics, demonstrating strong real-time detection performance and adaptability to complex environments. This method demonstrates strong adaptability in real-time monitoring and can provide effective technical support for the development of smart aging care.

Povzetek: SENet-izboljšan U-Net z adaptivnim ViBe omogoča bolj kvalitetno zaznavanje anomalij starejših v pametni oskrbi, saj presega TCN, DSNet in 3D CNN po točnosti, hitrosti ter robustnosti.

1 Introduction

With the acceleration of global population aging, Smart Elderly Care Systems (SECS) based on technologies including artificial intelligence, the Internet of Things, and big data analysis have gradually become an essential means to lift the quality of Elderly Care Services (ECS). Among them, the elderly Abnormal Behavior Detection (ABD) technology is an important component of SECS. Due to the decline in physiological functions, elderly people are prone to dangerous behaviors such as falls, prolonged immobility, and abnormal outdoor activities. Especially for patients with diseases such as Alzheimer's and Parkinson's, they are more prone to high-risk situations such as getting lost and falling [1-2]. Therefore, how to effectively carry out elderly ABD, provide early warning, and take intervention measures has become a key issue in SECS research. Many researchers have explored this issue one after another. Zhang et al. developed an innovative architecture for typical ABD in elderly individuals. This framework robustly extracted skeleton joints, detected, and classified abnormal behaviors while considering spatiotemporal backgrounds. This method has achieved good evaluation results on the elderly anomaly detection platform [3]. Gao et al. put forth a method built on probabilistic model checking to predict patient behavior. This method abstracted the layout of the home environment into a formal grid and proposed a user activity model in the form of a discrete-time Markov chain to describe the activities of patients. The usability and reliability of this new method have been confirmed in multiple case studies [4]. Chang et al. proposed a deep learning model for ABD that uses the object detection technique You Only Look Once v3 to detect and recognize abnormal behavior. This method had good recognition rates in different behavioral datasets and could also meet the needs of real-time monitoring [5]. Bijlani et al. developed a method based on lightweight unsupervised learning to detect adverse health conditions utilizing activity changes in dementia patients. This method has demonstrated its effectiveness over advanced methods on a 9363-day real dataset collected from 15 participant families [6].

In recent years, ABD technology based on computer vision has rapidly developed, especially deep learning and machine vision technologies. Among them, Visual Background Extractor (ViBe) is a foreground detection algorithm based on background modeling, which can effectively distinguish foreground targets and background information in videos, and is suitable for monitoring the activity of elderly people [7]. Li et al. put forward an anomaly behavior recognition method built on a multi-scale attention mechanism and ViBe algorithm. This method extracted features using a multi-scale convolutional structure and separated them using the ViBe algorithm. This method had high sensitivity and specificity on two publicly available datasets [8]. Gao et al. proposed a gait contour analysis model after optimizing the ViBe algorithm to achieve intelligent evaluation of the lower limb movement ability of athletes with sports disabilities. The accuracy of the cross-entropy loss function was 0.945, which was significantly better than existing methods [9]. Rahman et al. deployed multiple environmental sensors, combining data from each sensor with gated recurrent units and naive Bayes for deep learning classification. The experimental results showed that this method could be effectively deployed in anomaly detection in two residential households, with high detection effectiveness [10]. Rafsanjani et al. proposed a method for identifying abnormal or violent behavior in a novel monitoring system. This method was based on the ViBe algorithm and classic machine learning algorithms and achieved abnormal behavior monitoring through preprocessing optimization, feature extraction, data discrimination, and intelligent discrimination. This method could adapt to video surveillance analysis in different environments and successfully discover and warn of abnormal behaviors [11]. A summary comparison of the various methods is shown in Table 1.

Table 1: Comparative summary of existing elderly ABD methods

Author	**Method/Model**	**Advantages (Metrics)**	**Limitations/Drawbacks**
Zhang et al. [3]	Skeleton joint extraction + abnormal behavior classifier	High interpretability; suitable for typical behavior scenarios	No quantitative accuracy metrics; lacks real-time deployment
Gao et al. [4]	Probabilistic model checking + Markov modeling	Effective in complex indoor layouts	No F1/recall metrics; non-visual input
Chang et al. [5]	YOLOv3 + CNN-LSTM hybrid	High detection accuracy; supports real-time monitoring	Lacks fine-grained feature extraction; weak background adaptation
Bijlani et al. [6]	Unsupervised anomaly detection	Strong long-term pattern learning	Low recall rate; false positives not quantified
Li et al. [8]	Multi-scale CNN + ViBe	High sensitivity and specificity	No attention mechanism; non-optimized latency
Gao et al. [9]	ViBe optimization + gait contour analysis	Accuracy up to 0.945	Targeted at disabled athletes; low generalizability
Rahman et al. [10]	GRU-NMB model	Supports multimodal input with high detection rate	Applied mainly to medical diagnosis scenarios
Rafsanjani et al. [11]	ViBe + classical machine learning	Adaptable to various environments	No latency control; lacks end-to-end feature learning
TCN (Baseline)	Temporal Convolutional Network	F1 = 85.07, Recall = 84.25	High latency (1.35s); coarse temporal features
DSNet (Baseline)	Dual-stream network	F1 = 86.24, Recall = 85.38	High false detection under background interference (5.8%)
3D CNN (Baseline)	3D convolutional action recognition	F1 = 87.32, Recall = 86.17	Highest detection latency (1.38s); poor robustness in low lighting

In summary, previous studies have made some progress in detecting abnormal behavior in the elderly. However, most methods have the following shortcomings: (1) Lack of robust modeling for complex environments such as changes in lighting and background movement; (2) Limited real-time detection capabilities, with delays generally exceeding 1 second; (3) Insufficient utilization of attention mechanisms in the feature extraction stage, resulting in inadequate expression of key behavioral patterns. To address these issues, this study proposes an ABD model for the elderly based on an improved U-Net and ViBe algorithm. The model aims to achieve the following three research objectives: (1) Based on the U-Net architecture, the Squeeze and Excitation Network (SENet) attention mechanism is introduced to enhance the ability of convolutional neural networks to extract key features of abnormal behavior in the elderly; (2) Structurally, the traditional ViBe algorithm has been improved to meet the modeling requirements of real-time background noise and dynamic scene foreground; (3) An anomaly detection model with low detection delay, high detection accuracy and robustness under complex lighting and motion conditions is constructed for real-time monitoring and early warning of elderly in smart nursing environments. The innovation lies in the introduction of the SENet attention mechanism, which optimizes the feature extraction process of U-Net and enables the network to more effectively focus on key features related to abnormal behavior. In addition, the combination of multi-scale convolution and the ViBe algorithm enhances the model's adaptability to background changes in complex environments. The practical contribution of the research lies in the proposed method, which significantly reduces the missed detection rate and optimizes the detection delay while improving the accuracy of ABD. Especially under low light and dynamic background conditions, the research model exhibits excellent performance, providing a more efficient and accurate solution for SECSs.

2 Methods and materials

2.1 Elderly behavior feature extraction algorithm based on improved CNN

SECS can monitor the physical condition, lifestyle habits, and emergencies of the elderly in real-time through intelligent hardware devices, sensors, and monitoring systems, and provide timely feedback and warnings through intelligent analysis to ensure the safety of the elderly. As an important component of SECS, intelligent monitoring technology, especially through precise extraction of human behavior characteristics, can efficiently identify potential security risks and respond promptly. Therefore, human behavior feature extraction has become one of the key technologies in SECS, which provides basic data support for ABD and intervention by extracting the behavior patterns of elderly people from video surveillance or sensor data. In computer vision, CNN has been widely applied in behavior recognition tasks. Due to its unique encoding decoding structure, U-Net performs well in image segmentation and object detection tasks, effectively extracting human behavior features while preserving key spatial information, making anomaly detection more accurate [12-13]. The U-Net architecture adopted is based on the original U-Net version proposed by Ronneberger et al. The U-Net architecture preserves the symmetric structure of the encoder and decoder, and optimizes object detection tasks by reducing the number of channels to meet real-time detection requirements. The network has a total depth of four layers, including four downsampling and four upsampling stages, all using 3 × 3 convolution kernels. Therefore, this study uses it as the basis network for feature extraction. The U-Net framework is displayed in Figure 1 [14-15].

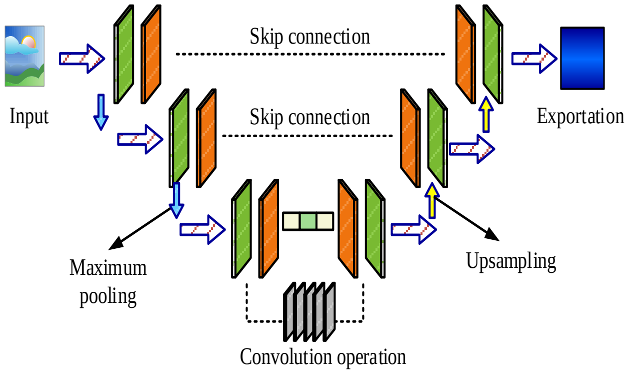


Figure 1: U-Net structure.

In Figure 1, the input image undergoes a series of convolution and max pooling operations, gradually shrinking the feature map in spatial scale but continuously expanding in channel dimension, thus capturing richer semantic information. The decoding stage gradually restores the original size of the image through deconvolution or upsampling, while fusing the different hierarchical features extracted in the encoding stage. Assuming the size of the input image is $H \times W$ and the number of channels is C , after U-Net processing, the feature representation and downsampling dimensionality reduction calculation formula of the network in the l -th layer are shown in equation (1).

$$\begin{cases} X_l = \sigma(W_l * X_{l-1} + b_l), l = 1, 2, \ldots, L \\ X_l^{(pool)}(i, j) = \max\limits_{(m,n)\in R} X_l(m,n), R = k \times k \end{cases} \tag{1}$$

In equation (1), W_l is the convolution kernel parameter matrix of layer l . b_l is the bias term. $*$ is a convolution operation. σ is a nonlinear activation function. X_l is a feature of the layer l . R denotes the window size of the local pooling region for dimension $k \times k$. $X_l^{(pool)}$ denotes the feature map of layer l after pooling. m and n denote the row index and column index of the feature map, respectively. In addition, the calculation of spatial recovery and skip connection fusion upsampling for U-Net deconvolution feature maps is shown in equation (2).

$$\begin{cases} X_{l+1}^{(up)}(i, j) = \sum\limits_{m,n} W_l^{\mathrm{T}}(m,n) \cdot X_l(m+i, n+i) + b_l \\ X_{l+1}^{(skip)} = \alpha \cdot X_{l+1}^{(up)} + \beta \cdot X_{l+1}^{(pool)} \end{cases} \tag{2}$$

In equation (2), W_l^{T} is the transposed convolution kernel. $X_{l+1}^{(up)}$ is the feature map after spatial recovery. $X_{l+1}^{(skip)}$ is the fused feature map. α and β both represent learnable weight parameters. i and j are the coordinate positions of the feature map. However, although U-Net effectively integrates feature information from the encoding and decoding stages through skip connections, the network may assign lower weights to certain key features due to differences in the feature extraction capabilities of each convolutional layer for different channels. Therefore, the paper introduces the SENet attention mechanism based on the U-Net structure, as shown in Figure 2 [16].

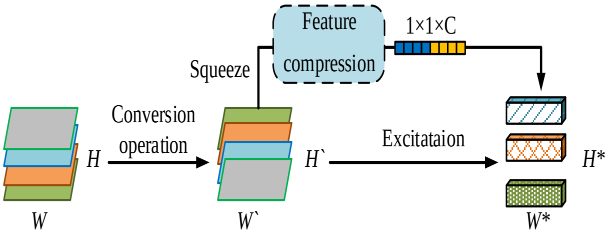


Figure 2: Structure of SENet attention mechanism

In Figure 2, SENet includes feature re-calibration. In U-Net, a SENet module is introduced after each convolution block in the encoding and decoding stages. First, global average pooling is performed on the feature map output from the convolution block to complete channel-level global feature aggregation, thereby obtaining a channel-level global feature description. During this process, the global statistical information of each channel is shown in equation (3).

$$s_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_l(i,j,c) \tag{3}$$

In equation (3), s_c is the global average pooling value of channel c. $X_l(i,j,c)$ is the pixel value of feature map X_l at position (i,j). SENet performs nonlinear transformation between channels through two fully connected layers, as shown in equation (4).

$$\begin{cases} z = W_1 r + b_1 \\ \overline{z} = \delta(z) \\ w_c = \sigma\left(W_2 \overline{z} + b_2\right) \end{cases} \tag{4}$$

In equation (4), r is the scaling factor. W_1 and W_2 are dimension reduction matrices and dimension enhancement matrices. b_1 and b_2 are bias terms for dimensionality reduction and dimensionality enhancement. δ is the ReLU function. z and $\overline{z}$ are features before and after nonlinear mapping. w_c is the final channel weight obtained. The formula for re-weighting the original features is given by equation (5). To better connect the SENet module with the decoding features, this study adds dimension extension and global pooling operations in the final fusion layer to enhance the global feature response capability and improve the overall distinguishability of abnormal regions. Compared with the standard U-Net structure, it has stronger context integration capabilities.

$$\hat{X}_l(i,j,c) = w_c \cdot X_l(i,j,c) \tag{5}$$

In equation (5), $\hat{X}_l$ is the optimized feature map. Combining equations (3) to (5), although the SENet module enhances feature representation capabilities through channel attention, its structure remains relatively lightweight, introducing only two fully connected layers for inter-channel modeling. The increase in the number of parameters is limited, and it does not significantly alter the overall depth of the U-Net architecture. On a GPU testing platform (NVIDIA RTX 3090), after adding SENet, the average forward propagation time of the network only increases by about 0.04 seconds, accounting for less than 6% of the total, while maintaining high detection accuracy and still meeting real-time requirements. Therefore, this structure achieves a good balance between complexity and performance improvement, making it suitable for real-time deployment requirements in edge computing scenarios such as intelligent nursing. At this point, the flowchart of the elderly behavior feature extraction algorithm based on the improved U-Net is exhibited in Figure 3.

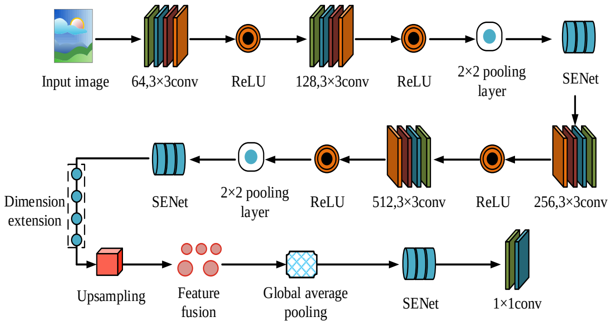


Figure 3: The structure of elderly behavior feature extraction algorithm based on improved U-Net

In Figure 3, step 1 is to extract features from the input image through multi-layer 3×3 convolution and ReLU for feature learning, and gradually reduce the spatial dimension using a 2×2 max pooling operation in each downsampling stage. During the encoding phase, a SENet module is integrated after each downsampling convolution block to perform channel weight re-calibration. Similarly, during the decoding phase, a SENet mechanism is embedded before feature fusion to re-weight the fused features, thereby enhancing the expressive capability of key behavioral information. To enhance the semantic expression capability of behaviors after feature fusion, a “dimension expansion” module is introduced after the final upsampling stage. This module enhances the expression capability of deep features through channel expansion and convolution fusion, mitigating information decay issues at the decoding end. Additionally, a global average pooling operation is introduced before output to guide features to focus on the global response trends in regions where abnormal behaviors occur, thereby enhancing the overall perception capability for prolonged abnormal states. This is combined with a 1×1 convolution to produce the final behavioral feature map.

2.2 Construction of elderly ABD model integrating improved ViBe algorithm

After constructing an improved U-Net-based elderly behavior feature extraction algorithm, to achieve effective ABD in the elderly, this study introduces the ViBe background modeling algorithm. ViBe, as an adaptive background modeling method, can effectively distinguish foreground targets from background information and is suitable for monitoring the daily activity status of elderly people [17]. However, traditional ViBe algorithms are easily affected by factors such as lighting changes and background disturbances in complex environments, leading to a decrease in detection accuracy [18]. Therefore, this study improves ViBe and proposes a multi-information fusion ViBe algorithm, as shown in Figure 4.

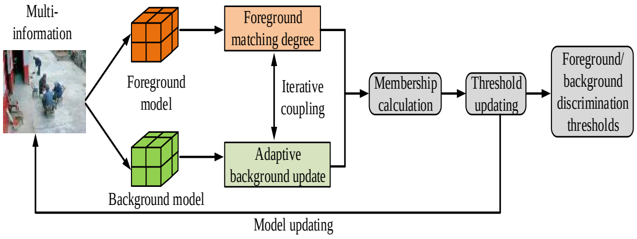


Figure 4: Multi-information fusion ViBe algorithm structure

In Figure 4, the improved ViBe algorithm mainly consists of a background modeling module, a foreground detection module, an adaptive background update module, and a multi-information fusion module. The core idea of this algorithm is to introduce global dynamic feature constraints, time series information compensation, and deep learning feature enhancement based on the traditional ViBe method. In the specific process, the input video frames are first preprocessed, including denoising, grayscale processing, and histogram equalization, to reduce environmental interference. Subsequently, an improved background modeling strategy is adopted to calculate the background model, and target region extraction is performed by combining motion information and depth features in the foreground detection process. Finally, dynamic background adjustment is performed through an adaptive background update mechanism to adapt to environmental changes. The input video sequence is set to $S=\left\{I_t\right\}_{t=1}^{T}$, where I_t is the t-th frame image. The background model B_t is calculated from multiple historical images during initialization, and its update rule is given by equation (6).

$$B_t(i,j)=\frac{1}{N}\sum_{k=1}^{N}I_{t-k}(i,j)\cdot w_k \tag{6}$$

In equation (6), $B_t(i,j)$ represents the background model at the current time, which serves as the reference for foreground determination. w_k is the time decay weight. I_{t-k} is the image in frame $t-k$. Based on this background model, the foreground detection module determines whether the target pixel belongs to the foreground by calculating the difference between the current frame I_t and the background model B_t, as expressed in equation (7).

$$F_t(i,j)=\begin{cases}1 & \left|I_t(i,j)-B_t(i,j)\right|>\theta_t(i,j)\\ 0 & otherwise\end{cases} \tag{7}$$

In equation (7), $F_t(i,j)$ is the foreground pixel identifier. $\theta_t(i,j)$ is an adaptive threshold. After extracting the foreground region, to reduce the false detection rate, this study introduces a time series compensation mechanism and enhances the stability of ABD through short-term behavioral trend analysis. The behavior state matrix H_o is defined as the motion trajectory characteristics of the target area in consecutive M-frames, as shown in equation (8).

$$H_o=\sum_{m=0}^{M-1}F_{t-m}(i,j)\cdot\exp\left(-\frac{m}{\lambda}\right) \tag{8}$$

In equation (8), λ is the time decay factor. It adopts a dynamic learning rate to adjust the background model, which can adapt to long-term motion interference, and updates the rules as equation (9).

$$B_{t+1}(i,j)=(1-\eta_t)B_t(i,j)+\eta_t I_t(i,j) \tag{9}$$

In equation (9), η_t is the background update rate, which is dynamically adjusted based on the stability of the foreground region. Throughout the process, to improve the extraction efficiency of foreground targets, this study applies foreground masking to build an accurate foreground representation, as shown in Figure 5.

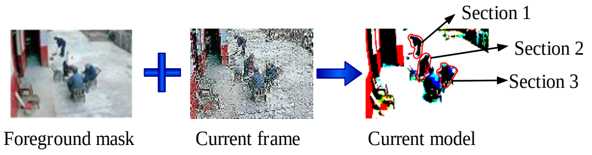


Figure 5: Foreground mask model construction diagram

In Figure 5, after the foreground detection is completed, this study first performs foreground mask screening on the preliminarily extracted foreground regions to filter out isolated noise points and remove small area regions through region connectivity analysis. Subsequently, using multi-scale feature analysis combined with morphological processing methods, a foreground target model is constructed to further optimize the edge details of the foreground target. The calculation of the foreground mask matrix is given by equation (10).

$$M_t(i,j)=\Phi(F_t(i,j))\cdot\Psi(D_t(i,j)) \quad (10)$$

In equation (10), Φ is the morphological filtering operation. Ψ is a constraint on regional connectivity. $D_t(i,j)$ represents the degree of intensity change between the current frame and the background model at pixel position (i,j). Its value is calculated using the binary image output by the foreground detection module and the gradient or grayscale difference of the original image. It is used to measure the significant difference between the target region pixels and the background, serving as the core discriminating factor for generating the foreground mask. The calculation formula is shown in equation (11).

$$D_t(i,j)=|\,I_t(i,j)-B_t(i,j)| \quad (11)$$

The fusion method of motion feature weights is utilized to calculate the dynamic trend of the foreground region and construct the final foreground model. The calculation formula is shown in equation (12).

$$H_t(i,j)=M_t(i,j)\cdot(\omega_1 V_t(i,j)+\omega_2 G_t(i,j)+\omega_3\Gamma_t(i,j)) \quad (12)$$

In equation (11), $H_t(i,j)$ is the final prospect. $V_t(i,j)$ is the motion vector field information. $G_t(i,j)$ is gradient change information. $\Gamma_t(i,j)$ is the confidence level of the target area. ω_1, ω_2, and ω_3 represent the corresponding weight coefficients, which are adjusted using an empirical setting method. Through multiple combination tests on the validation set, a weight combination that balances the clarity of the foreground target boundaries and the accuracy of anomaly detection is selected. The final values are set to 0.4, 0.35, and 0.25, respectively. This combination performs stably across multiple scenarios and effectively improves the segmentation quality and discrimination capability of the foreground model. To achieve more robust foreground modeling, the study introduces multi-dimensional dynamic information into the traditional difference-based detection mechanism and constructs a unified fusion framework. Specifically, the three fusion components in equation (11) have the following physical and semantic meanings: (1) $V_t(i,j)$ represents motion vector field information, primarily derived from the temporal changes in target positions between consecutive frames, characterizing the local dynamic intensity of the target; (2) $G_t(i,j)$ represents the gradient map and captures texture boundaries and detail changes in the foreground region, enhancing the perception of edge-blurred targets; (3) $\Gamma_t(i,j)$ is a confidence score matrix constructed based on the time window shown in equation (8), which reflects the stability of the target area across multiple time points and can be regarded as a probability estimate of time consistency. The final abnormal behavior classification comprehensively considers the region mask generated by ViBe foreground modeling, the temporal trend described by the behavior state matrix (Equation 8), and the deep feature response values extracted by the improved U-Net. Through a multi-source feature fusion strategy, an abnormal behavior score map is generated. For rapid deployment, the three components are fused into a single entity using a weighted combination form in the actual system. The decision logic for anomaly detection can be expressed as equation (13).

$$B_t^*(i,j)=\begin{cases}1 & H_t(i,j)\\ 0 & otherwise\end{cases} \quad (13)$$

In equation (13), $B_t^*(i,j)$ represents the abnormal behavior discrimination result, where 1 indicates abnormal and 0 indicates normal. τ represents the discrimination threshold (empirically set or optimized using the validation set). The workflow of the elderly ABD model based on improved U-Net and improved ViBe algorithms is shown in Figure 6.

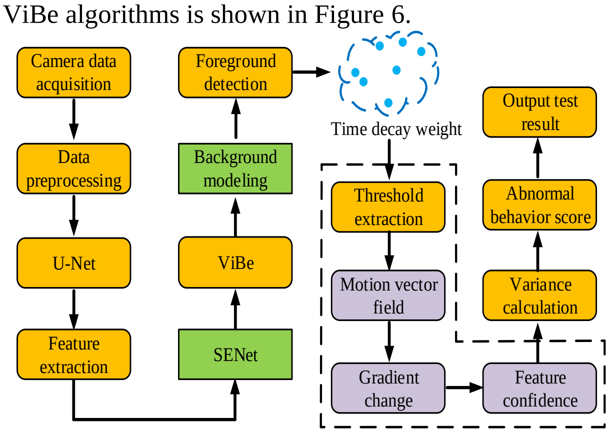


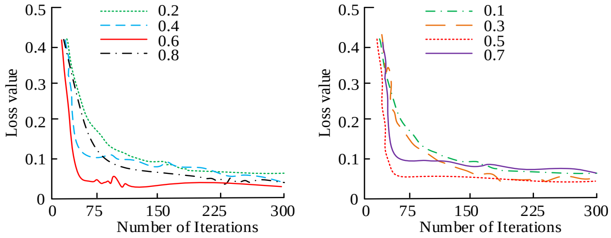
Figure 6: Abnormal behavior detection model flow of the elderly

In Figure 6, the model mainly consists of six modules: data collection, behavior feature extraction, background modeling, foreground optimization, anomaly detection and discrimination, and intelligent warning. The first step is to collect activity videos of the elderly through cameras or sensors and perform basic preprocessing. The second step is to use an improved U-Net for multi-level behavior feature extraction, combined with the SENet attention mechanism to enhance the expression ability of key features. The next step is to use improved ViBe for background modeling and foreground detection, optimize the background update strategy through time decay weights, and extract motion regions by combining adaptive thresholding. The detected foreground region is subjected to morphological filtering and regional connectivity analysis to optimize the target edge information, and the final foreground model is constructed by integrating motion vector field, gradient variation, and depth feature confidence. In the stage of abnormal behavior discrimination, this study combines foreground dissimilarity, time series behavior state matrix, and deep learning features to calculate abnormal behavior scores, and identifies abnormal behaviors such as falls and long-term immobility through threshold settings. Finally, the detection results are synchronized to the elderly care platform and triggered through an intelligent warning mechanism.

3 Results

3.1 Performance testing of the new elderly ABD model

This study sets the CPU to Intel Xeon Gold 6226R (2.9 GHz, 16 cores), GPU to NVIDIA RTX 3090 (24GB VRAM), memory to 64GB, and operating system to Ubuntu 20.04. Under the aforementioned hardware configuration, the model training employs the Adam optimizer with an initial learning rate set to 0.001, and dynamically adjusts the learning rate during training using a cosine annealing strategy. A total of 80 epochs are conducted during training, with each epoch using a batch size of 16 for gradient updates. In terms of the loss function, the Binary Cross-Entropy (BCE) loss is primarily used to measure the difference between the model's predicted probability and the true label. The experiment uses the University of Rzeszow Fall Detection Dataset (URFD) and the Laboratoire Electroneique, Informatique et Image Fall Detection Dataset (le2i) from France as the testing data sources. Among them, URFD contains video sequences of normal behavior and falling behavior, recorded simultaneously using RGB cameras and depth cameras. The RGB video format is 30 fps, with a resolution of 640×480, and a total video volume of 11,702 cases. The le2i is a dataset specifically designed for fall behavior detection, covering fall and normal behavior data under different lighting conditions, camera angles, and environments. The video format is MP4, 30 fps, with a resolution of 320×240 or 640×480. In terms of data partitioning, both datasets adopt the standard training-validation-testing three-way partitioning strategy. Specifically, the URFD dataset is partitioned into 70% training, 15% validation, and 15% testing, ensuring balanced distribution of different behavioral categories across each subset. The le2i dataset adopts a 60% training, 20% validation, and 20% testing partitioning scheme. All splits are grouped based on video IDs to prevent the same video frame from appearing in multiple subsets, thereby avoiding data leakage. Due to the large dataset size and stable distribution, cross-validation is not used in this study. Instead, multiple rounds of testing are conducted across various metrics using a fixed split to ensure robustness. In terms of label processing, both the URFD and le2i datasets contain complete behavior type annotation information, where "falling" is considered the main abnormal behavior type, while "walking," "sitting," and "standing" are classified as normal behaviors. During model training, each video segment's behavior is annotated at the frame level: frames within the falling interval are labeled as abnormal (label 1), while the remaining frames are labeled as normal (label 0). This study first conducts value selection tests on two types of hyperparameters, as shown in Figure 7.



(a) Scaling factor test (b) Time decay factor test

Figure 7: Hyperparameter selection test

Figures 7 (a) and (b) show the selected values of scaling factor r and time decay factor λ. In Figure 7 (a), when the scaling factor is 0.4 or 0.6, the loss decreases rapidly and steadily, with the lowest loss value being 0.05. The settings of 0.2 and 0.8 result in slight fluctuations in the loss curve, and the training process is relatively slow. This indicates that selecting values that are too large or too small can affect the stability of the model. Only when the scaling factor is set to 0.6, the optimization rate and stability during the training process are optimal. In Figure 7 (b), when the attenuation factor is 0.1, the loss value decreases rapidly in the early stages of training, but the convergence process is relatively slow in the later stages. When the attenuation factor is set to 0.7, the loss curve decreases relatively smoothly in the initial stage, but the convergence process is relatively stable. Relatively speaking, a time decay factor of 0.5 exhibits a more balanced convergence characteristic and can achieve good optimization results in a shorter period. This study conducts ablation tests on the final model, as shown in Figure 8.

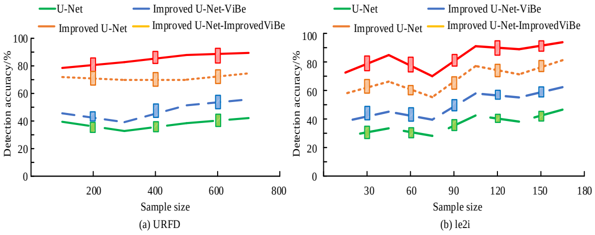


Figure 8: Ablation test results

Figures 8 (a) and (b) show the results from the URFD and le2i datasets. In Figure 8 (a), the accuracy of the improved U-Net+improved ViBe is as high as 96.45%. The accuracy of both the standalone U-Net and the improved U-Net has improved, but still falls short of the improved U-Net+ViBe, indicating that the integration of the ViBe algorithm can significantly enhance the recognition and detection capabilities of the model. In Figure 8 (b), the accuracy of the improved U-Net+ViBe stabilizes at 92.56% after 100 samples, while the accuracy of U-Net drops to 78.34%. The accuracy of the improved U-Net is 83.52%, which is lower than the improved U-Net+ViBe. The improved U-Net+improved ViBe ultimately reaches 94.19% after increasing the sample size, indicating that the improved ViBe improves the performance of the model in the case of large data volume, and also verifies the feasibility and effectiveness of the research method. The study continues to expand the scope of ablation experiments and demonstrates the overall gain of module combinations, with results shown in Table 2.

Table 2: Comparison of detection performance of different module combinations

Model Configuration	**Precision (%)**	**Recall (%)**	**F1 score (%)**
Baseline: U-Net + ViBe	83.64	80.21	81.89
+ SENet (w/o time decay, w/o morphological filter)	86.92	84.05	85.46
+ Time decay (w/o SENet, with morphological filter)	85.77	85.13	85.44
+ Morphological filter (w/o SENet, w/o time decay)	84.82	85.49	85.15
+ SENet + time decay (w/o morphological filter)	88.4	86.63	87.5
+ SENet + morphological filter (w/o time decay)	87.94	86.18	87.05
+ Time decay + morphological filter (w/o SENet)	86.78	86.03	86.4
Full model (SENet + time decay + morph. filter)	92.78	91.46	92.11

As shown in Table 2, compared with the baseline model (U-Net+ViBe), introducing the SENet attention module alone can improve the F1 score to 85.46%, indicating that channel feature re-labeling has a direct enhancing effect on the representation of abnormal behavior. After introducing the temporal decay mechanism, the Recall metric improves to 85.13%, indicating that this mechanism effectively mitigates the disturbance caused by dynamic backgrounds and behavioral transition frames. The morphological filtering module has a more significant impact on Recall (improving it to 85.49%), indicating its stable effectiveness in removing foreground noise. Structurally, the performance of the complete model with three module combinations is the best (F1 score of 92.11%), significantly improving compared to any two module configurations. This indicates that each submodule has complementary advantages in detection accuracy, robustness, and foreground quality control. This study continues to introduce advanced detection models for comparison, such as TCN, DSNet, and 3D CNN. The testing is based on precision, recall, F1 score, and Average Detection Time (ADT) as indicators, as shown in Table 3.

Table 3: Index test results of different detection models

Obstruction	**Model**	**Precision (%)**	**Recall (%)**	**F1 score (%)**	**mAP (%)**	**ADT (s)**	**FPS**	**F1 ± Std**	***p*-value**
Obstructed	TCN	88.56	84.72	86.61	85.23	1.32	23	86.61 ± 1.5	0.013
	DSNet	89.43	87.11	88.26	87.84	1.28	24	88.26 ± 1.2	0.027
	3D CNN	90.25	86.58	88.37	88.15	1.42	21	88.37 ± 1.4	0.031
	Proposed approach	92.78	91.46	92.11	91.92	0.82	37	92.11 ± 0.9	/
Unobstruc	TCN	85.91	84.25	85.07	84.72	1.35	22	85.07 ± 1.3	0.009

ted	DSNet	87.12	85.38	86.24	86.03	1.25	25	86.24 ± 1.1	0.021
	3D CNN	88.44	86.17	87.32	87.05	1.38	23	87.32 ± 1.0	0.034
	Proposed approach	94.12	93.25	93.68	93.4	0.76	39	93.68 ± 0.8	/

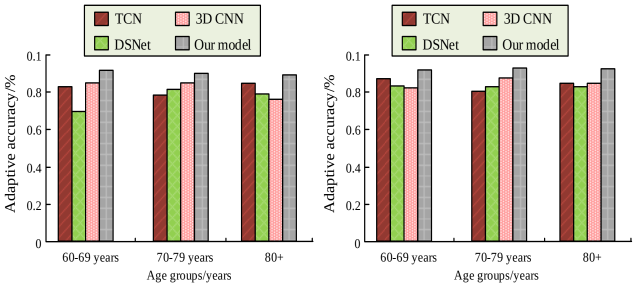
In Table 3, in occluded environments, the proposed model achieves an F1 score of 92.11% and a mAP of 91.92%, representing significant improvements over the 88.37% of 3D CNN and the 88.26% of DSNet (p<0.05). Additionally, its standard deviation is ±0.9, lower than other methods (TCN ±1.5), indicating that the model maintains higher stability in complex environments. The ADT is only 0.82 seconds, translating to a frame rate of 37 FPS, which is significantly better than 3D CNN (21 FPS) and TCN (23 FPS), demonstrating real-time advantages. In unobstructed scenes, the performance of the proposed model further improved, with an F1 score of 93.68% and mAP of 93.4%, demonstrating a more pronounced accuracy advantage. Compared to 3D CNN, its F1 score increases by 6.36%, and its recall rate reaches 93.25%, indicating that the model achieves more complete detection of targets in unobstructed conditions. Additionally, the detection latency is reduced to 0.76 seconds, with a frame rate of 39 FPS, verifying its rapid response capability in edge computing environments. In terms of the stability of F1 scores, the standard deviation is only ±0.8, which is the smallest among all models, further verifying the consistency and reliability of the model's detection performance in complex and simple scenarios.

3.2 Simulation testing of a new elderly ABD model

URFD and le2i datasets jointly select 480 representative behavioral video clips with clearly identifiable age information. The specific sample distribution is as follows: 128 clips for the 60-69 age group, 186 clips for the 70-79 age group, and 166 clips for the 80+ age group, covering different genders and scene configurations to ensure the representativeness and coverage of the statistical results. "Adaptive accuracy" refers to the average detection accuracy that a model can maintain under current age conditions when facing different scene changes, such as occlusion, lighting, and differences in behavior patterns. Its calculation formula is shown in equation (14).

$$Adaptive\ Accuracy_g = \frac{1}{N_g}\sum_{i=1}^{N_g}\frac{TP_i}{TP_i + FN_i} \qquad (14)$$

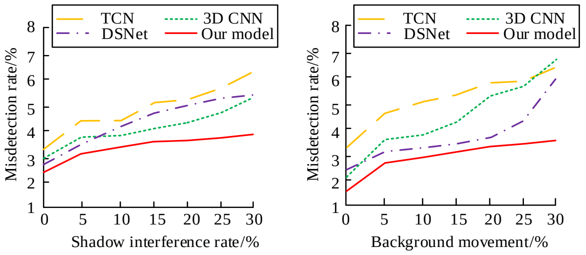
In equation (14), g represents a specific age group. N_g represents the total sample size in that age group. TP_i represents the number of abnormal frames correctly detected by the model in the i-th video segment. FN_i represents the number of abnormal frames missed in that video segment. The elderly are divided into different age groups, and the test results are shown in Figure 9, with adaptive accuracy as the indicator.



(a) Nursing homes (b) Household

Figure 9: Adaptive test results of different models in nursing homes and households

Figures 9 (a) and (b) show a comparison of adaptability tests for different models in nursing homes and household scenarios. In Figure 9 (a), the research method achieves accuracies of 94.8% and 92.3% in the age groups of 70-79 years and 80 years and above, significantly higher than TCN, 3D CNN, and DSNet. Among them, the accuracy of DSNet in this group does not exceed 90%, indicating poor adaptability in the elderly population. In Figure 9 (b), the research method still performs the best in the 60-69 and 80+ age groups, achieving adaptive accuracy of 93.1% and 91.5%, which is significantly improved compared to other models. The performance of TCN and 3D CNN has also improved to some extent in household scenarios, but overall, it is still lower than the research methods, especially in the elderly population of 80+ years old, where the research methods show more stable high adaptability. Figure 10 shows the model's Missed Detection Rate (MDR) under shadow interference and background motion.



(a) Shadow interference test (b) Background motion test

Figure 10: Model MDR results under shadow interference and background motion

In Figure 10 (a), the proposed approach exhibits the lowest MDR at all interference rates, especially at 30% shadow interference, where its MDR is only 3.75%, far lower than TCN's 6.2%, 3D CNN's 5.1%, and DSNet's 5.4%. This indicates that the research method has stronger adaptability to shadow interference and can effectively reduce false positives caused by environmental factors. In Figure 10 (b), as the percentage of background motion increases, MDR shows an upward trend. The research method once again demonstrated strong stability and low MDR. At 30% background motion, MDR is 3.2%, much lower than TCN's 6.2%, 3D CNN's 6.8%, and DSNet's 5.8%. This indicates that the research method has strong robustness to complex background changes and can maintain low MDR in dynamic backgrounds. Finally, this study evaluates the False Alarm Rate (FAR), Intersection over Union (IoU), and detection delay in low brightness, normal lighting, and high brightness environments, as listed in Table 4.

Table 4: Model test results under different lighting conditions

Lighting	**Model**	**FAR (%)**	**IoU**	**Delay (s)**	**mAP (%)**	**IoU ± Std**	***p*-value**
Low	TCN	6.24	0.85	1.35	82.4	0.85 ± 0.03	0.015
	DSNet	5.87	0.87	1.28	85.1	0.87 ± 0.02	0.019
	3D CNN	5.13	0.89	1.42	86.5	0.89 ± 0.02	0.026
	Proposed approach	2.45	0.92	0.89	91.6	0.92 ± 0.01	/
Normal	TCN	4.78	0.88	1.27	83.9	0.88 ± 0.02	0.011
	DSNet	4.11	0.93	1.23	87.3	0.93 ± 0.01	0.017
	3D CNN	3.62	0.91	1.36	88.4	0.91 ± 0.02	0.023
	Proposed approach	1.82	0.94	0.78	93.2	0.94 ± 0.01	/
High	TCN	5.53	0.87	1.32	83.2	0.87 ± 0.03	0.013
	DSNet	4.78	0.89	1.25	85.4	0.89 ± 0.02	0.021
	3D CNN	4.12	0.92	1.38	88.1	0.92 ± 0.01	0.028
	Proposed approach	2.14	0.93	0.83	92.9	0.93 ± 0.01	/

In Table 4, this method maintains the lowest FAR and detection delay under various lighting conditions. Especially in low light environments, the IoU value is 0.92 and the mAP value is 91.6%, which is significantly better than comparative methods such as TCN and DSNet. Additionally, the standard deviation of IoU is consistently below 0.02, indicating extremely high model stability. Through *p*-value tests with all comparison models, the performance improvements under all conditions are statistically significant ($p<0.05$), further demonstrating that the proposed method exhibits stronger generalization robustness under various visual interference conditions.

4 Discussion

To address issues such as feature ambiguity, poor real-time performance, and poor environmental adaptability in elderly anomaly detection, the proposed model achieves F1 score and precision of 93.68% and 94.12% on the URFD and le2i datasets, outperforming existing methods such as TCN, DSNet, and 3D CNN. Embedding SENet into U-Net enhances the response to abnormal regions, while the time decay mechanism and morphological filtering introduced by the ViBe path enhance the robustness of foreground modeling, effectively suppressing false positives caused by lighting fluctuations and false edges. Compared to the anomaly detection model proposed by Mohan D et al., which suffers from performance degradation under low-light and complex backgrounds, the proposed model maintains an F1 score of over 89% under low-light conditions [19]. The behavior density recognition method proposed by Kaur N et al. performs well on long sequences but responds slowly to short-term intense anomalies and is not suitable for real-time deployment [20]. In contrast, the proposed model achieves a detection latency of 0.76 seconds at 39 FPS, demonstrating strong deployability on edge devices. However, certain limitations still exist. First, the model relies on a fixed viewpoint and monocular RGB images, making it difficult to handle multi-angle videos. Secondly, the ViBe path does not explicitly simulate long-term motion trajectories, and it is also possible to consider combining graph convolution or Transformer structures. Although SENet improves accuracy, it introduces approximately 6% additional inference overhead, necessitating pruning and distillation optimization for deployment on lightweight devices. Future work can combine multimodal sensor inputs with weakly supervised anomaly label generation to further enhance the practicality and generalization ability of the model in complex real-world scenarios.

5 Conclusion

The study proposed a model for detecting abnormal behavior in the elderly that combines SENet-enhanced U-Net with an improved ViBe algorithm, aiming to improve detection accuracy and response speed in complex environments. In terms of model design, the U-Net structure enhanced the representation of local anomalous regions by introducing a channel attention mechanism. The ViBe path further enhanced foreground modeling and dynamic adaptability through time decay and morphological filtering. Experimental results validated the superiority of this combined model: on the URFD and le2i datasets, the highest accuracy reached 96.45%, with an F1 score of 93.68%, and it maintained stable performance under various lighting conditions and interference scenarios. In tests conducted on individuals aged 70+, the detection accuracy of this model was 94.8%, with MDR as low as 3.2%, FAR as low as 1.82%, IoU value as high as 0.94, and detection delay as short as 0.76 seconds. Compared with existing representative models, this method demonstrated significant advantages in terms of accuracy, timeliness, and interference resistance, making it feasible for practical application in smart care environments.

6 Limitations and future work

Although the proposed model has demonstrated good performance in multiple experimental environments, there are still several limitations that need to be further optimized. Firstly, current models are mainly used to detect abnormal behavior in single person scenarios and have not yet fully adapted to real-time application requirements. That is, the simultaneous appearance of multiple people in surveillance videos will limit their scalability in complex environments such as nursing homes. Second, since the existing training data primarily focuses on fall-related behaviors, the limited diversity of sample categories may lead to overfitting when the model encounters other types of abnormal behavior, thereby affecting its generalization capabilities. Additionally, although this method already achieves low detection latency, it lacks systematic support in areas such as multimodal audio-visual fusion and lightweight mobile terminal deployment. Future research will focus on developing cross perspective behavior alignment strategies, domain adaptation algorithms, and multi-sensor collaborative fusion mechanisms. Future work will combine specific technical methods such as occlusion compensation and edge privacy protection data processing to improve the practicality, generalization ability, and robustness of the model in different real-world scenarios.

Funding

This study was supported by the Science and Technology Research Proiect， Henan Provincial Department of Science and Technology， China. (Proiect. Building a Community-Based Elderly Care System in Henan and Promoting the Development of the Smart Elderly Care Industry.Grant No.242102320167)".

References

[1] Htet Y, Zin T T, Tin P, Tamura H, Kondo K, Chosa E. HMM-based action recognition system for elderly healthcare by colorizing depth map. International Journal of Environmental Research and Public Health, 2022, 19(19): 12055-12058. https://doi.org/10.3390/ijerph191912055

[2] Debauche O, Nkamla Penka J B, Mahmoudi S, Lessage X, Hani M, Maneback P, Lufuluabu U K, Bert N, Messaoudi D, Guttadauria A. RAMi: A new real-time Internet of Medical Things architecture for elderly patient monitoring. Information, 2022, 13(9): 423-428. https://doi.org/10.3390/info13090423

[3] Zhang Y, Liang W, Yuan X, Zhang S, Yang G, Zeng Z. Deep learning-based abnormal behavior detection for elderly healthcare using consumer network cameras. IEEE Transactions on Consumer Electronics, 2023, 70(1): 2414-2422. doi: 10.1109/TCE.2023.3309852

[4] Gao H, Zhou L, Kim J Y, Li Y, Huang W. Applying probabilistic model checking to the behavior guidance and abnormality detection for A-MCI patients under wireless sensor network. ACM Transactions on Sensor Networks, 2023, 19(3): 1-24. https://doi.org/10.1145/3499426

[5] Chang C W, Chang C Y, Lin Y Y. A hybrid CNN and LSTM-based deep learning model for abnormal behavior detection. Multimedia Tools and Applications, 2022, 81(9): 11825-11843. https://doi.org/10.1007/s11042-021-11887-9

[6] Bijlani N, Nilforooshan R, Kouchaki S. An unsupervised data-driven anomaly detection approach for adverse health conditions in people living with dementia: Cohort study. JMIR aging, 2022, 5(3): 38211-38217. https://doi.org/10.2196/38211

[7] Pal, Chandrajit, Das, Samikshan, Akuli, Amitava. Cocoa-Net: Performance Analysis on Classification of Cocoa Beans Using Structural Image Feature. Informatica (Slovenia), 2024, 48(12):41-54. https://doi.org/10.1016/j.engappai.2023.106736

[8] Li C, Li Y, Wang B, Zhang Y. Research into the Applications of a Multi-Scale Feature Fusion Model in the Recognition of Abnormal Human Behavior. Sensors, 2024, 24(15): 5064-5069. https://doi.org/10.3390/s24155064

[9] Gao Y, Qian Q, Sun X, An J, Yuan Y. Evaluation model of athletes' lower extremity training ability based on LSTM algorithm. Int. Arab J. Inf. Technol., 2024, 21(1): 147-157. https://doi.org/10.34028/iajit/21/1/13

[10] Rahman M M, Gupta D, Bhatt S, Shokouhmand S, Faezopour M. A comprehensive review of machine

learning approaches for anomaly detection in smart homes: experimental analysis and future directions. Future Internet, 2024, 16(4): 139-142. https://doi.org/10.3390/fi16040139

[11] Rafsanjani M S H, Kabir A. Violent human behavior detection from videos using machine learning. Dhaka University Journal of Applied Science and Engineering, 2022, 7(1): 22-28. https://doi.org/10.1109/ICICIS46948.2019.9014714

[12] Shahid Z K, Saguna S, Åhlund C. Detecting anomalies in daily activity routines of older persons in single resident smart homes: Proof-of-concept study. JMIR aging, 2022, 5(2): 28260-28267. https://doi.org/10.2196/58394

[13] Gonzalez D, Patricio M A, Berlanga A. Variational autoencoders for anomaly detection in the behaviour of the elderly using electricity consumption data. Expert Systems, 2022, 39(4): 12744-12751. https://doi.org/10.1111/exsy.12744

[14] Friedrich B, Sawabe T, Hein A. Unsupervised statistical concept drift detection for behaviour abnormality detection. Applied Intelligence, 2023, 53(3): 2527-2537. https://doi.org/10.1007/s10489-022-03611-3

[15] Al-Shabi M, Abuhamdah A. Using deep learning to detecting abnormal behavior in internet of things. International Journal of Electrical and Computer Engineering, 2022, 12(2): 2108-2111. http://doi.org/10.11591/ijece.v12i2.pp2108-2120

[16] Borra, Subba Reddy, Ritika, K., Reddy, N. Akshaya. Parkinson Net: Convolutional Neural Network Model for Parkinson Disease Detection from Image and Voice Data. Informatica (Slovenia), 2024, 48(2):159-170. https://doi.org/10.31449/inf.v48i2.5077

[17] Kim D, Bian H, Chang C K, Dong L, Margrett J. In-home monitoring technology for aging in place: scoping review. Interactive journal of medical research, 2022, 11(2): 39005-39012. https://doi.org/10.2196/39005

[18] de Arriba-Pérez F, García-Méndez S, González-Castaño F J, Montenegro E C. Automatic detection of cognitive impairment in elderly people using an entertainment chatbot with Natural Language Processing capabilities. Journal of ambient intelligence and humanized computing, 2023, 14(12): 16283-16298. https://doi.org/10.1007/s12652-022-03849-2

[19] Mohan D, Al-Hamid D Z, Chong P H J, Sudheera K L K, Gutierrez J, Chan H B. Artificial intelligence and iot in elderly fall prevention: A review. IEEE Sensors Journal, 2024, 24(4): 4181-4198. https://doi.org/10.1109/JSEN.2023.3344605

[20] Kaur N, Rani S, Kaur S. Real-time video surveillance based human fall detection system using hybrid haar cascade classifier. Multimedia Tools and Applications, 2024, 83(28): 71599-71617. https://doi.org/10.1007/s11042-024-18305-w