Human activity recognition and fall detection using video and inertial sensors

by **Zhanggir Yergaliyev**
Supervisor: **Professor Adnan Yazici**
Co-supervisor: **Professor Enver Ever**

July 2022

THESIS DEFENSE

Introduction

PART 01

- According to population-based ecological studies, the most common injuries to the elderly (65 and over) and two-thirds of all serious injuries to individuals are caused by falls.
- The rate of falls are 28-35% for the population over age of 65 and 32-42% for the population over 60 years.

Objectives

WHAT WE WANT TO ACHIEVE

- To be able to properly analyze and identify various activities performed by humans in home conditions
- To prevent some health issues caused by falls by identifying them correctly in early stages

Main contributions

1. State-of-the-art vision-based activity recognition models for several datasets.
2. State-of-the-art inertial sensor based activity recognition model on UP-Fall dataset.
3. First multimodal activity recognition model which recognizes falls.

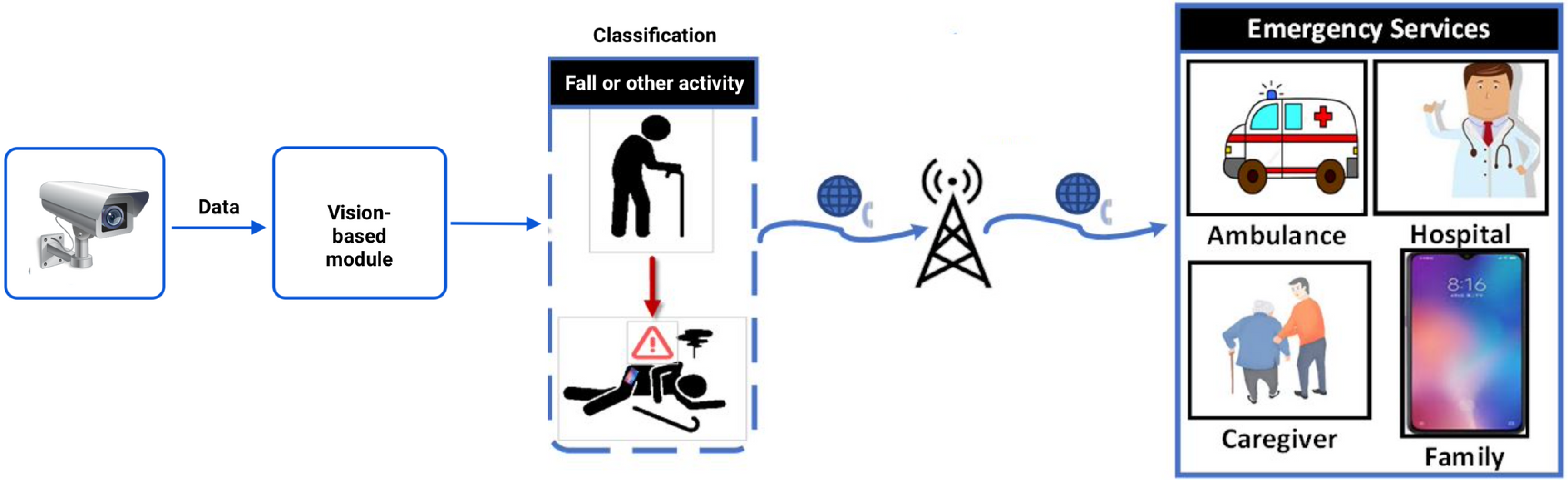
HAR approaches

Inertial sensor based

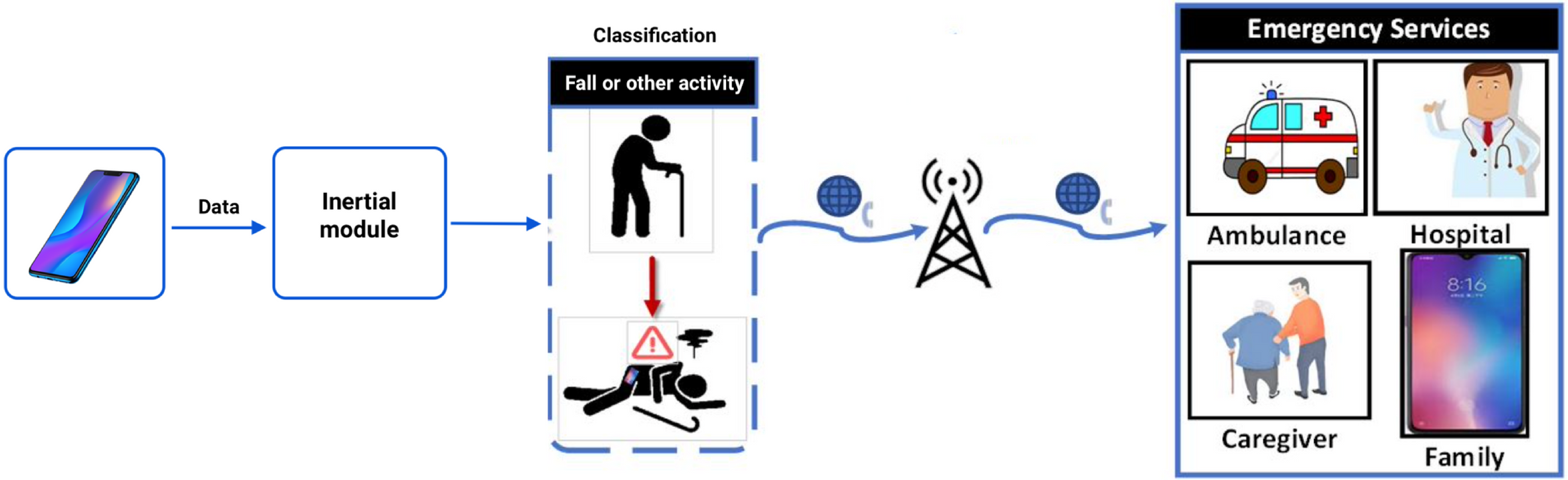
Video based

Hybrid

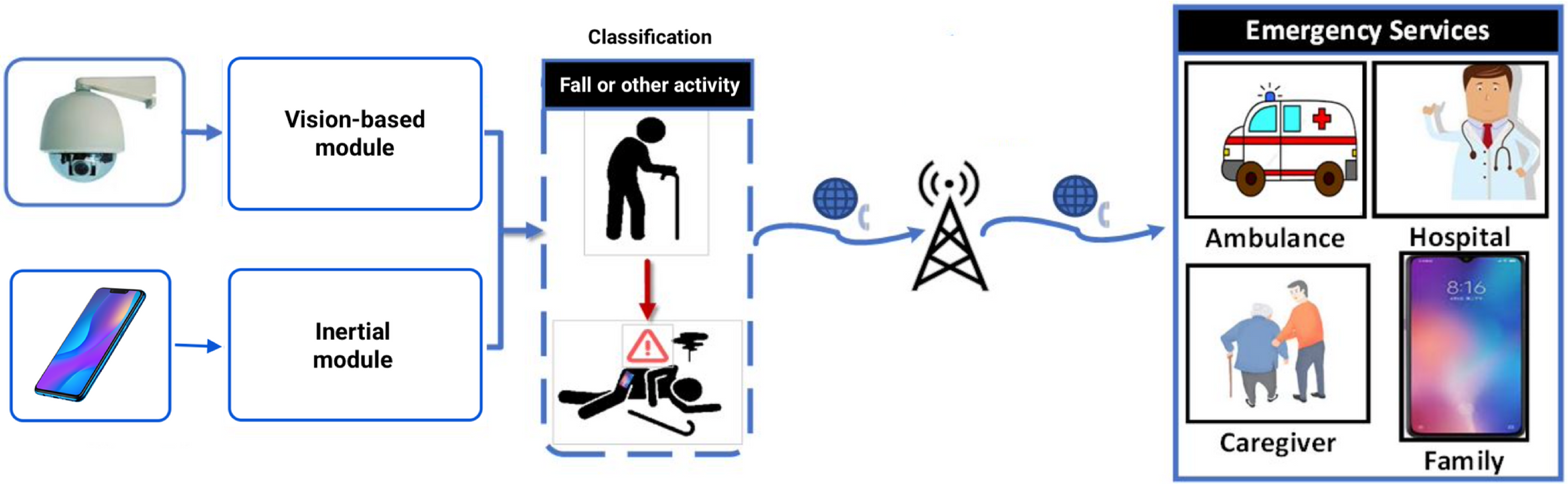
Overall structure of the proposed system (1)



Overall structure of the proposed system (2)



Overall structure of the proposed system (3)



Review of Related Literature

PART 02

Related Literature

HAR FROM VIDEO DATA

Reference	Year	Dataset	Classification Algorithm	Accuracy	Recognized actions
Amiri, et al.	2014	Collected data, DML SmartActions	SVM	58.20%	12 actions, including falls
Mehr, et al.	2019	DML SmartActions public dataset	CNN	82.41%	12 actions, including falls
Tsai, et al.	2020	NTURGB+D public dataset	3D ConvNet	90.79%	6 actions, including falls
Lv, et al.	2020	Collected data	LiteFlowNet	93.74%	5 actions, including falls

Related Literature

HAR FROM INERTIAL SENSOR DATA

Reference	Year	Dataset	Classification Algorithm	Accuracy	Recognized actions
Li, et al.	2019	Collected data	Bi-LSTM	96.00%	6 actions, including falls
Amara, et al.	2021	SisFall, UmaFall public datasets	LSTM	98.39%	4 actions, including falls
Alvarez, et al.	2017	USC-HAD, WISDM, Shoaib public datasets	Ameva algorithm	95.00%	7 actions, including falls

Related Literature

HYBRID HAR APPROACHES

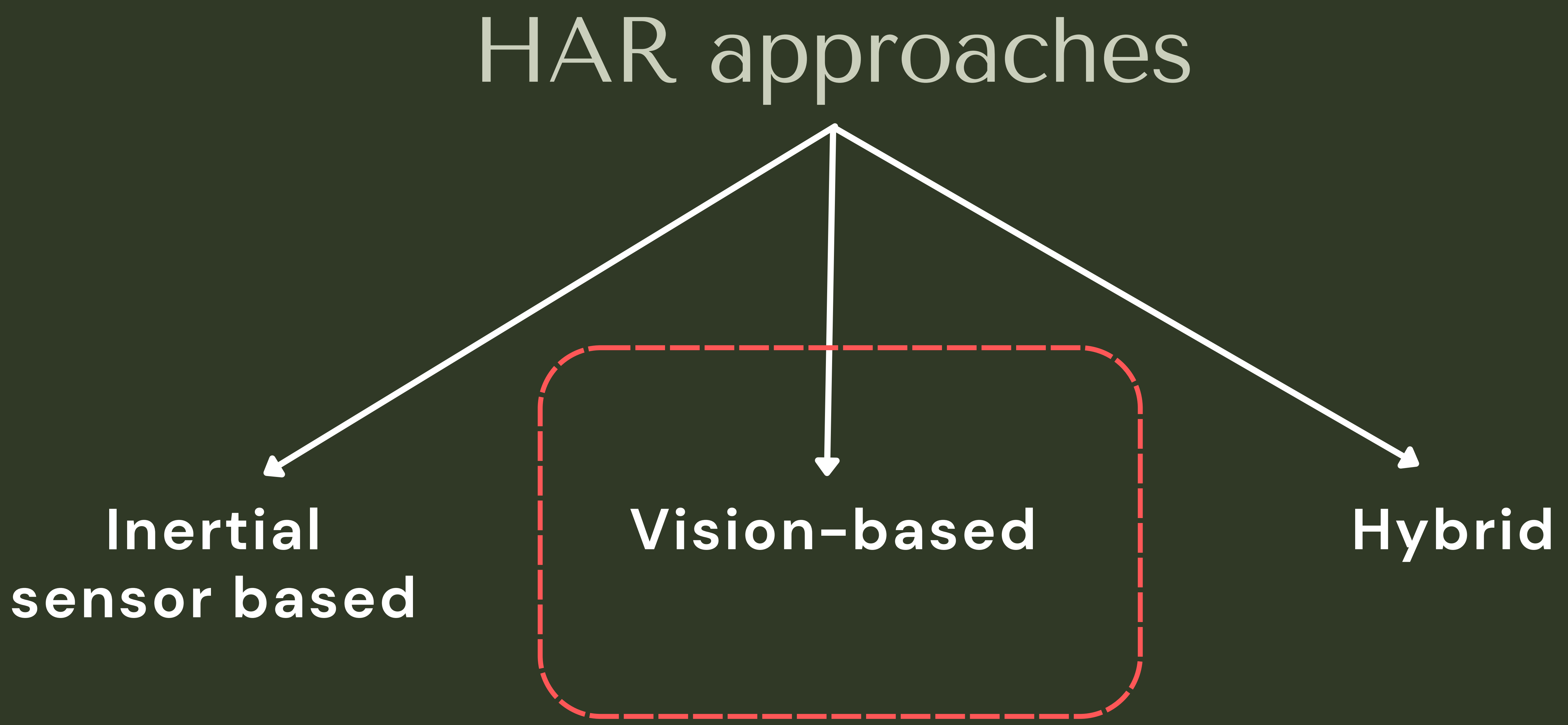
Reference	Year	Dataset	Classification Algorithm	Accuracy	Recognized actions
Kwolek, et al.	2014	UR Fall dataset	SVM	98.33%	Binary classification
Martínez-Villaseñor, et al.	2019	UP-Fall dataset	Random Forest, SVM, kNN, Multi-Layer Perceptron	95.88%	11 actions, including falls
Lee, et al.	2021	Collected data	RNN + CNN	100%	11 actions, including falls

Methods & Results

PART 03

Methods & Results

VISION-BASED MODULE



Datasets

MULTIMEDIA & HYBRID

falldataset

VISION-BASED

DMLSmartActions

VISION-BASED

UP FALL

HYBRID

Datasets

MULTIMEDIA & HYBRID

falldataset

VISION-BASED

DMLSmartActions

VISION-BASED

UP FALL

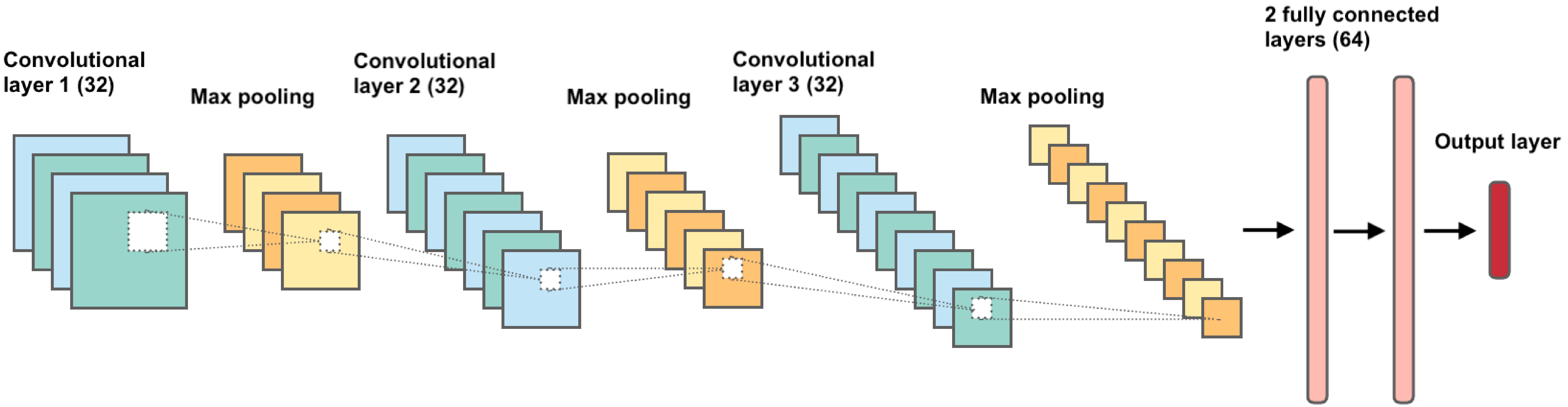
HYBRID

falldataset

Four classes:
Three of which correspond to common human body positions:
- ◈ Standing
- ◈ Sitting
- ◈ Sleeping (Lying)

One class which corresponds to an emergency:
- ◈ Falling down

CNN architecture - falldataset



Dealing with imbalanced dataset

Assigning class weights

Class weights were assigned based on the number of samples of each class.

Oversampling & undersampling

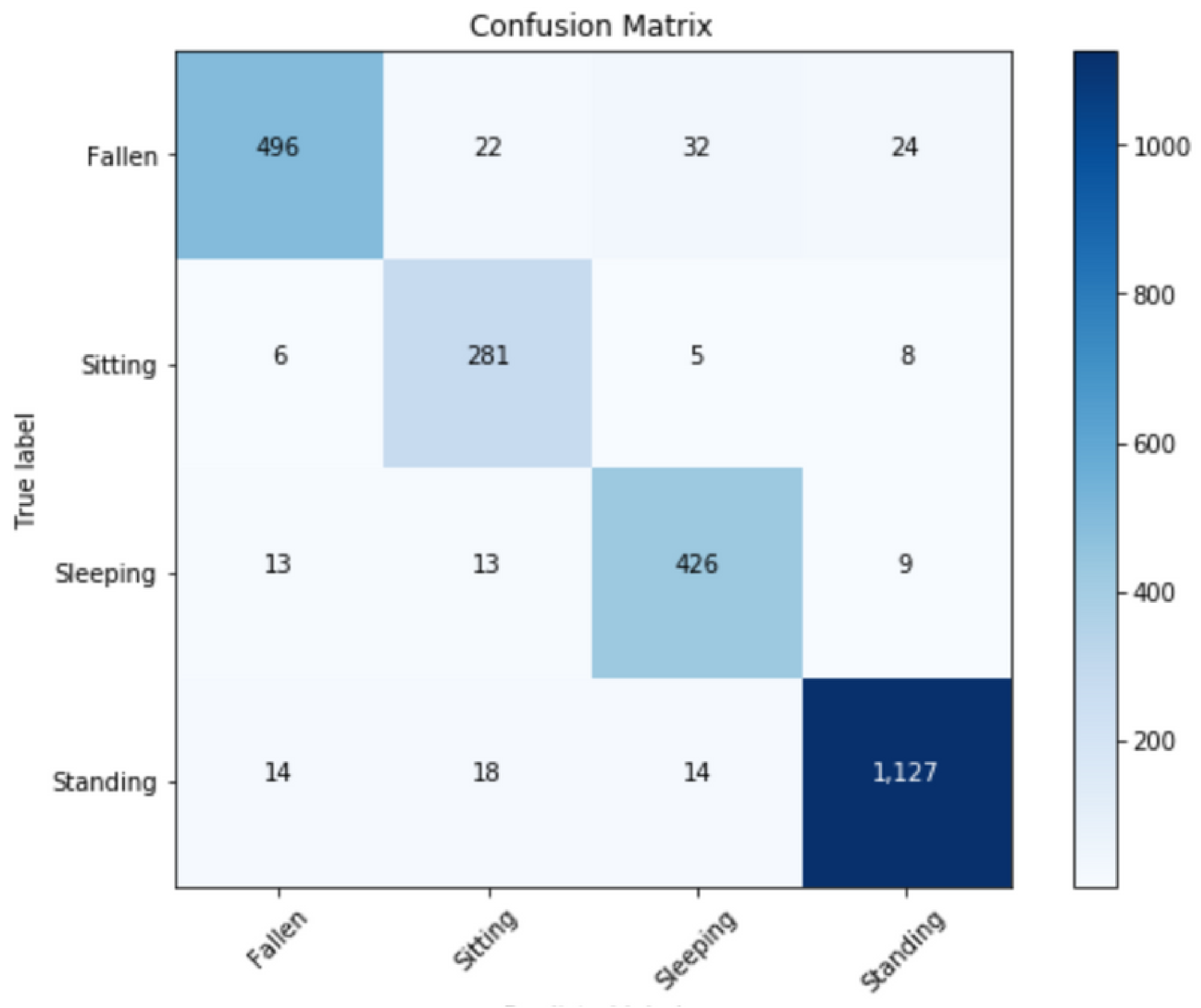
Oversampling for ‘Sitting’ class using PIL library: flipping all images horizontally
◈ Undersampling for ‘Standing’ class
◈ As a result, we have relatively balanced dataset: 2,300–2,900 images for each class

```
wj = n_samples / (n_classes * n_samplesj)
```

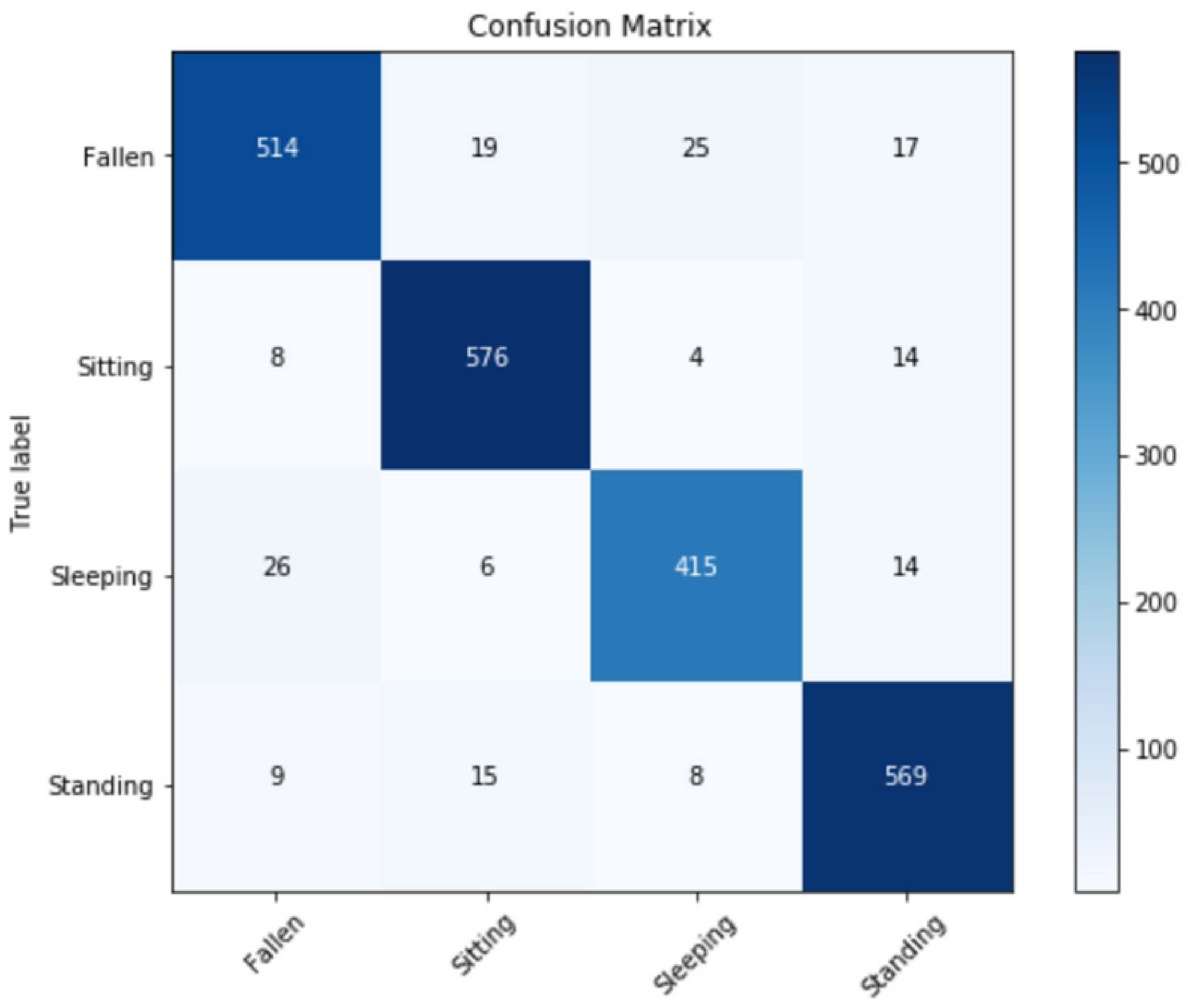
falldataset

Accuracy results obtained

Assigning class weights: 92.85%



Oversampling & undersampling: 92.68%



Datasets

MULTIMEDIA & HYBRID

falldataset

VISION-BASED

DMLSmartActions

VISION-BASED

UP FALL

HYBRID

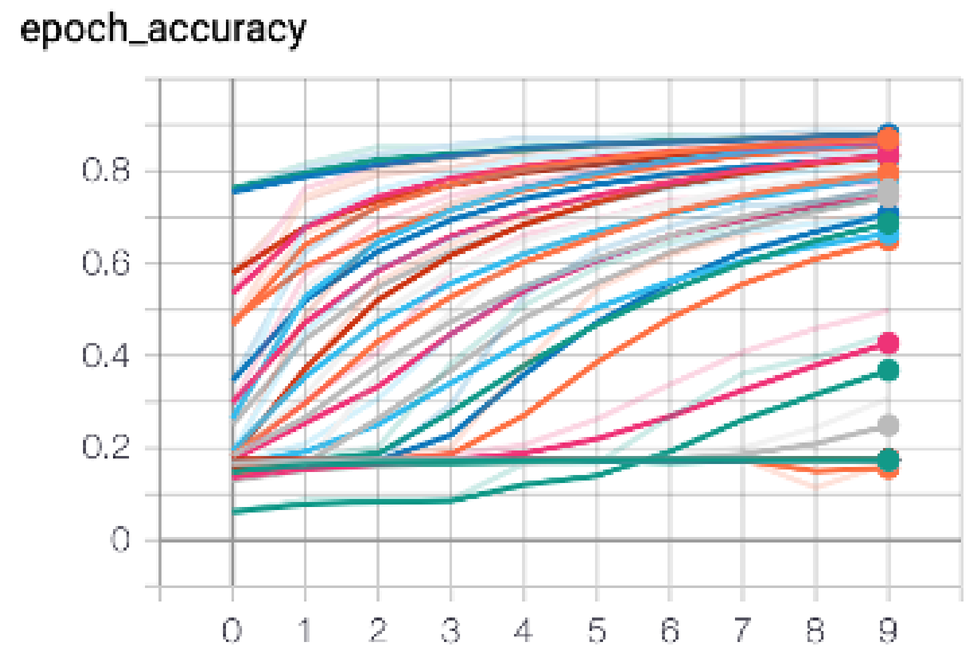
DMLSmartActions

Twelve classes:

- falling down
- drinking
- dropping and picking up something on the floor
- picking up something
- putting something
- cleaning the table
- reading
- sitting down
- standing up
- using a cellphone
- walking
- writing

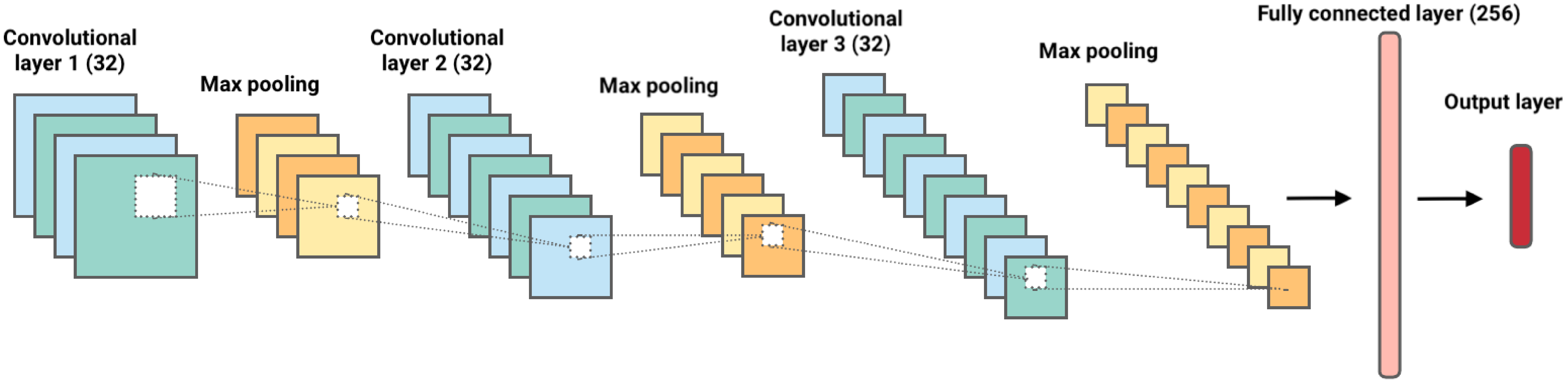
DMLSmartActions

To find an optimal architecture for the CNN model, we have implemented hyperparameter optimization with grid search and visualized performance metrics of our models.



Name	Smoothed	Value	Step	Time	Relative
2-conv-0.2-dropout-1-dense-1616311594/train	0.8699	0.88	9	Sun Mar 21, 13:38:00	10m 10s
2-conv-0.2-dropout-1-dense-1616311594/validation	0.8793	0.8848	9	Sun Mar 21, 13:38:00	10m 10s
2-conv-0.2-dropout-2-dense-1616317521/train	0.8644	0.8724	9	Sun Mar 21, 15:16:07	9m 39s
2-conv-0.2-dropout-2-dense-1616317521/validation	0.877	0.882	9	Sun Mar 21, 15:16:07	9m 39s
2-conv-0.5-dropout-1-dense-1616313610/train	0.7603	0.7994	9	Sun Mar 21, 14:10:53	9m 37s
2-conv-0.5-dropout-1-dense-1616313610/validation	0.7962	0.8282	9	Sun Mar 21, 14:10:53	9m 37s
2-conv-0.5-dropout-2-dense-1616319484/train	0.7864	0.8186	9	Sun Mar 21, 15:48:49	9m 39s
2-conv-0.5-dropout-2-dense-1616319484/validation	0.8349	0.8574	9	Sun Mar 21, 15:48:49	9m 39s
2-conv-0.8-dropout-1-dense-1616315562/train	0.1738	0.1738	9	Sun Mar 21, 14:43:26	9m 38s
2-conv-0.8-dropout-1-dense-1616315562/validation	0.1725	0.1749	9	Sun Mar 21, 14:43:26	9m 38s

CNN architecture - DMLSmartAction



Dealing with imbalanced dataset

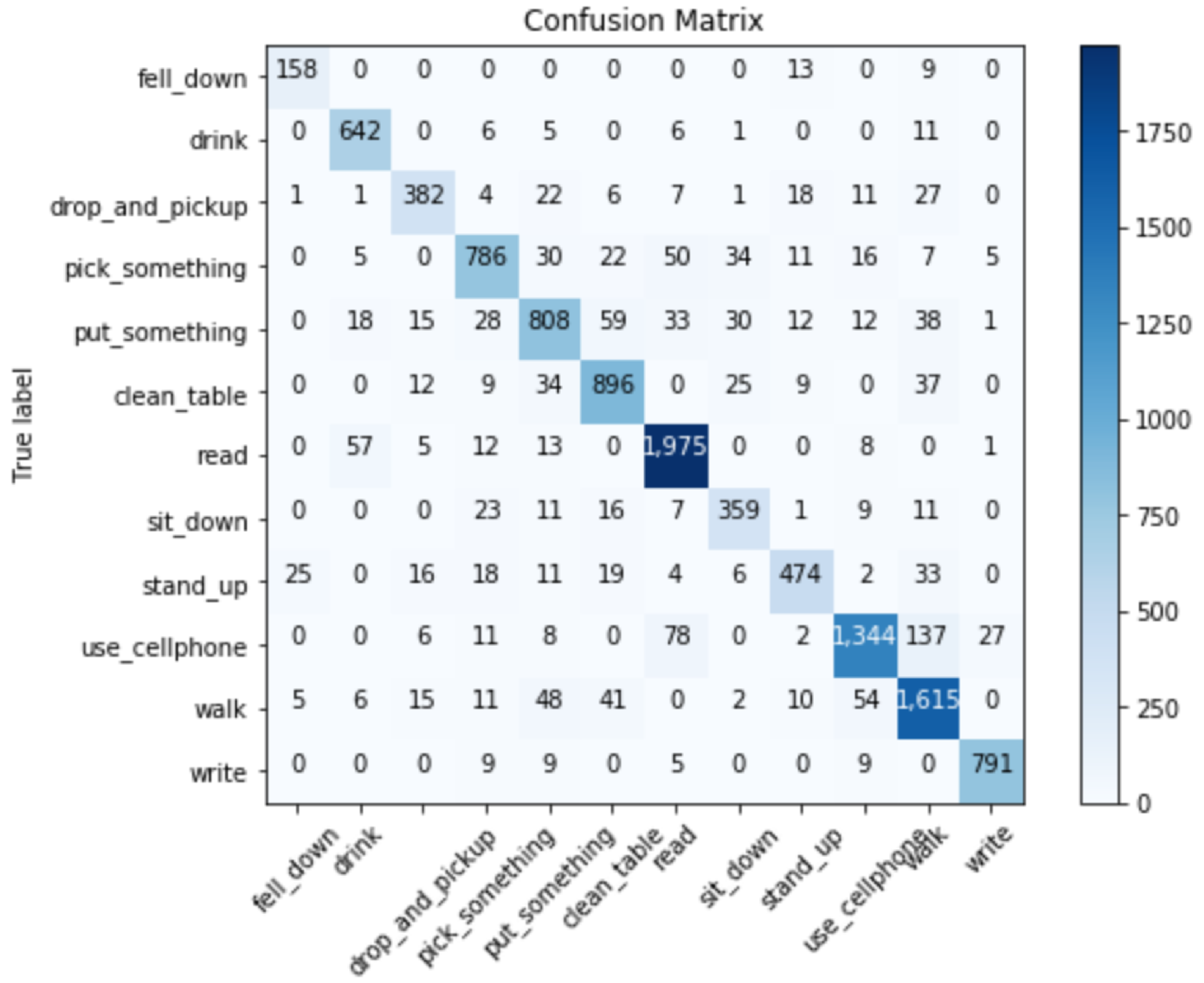
- Class weights were assigned based on the number of samples of each class:

$$w_j = n\_samples \,/\, (n\_classes * n\_samples_j)$$

DMLSmartActions

Accuracy results

Fold	Accuracy
1	86.22%
2	86.24%
3	87.29%
4	87.62%
5	87.02%
6	87.43%
7	87.02%
8	87.16%
9	87.20%
10	85.96%
avg	**86.97%**

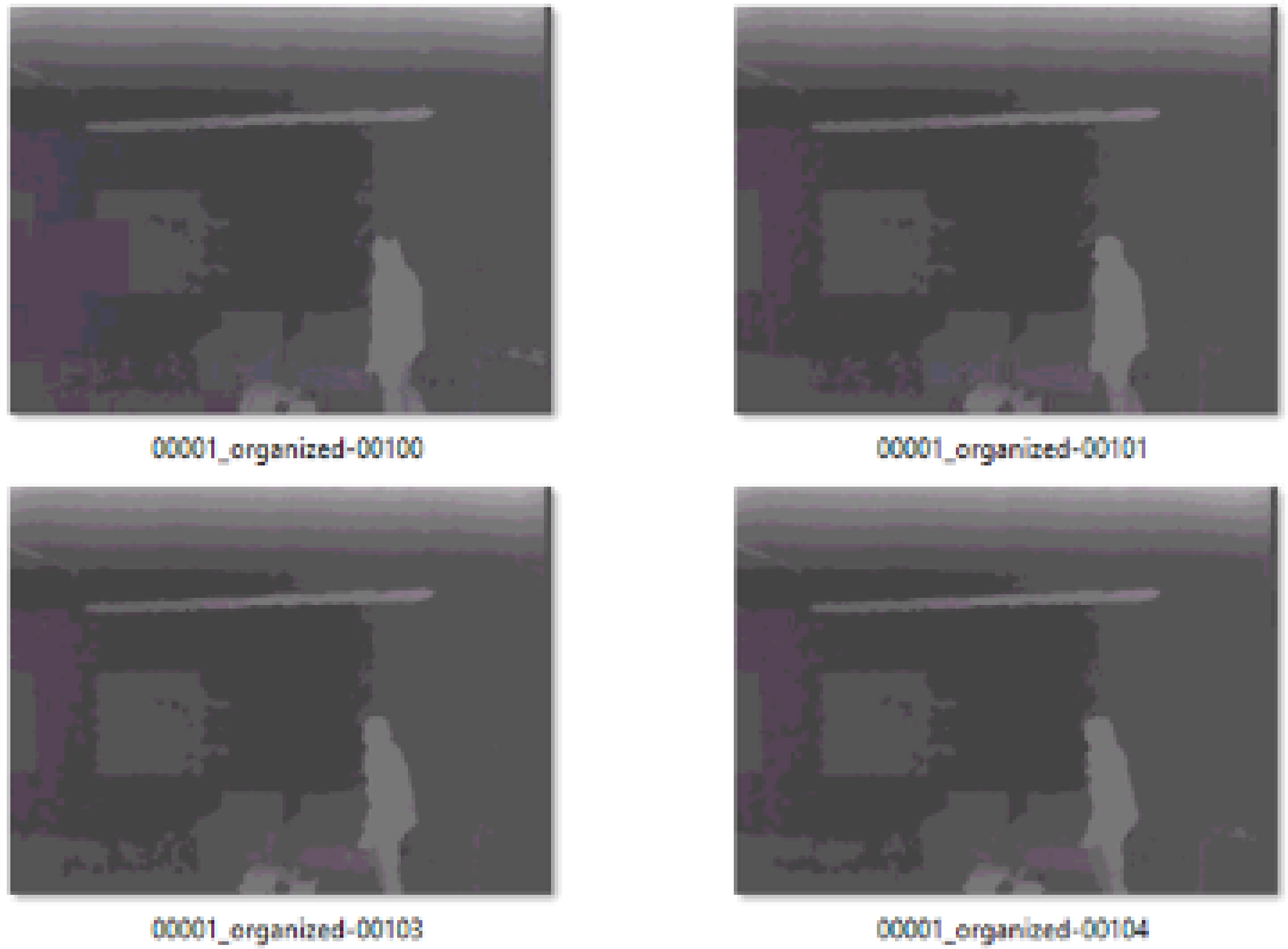


DMLSmartActions

Further work

1) Developing a script to extract a desired number of frames from any video.
2) Building a CNN model for DMLSmartActions depth videos.

As a result, the CNN model from depth videos obtained an accuracy of 83.14%.



Datasets

MULTIMEDIA & HYBRID

falldataset

VISION-BASED

DMLSmartActions

VISION-BASED

UP FALL

HYBRID

UP Fall

ONE OF FEW MULTIMODAL DATASETS FOR HUMAN ACTIVITY RECOGNITION AND FALL DETECTION

01

MULTIMODALITY

consists of the data from wearable sensors (the 3-axis accelerometer, the 3-axis gyroscope and the ambient light value), six infrared sensors, EEG headset, and two cameras.

02

ACTIVITIES

the dataset includes six different ADLs as well as five different kinds of falls

UP Fall

VIDEO DATA CLASSIFICATION

01

IMPLEMENTATION

After frames were preprocessed, I have utilized the architecture of CNN model that I built previously for DMLSmartActions dataset to perform activity classification.

02

RESULT

As a results, CNN model have reached an accuracy of 98.90% on 5 folds.

Fold	Accuracy
1	99.22%
2	98.05%
3	99.11%
4	99.21%
5	98.90%
avg	**98.90%**

UP Fall

VIDEO DATA CLASSIFICATION

01 IMPLEMENTATION

Alternatively, I have implemented **transfer learning** with ResNet50 (trained on ImageNet), with one dence layer of 128 nodes added in the end.

02 RESULT

As a results, transfer learning model have reached an accuracy of 99.6%.

UP Fall

TRANSFORMERS

DATA AUGMENTATION

We implemented some data augmentation to prevent overfitting:

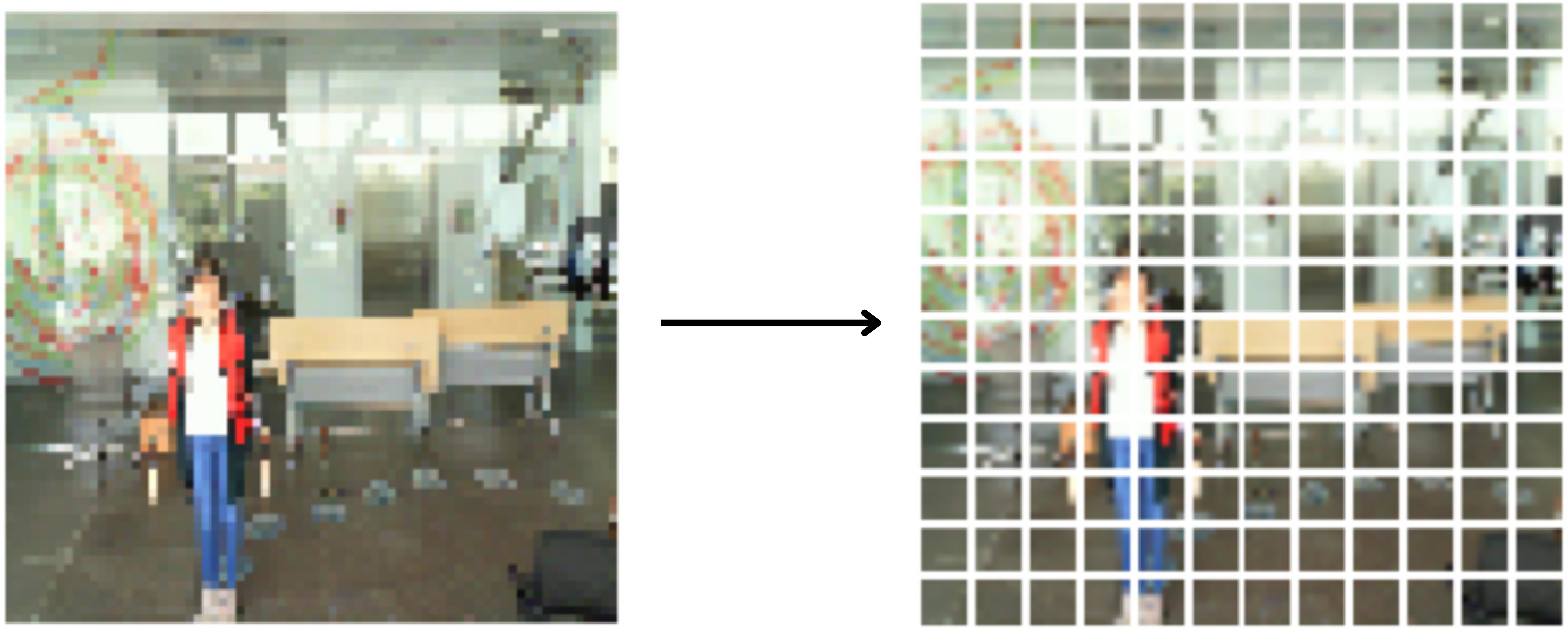
- random flip,
- random rotation,
- random zoom

UP Fall

TRANSFORMERS

CREATING PATCHES

Divided images into 144 patches of 12 x 12.



UP Fall

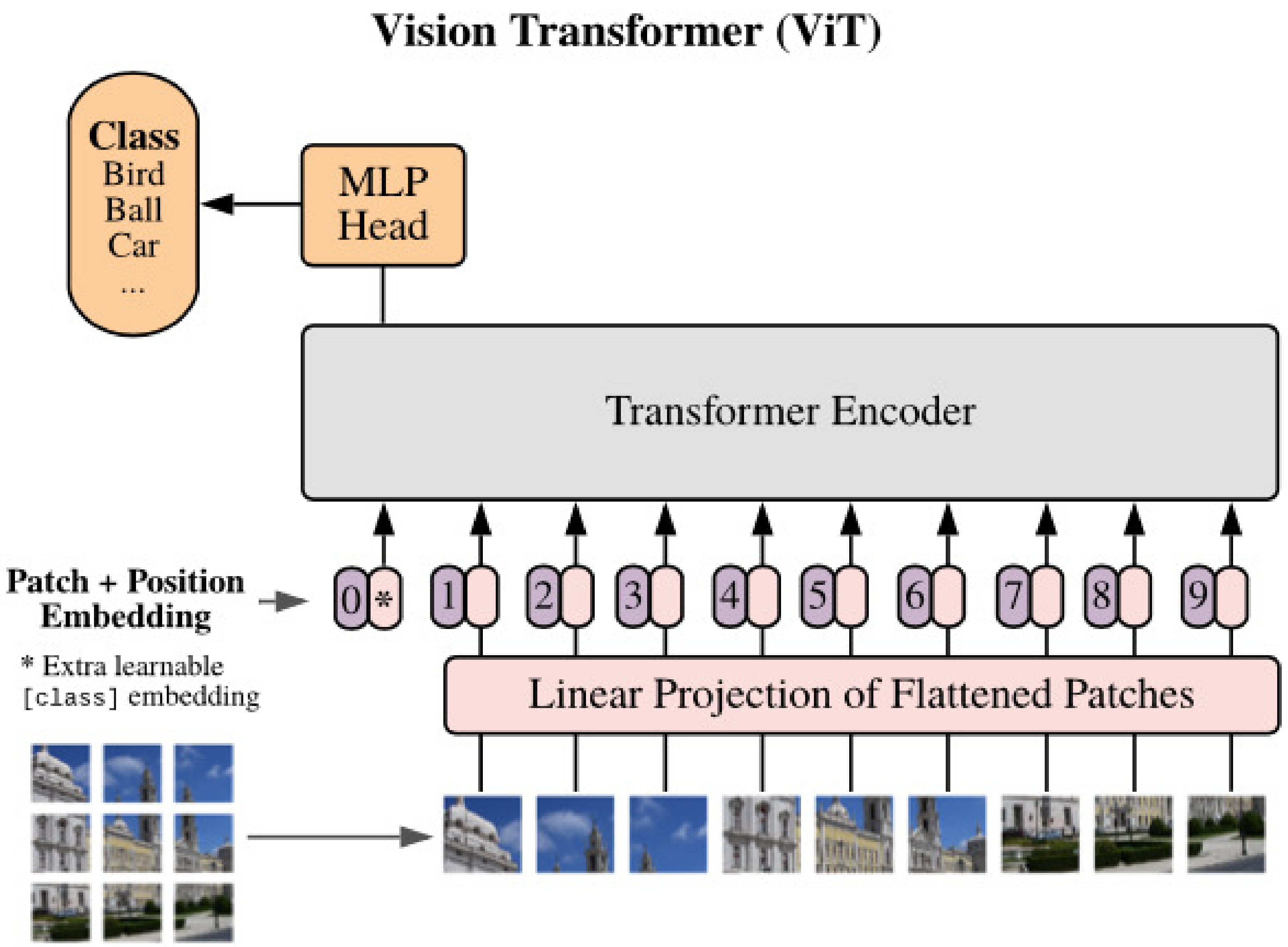
TRANSFORMERS

01 IMPLEMENTATION

Built a transformers model.

02 RESULT

As a results, transformer model have reached test accuracy of 99.87%.



Methods & Results

INERTIAL SENSOR BASED MODULE

HAR approaches

- **Inertial sensor based**
- **Vision-based**
- **Hybrid**

UP Fall

FEATURE EXTRACTION

01 IMPLEMENTATION

For each wearable, infrared sensor in the dataset, 12 temporal and 6 frequency features were extracted.

02 RESULT

Dataset with a total of 756 features.

Temporal:

- Mean
- Standard deviation
- Root mean squareMaximal amplitude
- Minimal amplitude
- Median
- Number of zero-crossing
- Skewness
- Kurtosis
- First quartile
- Third quartile
- Autocorrelatio

Frequency:

- Mean frequency
- Median frequency
- Entropy
- Energy
- Principal frequency
- Spectral centroid

UP Fall

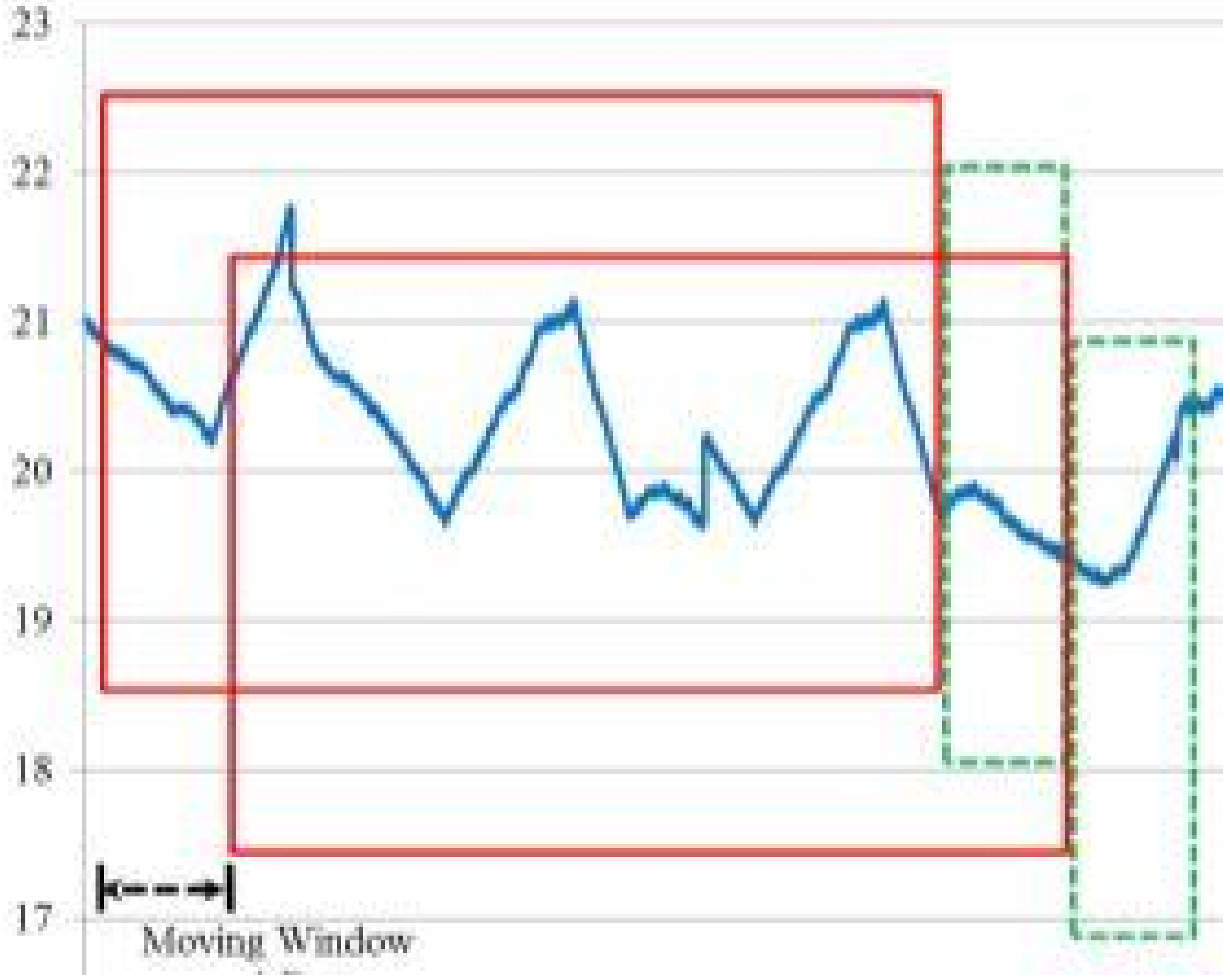
TIME WINDOW SELECTION

01 IMPLEMENTATION

I used three different feature datasets depending on the window size: (a) one-second, (b) two-second and (c) three-second. All the feature datasets consider 50% of overlapping.

02 RESULT

As a results, 1-s window length promotes the best performance on RF, SVM, KNN



UP Fall

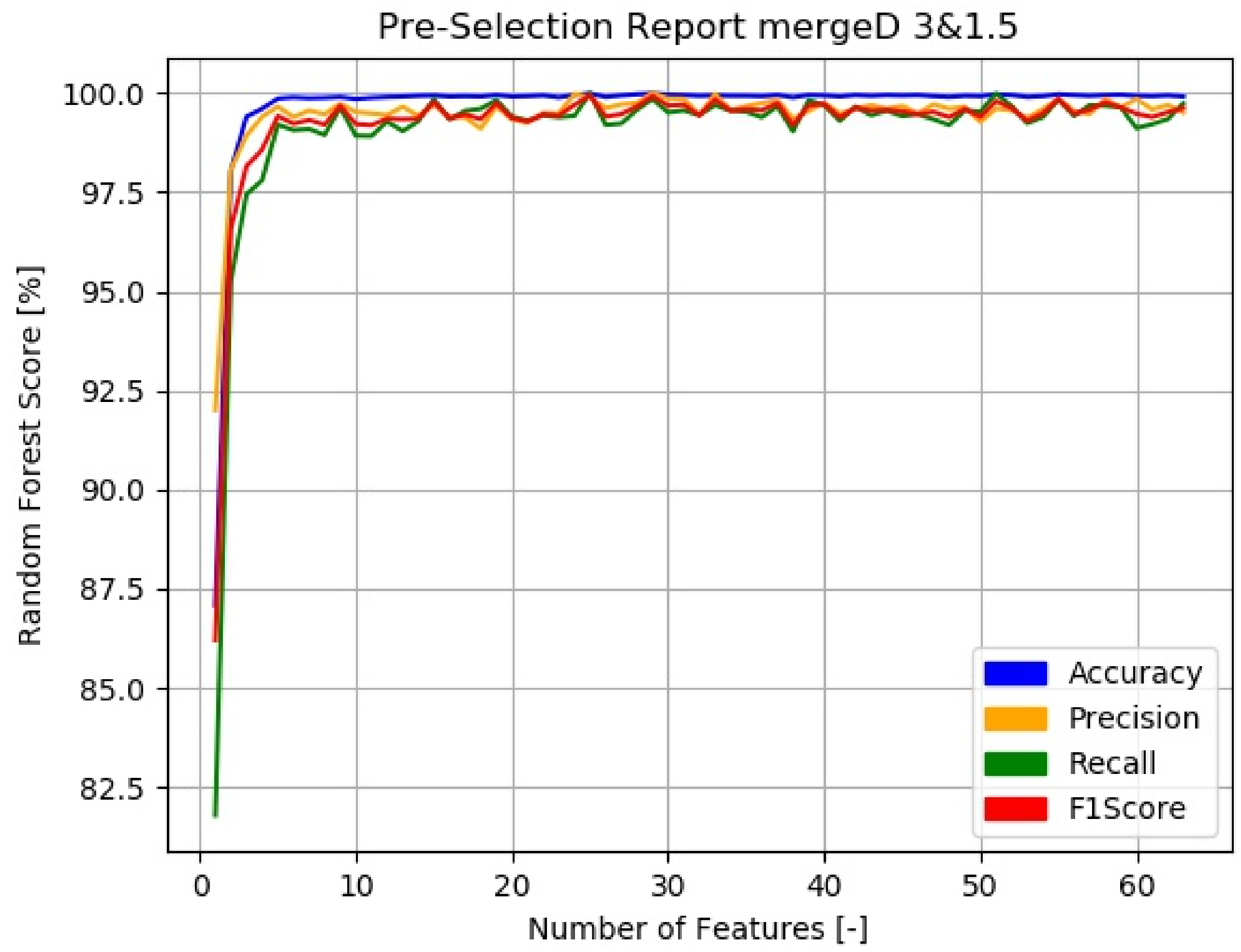
FEATURE SELECTION

01 FIRST PART

used correlation-based feature selection to reduce from 756 to 134 features.

02 SECOND PART

recursive feature elimination with Random Forest.
Final result: 63 features



UP Fall

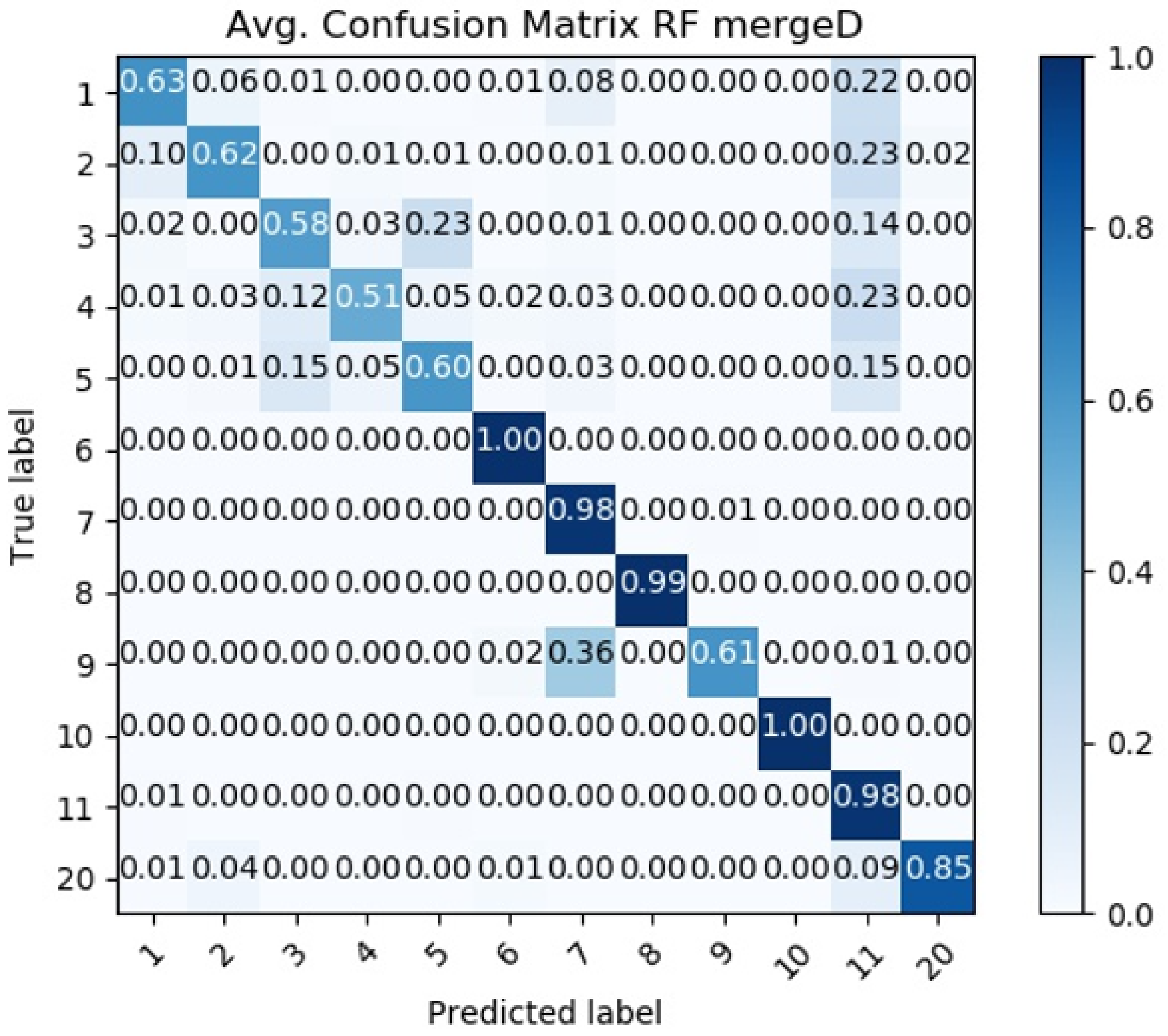
BASIC CLASSIFICATION MODELS

01 IMPLEMENTATION

The features selected by recursive feature elimination method were used for classification model training. The machine learning algorithm used for activity classification was Random Forest.

02 RESULT

The accuracy of the model reached 95.6% on 10-fold cross-validation



UP Fall

LONG SHORT-TERM MEMORY NETWORK

01 IMPLEMENTATION

1. Separated out last 10% of the data for testing
2. Preprocessed the data into sequences of 60
3. Built an LSTM model

02 RESULT

The accuracy of the model reached 92.73% on 10-fold cross-validation

Model architecture:
- LSTM with 128 nodes
- Dropout layer with rate of 50%
- LSTM with 128 nodes
- Dropout layer with rate of 50%
- LSTM with 128 nodes
- Dropout layer with rate of 50%
- Fully connected layer with 100 nodes

Methods & Results

MULTIMODAL ACTIVITY RECOGNITION

HAR approaches

Inertial sensor based

Vision-based

Hybrid

UP Fall

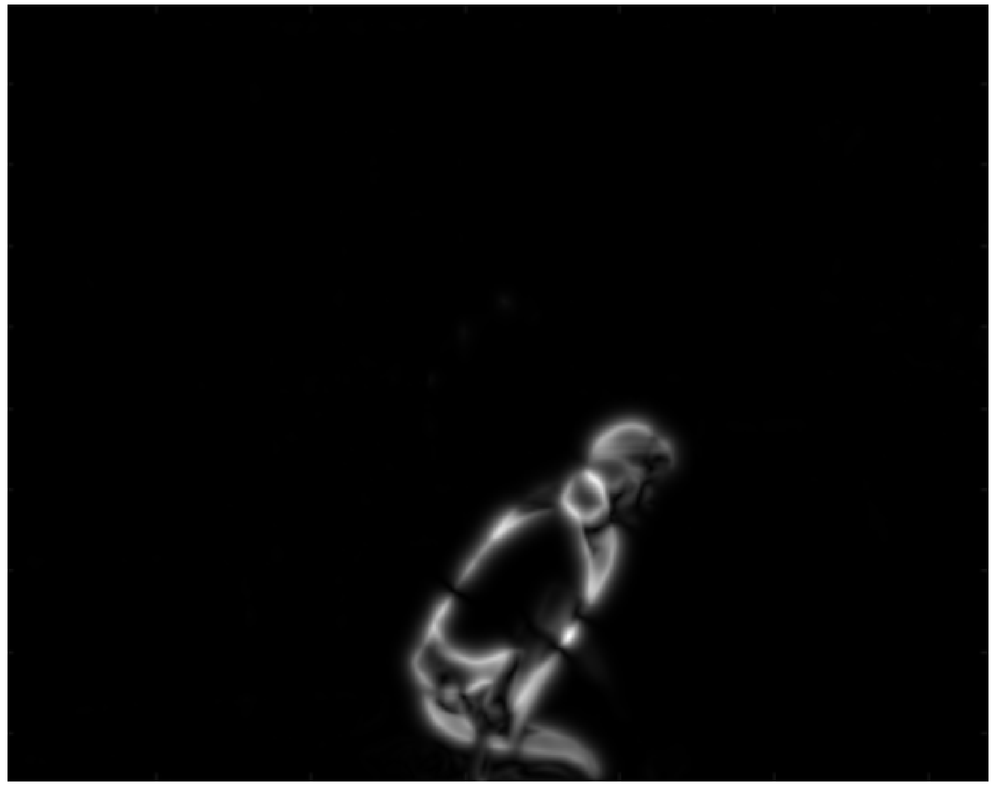
VIDEO PREPROCESSING METHOD

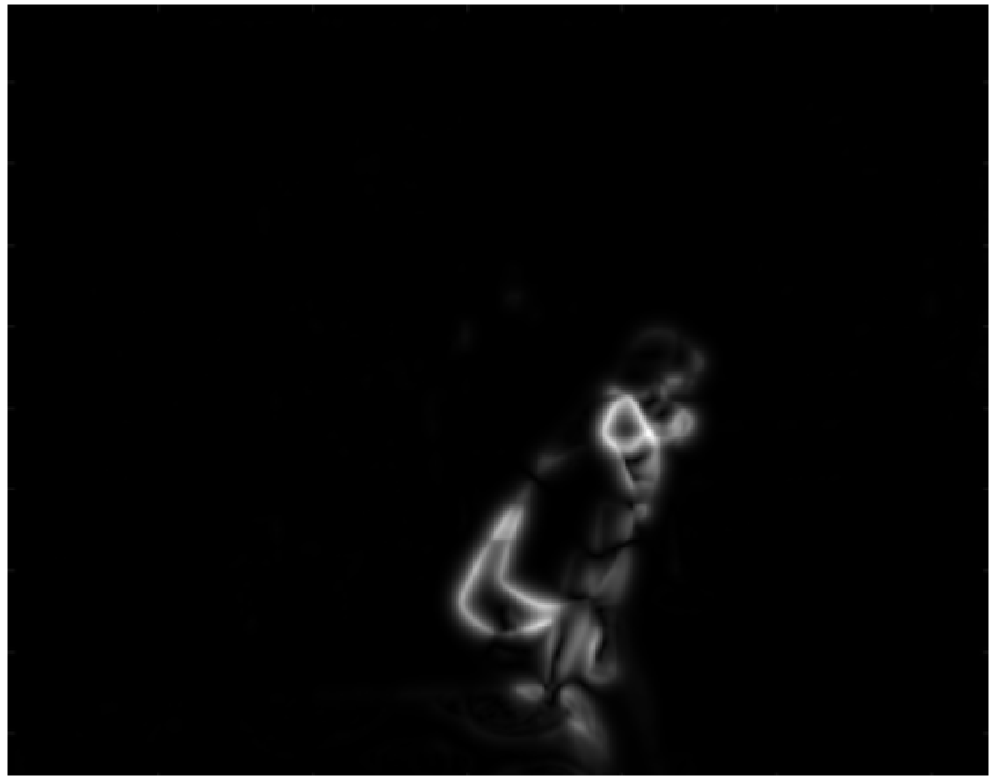
01 IMPLEMENTATION

Optical flow method is a methodology that allows calculating the apparent displacements of objects in an image sequence.

02 RESULT

Feature dataset with 800 features from the two cameras





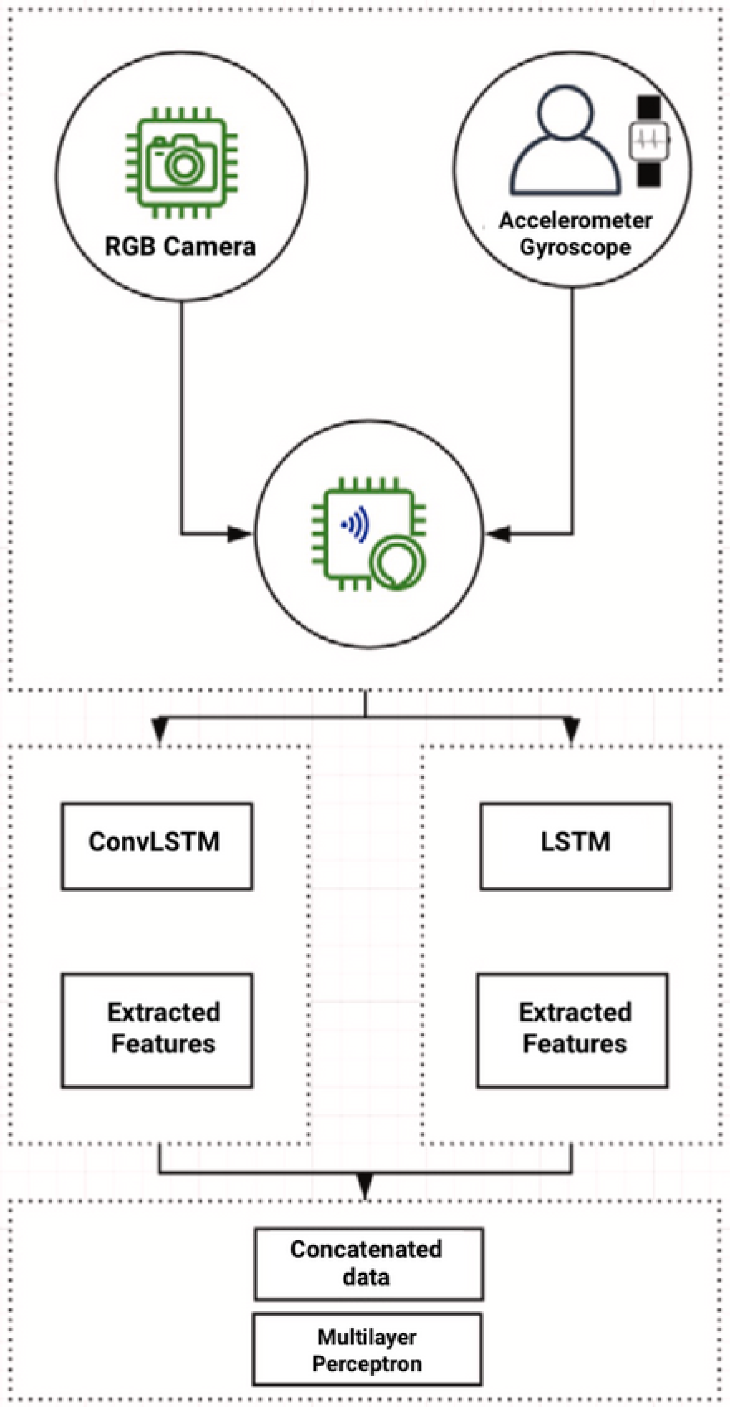
UP Fall

ConvLSTM network

01

IMPLEMENTATION

1. Separated out last 10% of the data for testing
2. Preprocessed the data into sequences of 60
3. Extracted features from a ConvLSTM model



02

RESULT

6512 sequences of features for training and 671 sequences for testing

UP Fall

FEATURE-LEVEL FUSION

01

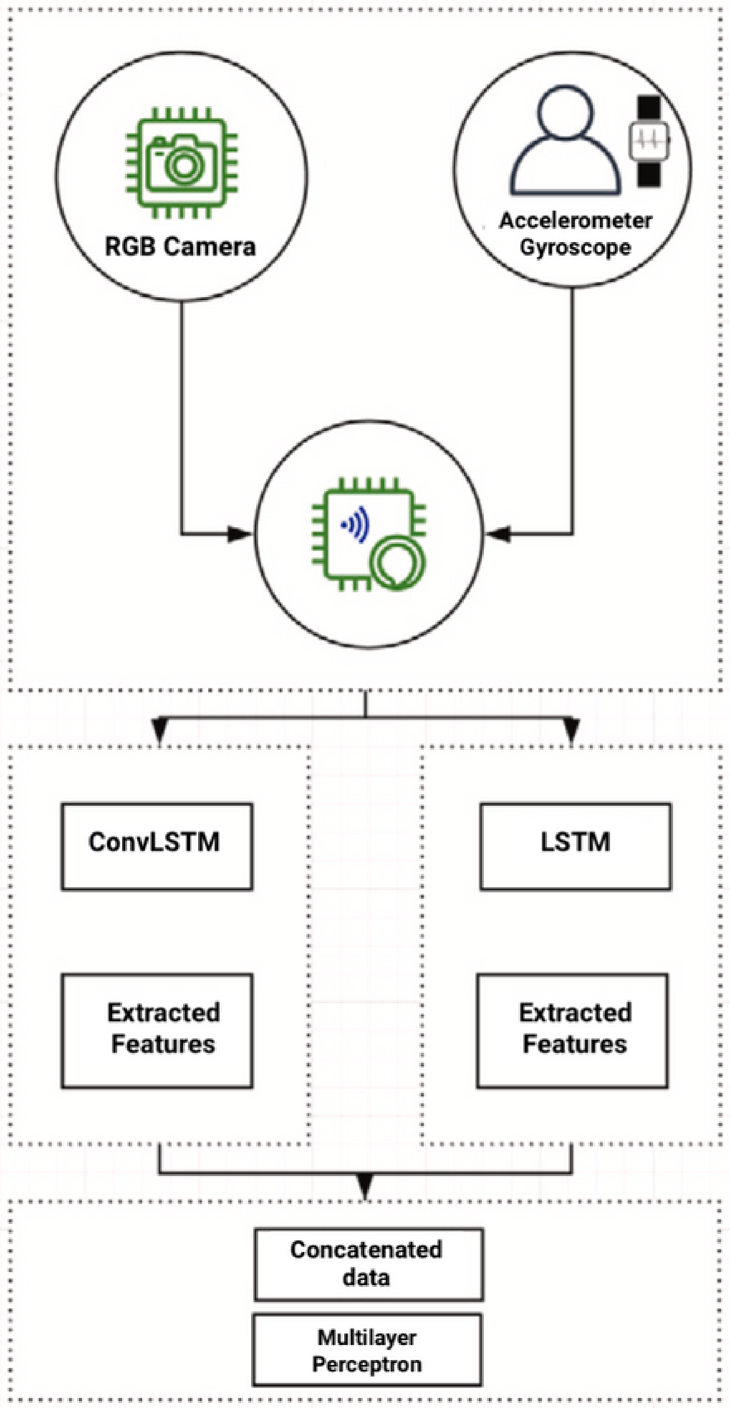
IMPLEMENTATION

After concatenating features obtained from LSTM and ConvLSTM, we defined a new model and trained a multilayer perceptron on the concatenated features.

02

RESULT

The feature-level fusion model obtained an accuracy of 85.84% and became the first multimodal model for fall classification.



UP Fall

COMPARISON WITH EXISTING STUDIES

Ref	Modality	Classification Type	Model	Accuracy
[1]	vision-based	binary	CNN	95.64%
[2]	inertial	binary	LSTM	93.17%
[3]	inertial	binary	SVM, LR, DT, RF, KNN, NB	96%-99%
proposed	vision-based	multi class	CNN	98.55%
proposed	vision-based	multi class	Transformer	99.87%
proposed	vision-based	multi class	Transfer learning, ResNet50	99.7%
proposed	multimodal	multi class	LSTM + ConvLSTM	85.84%

Conclusion

PART 04

- Falls is a crucial problem for elderly people. Early detection of falls may prevent or attenuate possible negative consequences for elderly people.
- While some of scientific articles focus on fall detection systems based on scalar body sensors, others apply vision based detection.
- We performed a fusion of inertial sensor based and vision-based modules for activity recognition and fall detection.

Thank You
for attention!

References

[1] Mobile activity recognition and fall detection system for elderly people using ameva algorithm. Pervasive Mob. Comput., 34(C):3–13, January 2017.
[2] Kripesh Adhikari, Hamid Bouchachia, and Hammadi Nait-Charif. Activity recognition for indoor fall detection using convolutional neural network. In Fifteenth IAPR International Conference on Machine Vision Applications, MVA 2017, Nagoya, Japan, May 8-12, 2017, pages 81–84. IEEE, 2017.
[3] Mohamed Ilyes Amara, Abderrahmane Akkouche, Elhocine Boutellaa, and Hakim Tayakout. A smartphone application for fall detection using accelerometer and convlstm network. In 2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH), pages 92–96, 2021.
[4] S. Mohsen Amiri, Mahsa Pourazad, and Panos Nasiopoulos. Improved human action recognition in a smart home environment setting. IRBM, 35, 11 2014.
[5] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos, and Victor C. M. Leung. A similarity measure for analyzing human activities using human-object interaction context. In 2014 IEEE International Conference on Image Process- ing, ICIP 2014, Paris, France, October 27-30, 2014, pages 2368–2372. IEEE, 2014.
[6] S. Mohsen Amiri, Mahsa T. Pourazad, Panos Nasiopoulos, and Victor C.M. Leung. Non-intrusive human activity monitoring in a smart home environment. In 2013 IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom 2013), pages 606–610, 2013.
[7] Morris Antonello, Marco Carraro, Marco Pierobon, and Emanuele Menegatti. Fast and robust detection of fallen people from a mobile robot. In Intelli- gent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on. IEEE, 2017.
[8] E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau. Multiple cameras fall dataset. In Technical report 1350, DIRO - Université de Montréale, 2010.
[9] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. Vision-based human activity recognition: a survey. Multim. Tools Appl., 79(41-42):30509–30555, 2020. 47
[10] Damien Bouchabou, Sao Mai Nguyen, Christophe Lohr, Benoit Leduc, and Ioan- nis Kanellos. A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning. CoRR, abs/2111.04418, 2021.
[11] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification. Journal of Electronic Imaging, 22(4):1 – 18, 2013.
[12] María de Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. Up-fall detection dataset: A multimodal approach. Sensors, 19(9):1988, 2019.

References

[13] Pedro Dias, Miguel Cardoso, Federico Guede Fernández, Ana Martins, and Ana Londral. Remote patient monitoring systems based on conversational agents for health data collection. In Nathalie Bier, Ana L. N. Fred, and Hugo Gamboa, editors, Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2022, Volume 5: HEALTH- INF, Online Streaming, February 9-11, 2022, pages 812–820. SCITEPRESS, 2022.
[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi- aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.
[15] Ricardo Espinosa, Hiram Ponce, Sebastián Gutiérrez, Lourdes Martínez- Villaseñor, Jorge Brieva, and Ernesto Moya-Albor. A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the up-fall detection dataset. Computers in Biology and Medicine, 115:103520, 2019.
[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9:1735–80, 12 1997.
[17] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. Artificial Intelligence, 17(1):185–203, 1981.
[18] Im Jung. A review of privacy-preserving human and human activity recognition. International Journal on Smart Sensing and Intelligent Systems, 13:1–13, 01 2020.
[19] Pekka Kannus, Harri Sievänen, Mika Palvanen, Teppo Järvinen, and Jari Parkkari. Prevention of falls and consequent injuries in elderly people. The Lancet, 366(9500):1885–1893, 2005. 48
[20] Bogdan Kwolek and Michal Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. Computer Methods and Programs in Biomedicine, 117:489–501, 10 2014.
[21] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recog- nition with gradient-based learning. In David A. Forsyth, Joseph L. Mundy, Vito Di Gesù, and Roberto Cipolla, editors, Shape, Contour and Grouping in Computer Vision, volume 1681 of Lecture Notes in Computer Science, page 319. Springer, 1999.
[22] Deok-Won Lee, Kooksung Jun, Khawar Naheem, and Mun Sang Kim. Deepneural network–based double-check method for fall detection using imu-l sensor and rgb camera data. IEEE Access, 9:48064–48079, 2021.
[23] Haobo Li, Aman Shrestha, Hadi Heidari, Julien Le Kernec, and Francesco Fio- ranelli. Bi-lstm network for multimodal continuous human activity recognition and fall detection. IEEE Sensors Journal, 20(3):1191–1201, 2020.
[24] Xiaojie Lv, Zongliang Gao, Changshun Yuan, Meng Li, and Chao Chen. Hybrid real-time fall detection system based on deep learning and multi-sensor fusion. In 2020 6th International Conference on Big Data and Information Analytics (BigDIA), pages 386–391, 2020

References

[25] Lourdes Martinez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Nuñez Martinez, and Carlos Peñafort-Asturiano. Up-fall detection dataset: A multimodal approach. Sensors, 19:1988, 04 2019.
[26] Homay Danaei Mehr and Huseyin Polat. Human activity recognition in smart home with deep learning approach. In 2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG), pages 149—153, 2019.
[27] Md Jaber Nahian, Tapotosh Ghosh, Mohammed Uddin, Md. Maynul Islam, Mufti Mahmud, and M. Shamim Kaiser. Towards artificial intelligence driven emotion aware fall monitoring framework suitable for elderly people with neuro- logical disorder. pages 275—286, 09 2020.
[28] Md. Jaber Al Nahian, Tapotosh Ghosh, Md. Hasan Al Banna, Mohammed A. Aseeri, Mohammed Nasir Uddin, Muhammad Raisuddin Ahmed, Mufti Mah mud, and M. Shamim Kaiser. Towards an accelerometer-based elderly fall detection system using cross-disciplinary time series features. IEEE Access, 9:39413—39431, 2021.
[29] Vinh The Nguyen, Tommy Dang, and Fang Jin. Predict saturated thickness using tensorboard visualization. In Karsten Rink, Dirk Zeckzer, Roxana Bujack, and Stefan Jänicke, editors, 6th Workshop on Visualisation in Environmental Sciences, EnvirVis@EuroVis 2018, Brno, Czech Republic, June 4, 2018, pages 35—39. Eurographics Association, 2018.49
[30] Kenneth Norberg. Audio Visual Communication Review, 1(3):190—194, 1953.
[31] Seong-Hi Park. Tools for assessing fall risk in the elderly: a systematic review and meta-analysis. Aging clinical and experimental research, 30(1):1—16, 2018.
[32] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Pro- cessing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 802—810, 2015.
[33] S. Singh, S. A. Velastin, and H. Ragheb. Muhavi: A multicamera human ac- tion video dataset for the evaluation of action recognition methods. In 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, pages 48—55, 2010.
[34] Jen-Kai Tsai, Chen-Chien Hsu, Wei-Yen Wang, and Shao-Kang Huang. Deep learning-based real-time multiple-person action recognition system. Sensors, 20:4758, 08 2020.
[35] Yan Yan, Tianzheng Liao, Jinjin Zhao, Jiahong Wang, Liang Ma, Wei Lv, Jing Xiong, and Lei Wang. Deep transfer learning with graph neural network for sensor-based human activity recognition. CoRR, abs/2203.07910, 2022.