ASEAN Engineering Journal

Full Paper

DEEP STAIRS DETECTION BASED ON CONVOLUTIONAL NEURAL NETWORKS FOR VISUALLY IMPAIRED PERSONS

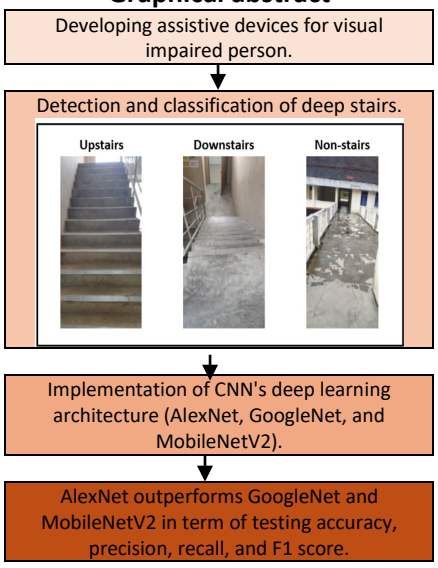
Nur Anis Jasmin Sufri[a*], Tie Heng En[a], Muhammad Amir As'ari[a,b], Muhammad Asraf Mansor[c]

[a]Department of Biomedical Engineering and Health Sciences, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia
[b]Sport Innovation and Technology Centre (SITC), Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia
[c]Micro-Nano System Engineering Research Group (MNSE), Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia

Article history
Received
04 September 2024
Received in revised form
24 December 2024
Accepted
24 January 2025
Published online
31 August 2025

*Corresponding author
najasmin2@graduate.utm.my

Graphical abstract



Abstract

In 2022, an alarming 2.2 billion people worldwide suffered from some form of vision impairment, with 237 million individuals experiencing moderate to severe visual impairment, putting them at a heightened risk of accidents and injuries. Clinical studies have shown that individuals with significant vision loss are at heightened risk of accidents, particularly during activities involving motion and orientation. Recent research on assistive devices has leveraged deep learning techniques to enhance safety, particularly by detecting stairs and preventing falls. This project aims to evaluate and compare the performance of well-known convolutional neural networks (CNNs) in detecting and classifying stairs, specifically MobileNetV2, GoogleNet, and AlexNet. A dataset of 3,000 RGB images, captured at a resolution of 2268 x 4032 pixels, was used, featuring images of stairs (upstairs, downstairs) and non-stair elements. The labeled dataset was augmented to match the input layer size of the pre-trained models and processed using MATLAB R2022b. Model performance was assessed by analyzing training and validation accuracy and loss through a training progress graph. Additionally, testing accuracy, precision, recall, and F1 score were evaluated using a confusion matrix. The results demonstrated that AlexNet outperformed GoogleNet and MobileNetV2, achieving an impressive 99% accuracy across all performance metrics.

Keywords: Visual impairment, Assistive technology, Convolutional neural networks, Stair fall risk, Deep learning

1.0 INTRODUCTION

Vision plays a significant role in human growth, development, and daily activities, as it is crucial for perceiving objects through the eyes. According to the Department of Ophthalmology at the University of Pittsburgh [1], visual impairment refers to an abnormal level of vision caused by a loss of visual field or damage to visual acuity, making it difficult for the eyes to perceive objects clearly. The World Health Organization [2] estimates that approximately 237 million people globally suffer from moderate or severe visual impairment, with the population projected to increase by 55% over the next 30 years.

Numerous studies have shown that visually impaired individuals are more prone to falling down stairs, leading to emergency room visits. The loss of vision impairs balance and walking ability [3]. Studies indicate that 59% [4] and 76% [5] of injury-related hospitalizations are linked to distance visual impairment or binocular visual acuity of 6/12 and 6/18, respectively.

As a result, assistive technology plays a vital role in helping people with vision loss navigate and move about. To assist blind individuals in walking independently, Sharma et al. developed a smart stick that integrates a wireless remote system via a radiofrequency module, a pothole detection and avoidance system using ultrasonic and moisture sensors, and an alarm system with a buzzer and vibration motor [6]. However, further

advancements are necessary, especially in the area of orientation and mobility aids, as stairs represent a major risk for individuals with visual impairments. Without adequate visual information, these individuals struggle to recognize the edges, depths, and potential obstructions on steps, increasing the likelihood of tripping, falling, or losing balance.

Recent studies on assistive devices for the visually impaired have applied deep learning techniques to detect staircases and prevent falls. Early detection of stairs can trigger an alert, such as a loud alarm or gentle vibration, to warn the person. A high-accuracy and high-precision staircase detection method, utilizing convolutional neural networks (CNNs), is essential to prevent staircase-related injuries for individuals with visual impairments.

After reviewing the literature, it was found that most existing studies focus on using RGB images as input rather than depth images or 3D point cloud data [7-8], likely due to the effectiveness of deep learning methods in staircase detection using RGB images [9-12]. Previous work can be categorized into semantic segmentation [9,12], object detection [8,10-11], and image classification [7] using deep learning techniques. Only one study introduced a new CNN model, PointNet, for staircase detection and classification based on depth images [7]. Furthermore, there is a lack of studies that explore the use of well-established CNN architectures—such as AlexNet, GoogleNet, and MobileNetV2—for stair detection using transfer learning.

Therefore, this project focuses on detecting and classifying staircases using the well-known CNN architectures, specifically AlexNet, GoogleNet, and MobileNetV2, to assist individuals with visual impairments. In addition,the study aims to demonstrate that well-established CNN models such as AlexNet, GoogleNet, and MobileNetV2 can effectively perform stair classification using an existing RGB camera, thereby eliminating the need to develop new models or invest in depth cameras.

2.0 METHODOLOGY

Figure 1 shows the top-level flow of the experiment setup for this project. The experiment begins with data acquisition and augmentation, which involves collecting and augmenting staircase and non-staircase images. This is followed by CNN model development, including the training and testing phases, and performance evaluation of each model.

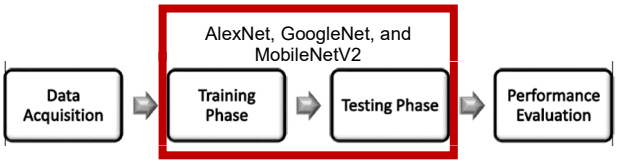


Figure 1 Main framework

The input dataset for this experiment was manually collected by capturing RGB images of staircases and non-staircases using a regular smartphone camera. The staircase images were taken at a height of 1.5 meters above ground and a distance of 1 meter from the stairs, while the non-staircase samples were captured randomly from different viewpoints and angles. To improve feature extraction during the training phase of the three CNN models, the absence of clutter and occlusion from people or objects was ensured. Several types of staircase designs, such as U-shaped stairs, double L-shaped stairs, straight stairs, stairs with intermediate landings, and spiral stairs, were also collected. Ultimately, the dataset consisted of three classes: upstairs, downstairs, and non-stairs, with each class containing 1,000 samples (the number of samples was inspired by the study in [7]).

The images were recorded at a resolution of 2268 by 4032 pixels, which was too large and unsuitable for the input layer of the CNN models. This was because the input layers of the CNN models used in this experiment required sizes of 227 x 227 x 3 pixels for AlexNet, and 224 x 224 x 3 pixels for both GoogleNet and MobileNetV2. Therefore, data augmentation was performed to resize the images to fit the input size requirements for each pre-trained network.

Before the training phase, the input dataset was split into training and testing datasets in a ratio of 0.9:0.1. The training dataset was further divided into training and validation datasets in a ratio of 0.7:0.3 [13]. Table 1 shows the distribution of images into the training, validation, and testing datasets, while Figure 2 illustrates a sample of the images.

Table 1 Number of images for each class

Dataset	Training	Validation	Testing	Total
Upstairs	630	270	100	1000
Downstairs	630	270	100	1000
Non-stairs	630	270	100	1000
Total	1890	810	300	**3000**

Figure 2 Random sample of image input data

This study employed three different pre-trained CNN models, which were trained, validated, and tested: AlexNet, GoogleNet, and MobileNetV2. The experimental evaluations were carried out using MATLAB R2022b software on a laptop equipped with an AMD Ryzen 7 5800H with Radeon Graphics processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3070 graphics processing unit, running on a 64-bit Windows 11 operating system. AlexNet was the first large-scale CNN, featuring deep convolutional layers stacked on top of each other [14]. Specifically, the convolution layers are combined with a max-pooling layer, three fully connected layers, and two dropout layers [15]. Rectified linear units (ReLU) are applied as the activation function in all layers, while the SoftMax function is used in the output layer [15]. AlexNet's design, along with its success in the ImageNet competition, accelerated the advancement of deep learning in computer vision. Figure 3 shows the architecture of AlexNet.

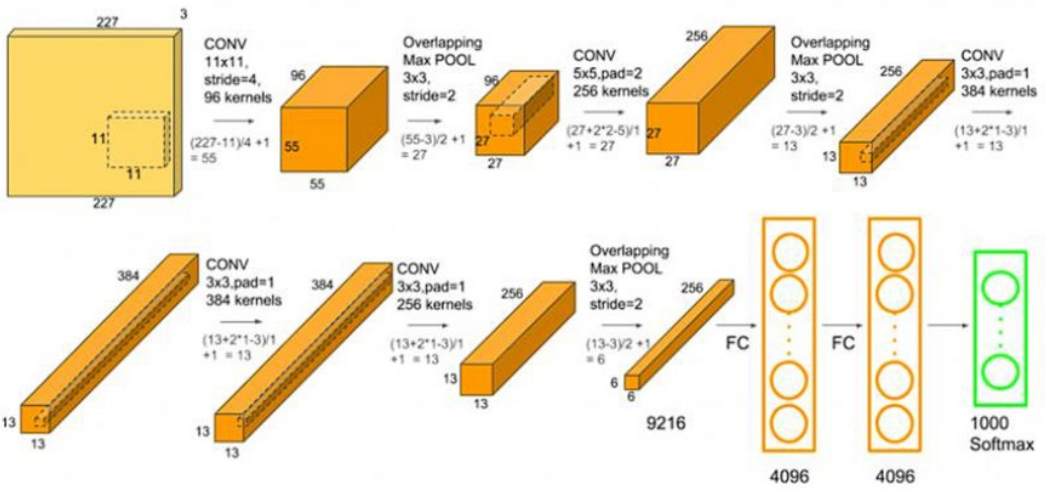


Figure 3 AlexNet structure

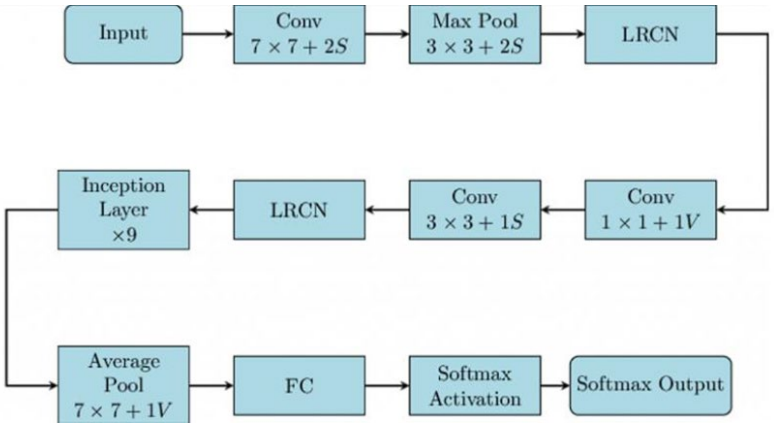


Figure 4 GoogleNet structure

GoogleNet was created to address the challenges of training deep neural networks with millions of parameters and achieving high accuracy [16]. The GoogLeNet architecture uses several distinct techniques, particularly 1x1 convolution and global average pooling, to significantly reduce the error rate compared to AlexNet [15]. Spatial redundancy during training is reduced by using un-pooling layers on top of CNNs, thus decreasing the number of learnable parameters [16]. Figure 4 shows the architecture of GoogleNet.

MobileNetV2 is a deep neural network architecture specifically designed for efficient mobile and embedded devices. It is an evolution of Google's original MobileNet design, which was introduced in 2017. MobileNetV2 aims to deliver excellent accuracy while minimizing computational and memory requirements for a range of computer vision tasks [15]. The MobileNetV2 architecture features two types of blocks: one with a stride of 1 and another with a stride of 2 for down-sampling purposes [15]. Both blocks consist of three layers: 1x1 convolution with ReLU activation, depth-wise convolution, and 1x1 convolution without any non-linearity [16]. Figure 5 shows the architecture of MobileNetV2.

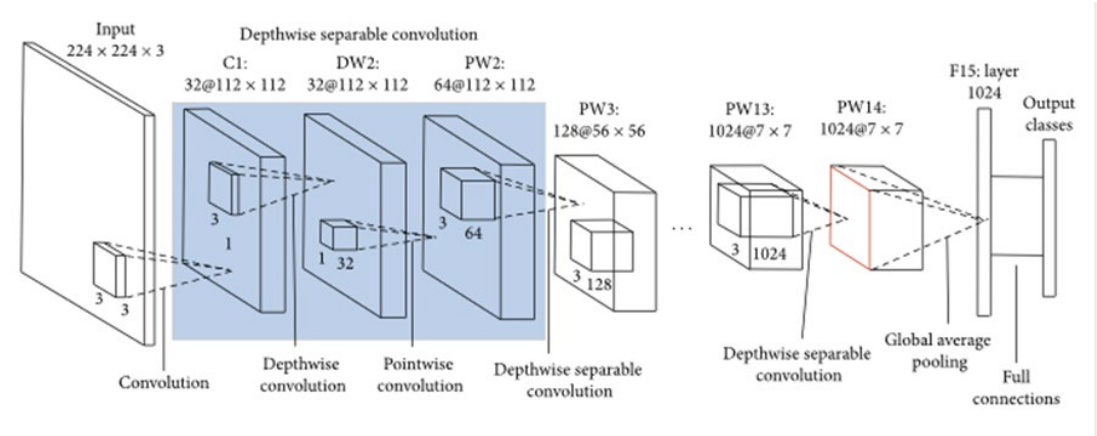


Figure 5 MobileNetV2 structure

The training hyperparameters, such as mini-batch size, maximum epochs, and learning rate, were set before training the network. After several trials, the optimized hyperparameters were chosen for training the three pre-trained models to achieve better performance. The hyperparameters are listed in Table 2.

The transfer learning technique was essential, as the pre-trained models were used as the foundation for learning a new classification task. In this project, staircases were detected and classified using three pre-trained CNN models: AlexNet, GoogleNet, and MobileNetV2. AlexNet's 25 layers include an image input layer, a convolution layer with ReLU activation, a normalization layer, a pooling layer, a dropout layer, a fully connected layer with a SoftMax function, and an output layer. Similarly, GoogleNet consists of 144 layers, including the input layer, convolution layer with ReLU activation, pooling layer, normalization layer, depth concatenation, fully connected layer with a SoftMax function, and output layer. The last pre-trained model used, MobileNetV2, involves 154 layers, including the input layer, convolution layer with clipped ReLU activation, batch normalization, element-wise addition, pooling layer, fully connected layer with a SoftMax function, and output layer.

Table 2 Training hyperparameters

Training hyperparameter	Mini batch size	Max epoch	Learning rate	Validation patience
Values	32	6	1 x 10^ (-4)	Infinity

As mentioned earlier, 300 images (10% of the total input dataset) were allocated for the testing dataset. The performance of the trained network was assessed using a confusion matrix, which consisted of actual and predicted results. The parameters obtained were True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). TP indicates that the model correctly predicted the actual positive class, while TN indicates that the model accurately predicted the actual negative class. Similarly, FP and FN refer to incorrect predictions.

These outcomes were crucial for evaluating the trained model's performance by analyzing its accuracy, precision, recall, and F1 score using the following equations (1) to (4). Accuracy represents the correct classification rate (TP + TN) over the total outcomes, while precision is the fraction of TP over the sum of TP and FP. Recall indicates the ability to correctly detect staircase events, and the F1 score is the harmonic mean of precision and recall, which compares the classification performance.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$sensitivity = \frac{TP}{TP + FN} \quad (3)$$

$$F1\,score = \frac{2 \times precision \times sensitivity}{precision + sensitivity} \quad (4)$$

3.0 RESULTS AND DISCUSSION

The aim of this study is to successfully develop CNN models capable of performing staircase detection and classification with high accuracy. The trained model's performance was assessed after the testing phase by plotting the confusion matrix and measuring the testing accuracy, as well as other parameters such as precision, sensitivity, and F1 score. The confusion matrix was plotted for each trained model, and the results were analyzed based on accuracy, precision, sensitivity, and F1 score.

Overall, AlexNet outperformed the other models with a testing accuracy of 99.0%, while GoogleNet and MobileNetV2 achieved accuracies of 98.3% and 96.0%, respectively. The precision values for AlexNet, GoogleNet, and MobileNetV2 were 99.0%, 98.4%, and 96.2%, respectively, while the recall values for AlexNet, GoogleNet, and MobileNetV2 were 99.0%, 98.3%, and 96.0%, respectively. By comparing the values achieved by the three CNNs, AlexNet had the highest recall and precision percentages among all the networks. This suggests that AlexNet had a low rate of false positives (FP), demonstrating high effectiveness in making positive predictions while minimizing incorrect positive predictions. Additionally, AlexNet had a low rate of false negatives (FN), indicating it was effective at identifying the positive class without missing too many positive instances. AlexNet also had the highest F1 score compared to GoogleNet and MobileNetV2, with an F1 score of 99.0%, while GoogleNet and MobileNetV2 achieved F1 scores of 98.4% and 96.1%, respectively. Thus, AlexNet's highest F1 score indicates that it could make accurate positive predictions while correctly identifying a significant portion of positive instances.

Table 3 Summary of the network performance

Model	AlexNet	GoogleNet	MobileNetV2
Precision	99.0 %	98.4 %	96.2 %
Recall	99.0 %	98.3 %	96.0 %
F1 score	99.0 %	98.4 %	96.1 %
Accuracy	99.0 %	98.3 %	96.0 %

Furthermore, AlexNet also showed outstanding testing accuracy with a significant value of 99.0% compared to the testing accuracy of GoogleNet and MobileNetV2 with values of 98.3% and 96.0% correspondingly. The overall performances of three different CNN models, AlexNet, GoogleNet, and MobileNetV2 were summarized in terms of training accuracy, validation accuracy, training loss, validation loss, precision, recall, F1 score, and testing accuracy in Table 3. It could be concluded that the staircase detection and classification system implemented using the AlexNet model had shown a reliable and satisfying result.

Figure 6 displays the confusion matrix for the AlexNet-trained model, which was used to predict the unseen testing dataset during the testing phase. The overall testing accuracy, shown in the bottom-right diagonal cell, was 99.0%. For the AlexNet model, all the upstairs images were classified accurately, with an accuracy of 99.0%, followed by the "other" class at 99.0% and the downstairs class at 98.0%. The recall and precision for the upstairs class were 100% and 99.0%, respectively, because all upstairs samples were classified accurately. However, one downstairs sample was misclassified

as an upstairs image, resulting in one false positive (FP) for the upstairs class.

For the downstairs class, the recall and precision were 98.0% and 99.0%, respectively. Out of 100 downstairs images, 98 were classified correctly, while 2 were misclassified as upstairs or other classes. Additionally, one sample labeled as "other" was misclassified as "downstairs," resulting in false negatives (FN) and false positives (FP) of 2 and 1, respectively. According to the confusion matrix shown in Figure 6, 99 out of 100 samples in the "other" class were classified correctly, but one sample was misclassified as "downstairs." Moreover, one downstairs sample was misclassified as "other," leading to both FP and FN having a value of 1. As a result, the recall and precision for the "other" class were both 99%.

The outcome of this project was benchmarked against related deep learning approaches in staircase detection and classification. The benchmarking results are presented in Table 4. As shown in Table 4, the performance of AlexNet surpassed that of the other trained CNN models.

Table 4 Benchmarking performance

Author & Year	Method	Type of approach	Findings
Research outcome	AlexNet	Image classification	Accuracy= 99.0% Precision= 99.0% Recall= 99.0% F1 score= 99.0%
	GoogleNet		Accuracy=98.3% Precision=98.4% Recall=98.3% F1 score=98.4%
	MobileNetV2		Accuracy=96.0% Precision=96.2% Recall=96.0% F1 score=96.1%
[7]	PointNet	Image classification	Accuracy= 97.3% Precision= 96.2% Recall= 100%
[8]	StairNetV22	Object detection	Accuracy= 91.99% Recall= 93.15%
[10]	ImageNet YOLOv3	Object detection	Precision= 83% Recall= 78%
[11]	MobileNet	Object detection	Accuracy=93.81%
[9]	AlbuNet	Semantic segmentation	Accuracy= 97%
[12]	U-Net YOLOv5	Semantic segmentation	Precision= 79% Recall= 80.3%

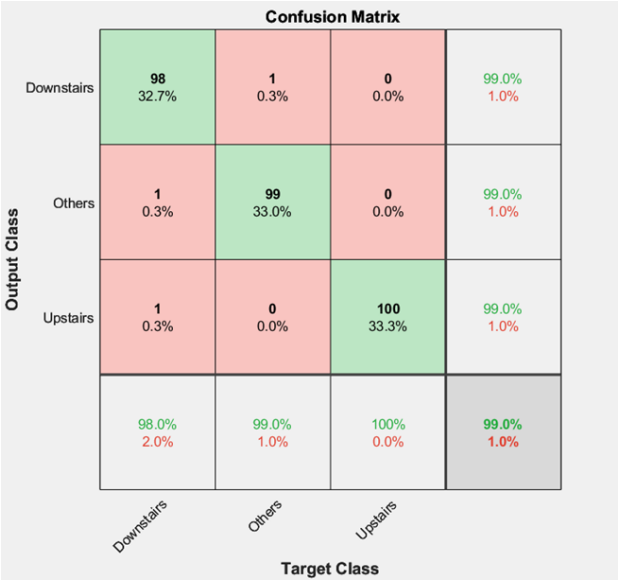


Figure 6 Confusion matrix of AlexNet

4.0 CONCLUSION

In summary, the main objectives of this project were to assess the performance of different pre-trained CNN models in detecting and classifying staircases, and to develop an automated staircase detection and classification system to assist the visually impaired in their daily lives. The reason CNN architecture was employed in this project is that CNNs are particularly well-suited for image classification tasks and outperform conventional machine learning methods. Transfer learning was used to train networks such as AlexNet, GoogleNet, and MobileNetV2 by feeding them an augmented dataset consisting of 3,000 images.

The results of this project demonstrate that the deep learning approach achieved higher performance compared to previous works in staircase detection and classification, with AlexNet delivering the best performance among the pre-trained networks used. The performance of AlexNet surpassed that of the other two trained networks, GoogleNet and MobileNetV2, with a testing accuracy of 99.0%, compared to 98.3% for GoogleNet and 96.0% for MobileNetV2.

Furthermore, the performance of each neural network was analyzed in terms of precision, recall, and F1 score using the confusion matrix. In comparison, AlexNet also scored the highest among the CNN models used in this project, with recall, precision, and F1 scores all reaching 99%. In conclusion, AlexNet was the most reliable model, delivering trustworthy results. The objectives of this project were successfully achieved, with staircases accurately detected and classified using the pre-trained CNN models and their performance evaluated.

Acknowledgement

The authors would like to express their gratitude to Universiti Teknologi Malaysia (UTM) and Ministry of Higher Education Malaysia under Fundamental Research Grant Scheme (FRGS/1/2023/ICT02/UTM/02/1) for funding this research.

Conflicts of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper

References

[1] Department of Ophthalmology, The University of Pittsburgh. 2022. What is Vision Impairment? Retrieved November 20, 2022, from http://ophthalmology.pitt.edu/vision-impairment/what-vision-impairment

[2] World Health Organization. 2022. Vision Impairment and blindness. World Health Organization. Retrieved November 20, 2022, from https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

[3] Elliott, D. B., Foster, R. J., & Whitaker, D. 2015. Falls and stair negotiation in older people and their relationship with vision. In Analysis of lower limb movement to determine the effect of manipulating the appearance of stairs to improve safety: A linked series of laboratory-based, repeated measures studies 8th ed., 3. Essay. NIHR Journals Library

[4] Close J, Ellis M, Hooper R, Glucksman E, Jackson S, Swift C. 1999. Prevention of falls in the elderly trial (PROFET): a randomised controlled trial. *Lancet*. 353: 93–7. DOI: https://doi.org/10.1016/S0140-6736(98)06119-4.

[5] Jack CI, Smith T, Neoh C, Lye M, McGalliard JN. 1995. Prevalence of low vision in elderly patients admitted to an acute geriatric unit in Liverpool: elderly people who fall are more likely to have low vision. *Gerontology*. 41: 280–5. DOI: https://doi.org/ 10.1159/000213695.

[6] Sharma, H., Tripathi, M., Kumar, A., & Gaur, M. S. 2018. Embedded Assistive Stick for visually impaired persons. *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. DOI: https://doi.org/10.1109/icccnt.2018.8493707

[7] Matsumura, H., & Premachandra, C. 2022. Deep-learning-based stair detection using 3D point cloud data for preventing visually impairing walking accidents. *IEEE Access,* 10: 56249–56255. DOI: https://doi.org/10.1109/access.2022.3178154

[8] C. Wang, Z. Pei, S. Qiu, and Z. Tang, 2022. "RGB-D-based stair detection using deep learning for autonomous stair climbing," arXiv preprint arXiv:2212.01098,

[9] Panchi, N., Agrawal, K., Patil, U., Gujarathi, A., Jain, A., Namdeo, H., & Chiddarwar, S. S. 2019. Deep learning-based stair segmentation and behavioral cloning for autonomous stair climbing. *International Journal of Semantic Computing,* 13(04): 497–512. DOI: https://doi.org/10.1142/s1793351x1940021x

[10] Patil, U., Gujarathi, A., Kulkarni, A., Jain, A., Malke, L., Tekade, R., Paigwar, K., & Chaturvedi, P. 2019. Deep learning-based stair detection and statistical image filtering for autonomous stair climbing. *2019 Third IEEE International Conference on Robotic Computing* (IRC). DOI: https://doi.org/10.1109/irc.2019.00031

[11] Ramalingam, B., Elara Mohan, R., Balakrishnan, S., Elangovan, K., Félix Gómez, B., Pathmakumar, T., Devarassu, M., Mohan Rayaguru, M., & Baskar, C. 2021. Stetro-deep learning powered staircase cleaning and maintenance reconfigurable robot. *Sensors,* 21(18): 6279. DOI: https://doi.org/10.3390/s21186279

[12] N. Rekhawar, Y. Govindani, and N. Rao, 2022. "Deep learning-based detection, segmentation and vision-based pose estimation of staircase," in *2022 1st International Conference on the Paradigm Shifts in Communication, Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, 78-83. IEEE.

[13] Russell, S. J., & Norvig, P. 2021. *Artificial intelligence: A modern approach* (4th ed.). Pearson

[14] Albawi, S., Mohammed, T. A., & Al-Zawi, S. 2017. Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET).* DOI: https://doi.org/10.1109/icengtechnol.2017.8308186

[15] Kumar, A. 2022. Different types of CNN Architectures explained: Examples. Data Analytics. Retrieved December 11, 2022, from https://vitalflux.com/different-types-of-cnn-architectures-explained-examples/

[16] Yamashita, R., Nishio, M., Do, R. K., & Togashi, K. 2018. Convolutional Neural Networks: An overview and application in Radiology. *Insights into Imaging*, 9(4), 611–629. DOI: https://doi.org/10.1007/s13244-018-0639-9