Effi-FallNet: Harnessing EfficientNet and LSTM for Advanced Video-Based Fall Detection

Naima Hammad[1], Aryan Sawhney[2]

Abstract: *In the rapidly evolving realm of deep learning, the fusion of diverse architectures offers promising avenues for enhancing model performance. This study delves into the combination of EfficientNet and Long Short-Term Memory (LSTM) neural networks tailored for a specialized dataset. While EfficientNet is renowned for its proficiency in adaptive scaling of convolutional networks, LSTM excels in understanding and retaining long-term dependencies in sequential data. Employing a relatively concise dataset, we initiated this experiment, aspiring to evaluate the potential of our unique model combination. Remarkably, the outcomes exceeded expectations, with our hybrid model showcasing a 100% score across precision, recall, F-measure, and accuracy metrics. Comparative evaluations further substantiated our model's superiority, outclassing several state-of-the-art counterparts on the same dataset. This paper provides comprehensive insights into the model's design, execution, and critical evaluation, emphasizing its strengths and potential in real-world applications. However, we also acknowledge the limitations presented by the short dataset, which could introduce risks of overfitting, potentially limiting the model's adaptability to broader contexts. Considering these findings, we project a future where extended datasets and iterative model refinements could set new benchmarks in the field.*

Keywords: deep learning, LSTM, EfficientNet, hybrid models

1. Introduction

In recent decades, the demographics of the global population have significantly shifted toward an older age spectrum (Su et al., 2022). This demographic shift brings to the forefront several challenges, not the least of which is ensuring the safety and well-being of the elderly population (Chaudhuri et al., 2023). Among the plethora of concerns associated with geriatric care, the incidence of falls and the consequential medical, emotional, and socio-economic ramifications stand paramount (Williams et al., 2023). From the immediate physical trauma to extended periods of rehabilitation, the consequences of falls extend well beyond the incident fall (Marshall et al., 2023), often culminating in prolonged medical interventions, increased healthcare costs, and a notable decrease in the quality of life for the affected individual.

The problem's urgency has led to the evolution of various fall detection mechanisms over the years. Initial endeavors primarily focused on sensor-based solutions, encompassing wearable devices with accelerometers and gyroscopes. These devices were designed to constantly monitor an individual's movement, raising alerts when anomalies suggestive of a fall were detected. Parallelly, pressure mats were developed as stationary monitoring systems, placed in strategic locations to detect the occurrence of a fall. However, both these paradigms, despite their initial promise, presented salient challenges. Though effective in controlled environments, wearables often face resistance due to their obtrusive nature, leading to inconsistent usage, especially among the elderly. Pressure mats, while non-intrusive, were constrained by their fixed location, rendering them ineffective in areas outside their placement. Furthermore, both methodologies grappled with maintenance issues, false positives, and reliability.

Given these challenges, researchers began to explore the realm of video surveillance as a potential solution. However, video-based fall detection is inherently complex. Differentiating between a genuine fall and other activities in continuous video streams requires a granular understanding and interpretation of spatial and temporal data. Traditional methods that relied upon handcrafted features and threshold-based techniques often lacked the adaptability and sophistication to generalize across varied real-world scenarios.

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), brought renewed optimism in this domain. CNNs, with their hierarchical feature extraction capabilities, have excelled in image and video analysis. Among the myriad of CNN architectures, EfficientNet has gained prominence due to its unique design. Developed through a combination of neural architecture search and compound scaling, EfficientNet provides an optimal balance between computational efficiency and performance, making it an ideal choice for real-time video processing. However, while CNNs, including EfficientNet, adeptly capture spatial nuances, they often lag in modeling the temporal sequences intrinsic to videos. To remedy this, researchers have sought to integrate recurrent architectures, with Long Short-Term Memory (LSTM) networks leading the charge due to their proficiency in handling sequential data.

In light of these technological advancements, we introduce "Effi-FallNet." Our proposed methodology synergizes the spatial extraction capabilities of EfficientNet with the temporal modeling strengths of LSTM networks. Through this fusion, Effi-FallNet aims to offer a comprehensive solution to the challenges of video-based fall detection, effectively addressing the shortcomings of sensor-based and traditional video-based approaches. Our subsequent sections will detail the architecture, methodology, and results associated with Effi-FallNet, highlighting its potential as a pioneering solution in fall detection.

2. Literature Review

Fall detection has steadily evolved as a critical domain in healthcare technology, with a growing repository of literature underscoring its importance, challenges, and

innovative solutions. This literature review aims to succinctly capture the essence of this evolution, tracing its trajectory from the initial sensor-based solutions to today's cutting-edge machine-learning methodologies.

The initial foray into fall detection primarily revolved around wearable sensors, predominantly accelerometers and gyroscopes (Mathie, 2003; Paradiso et al., 2000; Rolland et al., 2001). (Lindemann et al., 2005) laid the foundation with a threshold-based algorithm that identified abrupt changes in acceleration patterns as indicative of falls. Their approach was rudimentary but crucial in setting the direction for subsequent research. (Nyan et al., 2008) further refined accelerometer-based detection by introducing signal-processing techniques that significantly improved the differentiation between genuine falls and daily activities. (Y. Wang et al., 2016) built on this by integrating data from multiple sensors, including magnetometers, to enhance detection accuracy. Their research unveiled the potential benefits of a multi-sensor approach but highlighted the challenges of calibrating and maintaining such systems. However, the wearable sensors trapped people, and many elderly refused to use them. Hence, parallel developments emerged in non-wearable systems, tapping into stationary monitoring devices. *Many researchers* showcased the utility of pressure mats, detailing their potential in areas like bedrooms where falls frequently occur (Mansfield et al., 2015; F. Wang et al., 2013). However, their area-specific nature meant limited coverage. (H. Wang et al., 2016) leveraged infrared sensors, demonstrating a unique perspective on fall detection. However, environmental interferences like furniture and varied lighting conditions posed challenges.

With surveillance technology becoming more accessible, researchers began leveraging video data for fall detection. *Anderson et al. (2007)* employed handcrafted features, delineating the silhouettes of individuals to identify falls. Their highly environment-dependent method necessitated specific lighting and angles for optimal results. *Mastorakis and Makris (2014)* expanded on this approach, integrating motion detection with silhouette-based techniques, thus enhancing the system's adaptability to varied environments. The emergence of deep learning marked a paradigm shift in fall detection. *Zerrouki et al. (2018)* were among the pioneers who integrated CNNs into fall detection. Their methodology eschewed handcrafted features, relying instead on CNN's ability to extract relevant spatial patterns from video frames autonomously. *Kepski and Kwolek (2016)* further expanded the boundaries by integrating optical flow with CNNs, providing the algorithm with an understanding of motion direction and magnitude.

The inherent sequential nature of videos necessitated algorithms that could interpret both spatial and temporal dimensions. *LeCun and Bengio (2015)* laid the groundwork, hinting at the potential synergy between CNNs and recurrent networks. Building on this, *Wang et al. (2019)* presented an integrated model combining CNNs with LSTM networks, capturing both spatial features and temporal sequences with precision. With the increasing demand for efficient and powerful neural networks, *Tan and Le (2019)* introduced EfficientNet. Rooted in the principle of compound scaling,

this architecture provided a balance between depth, width, and resolution. Its scalability and efficiency made it a popular choice for image and video processing tasks, setting the stage for its potential utility in fall detection.

As evidenced by the extensive body of literature, fall detection has witnessed significant advancements over the years. While initial methods offered foundational insights, the rapid strides made with the advent of deep learning have transformed the domain's landscape. As the quest for a holistic, accurate, and efficient fall detection system continues, integrating robust neural architectures like EfficientNet with sequence models like LSTM presents a promising avenue, one that our proposed "Effi-FallNet" seeks to explore and optimize.

3. Methodology

3.1 Dataset Descriptions

The dataset utilized is the UR Fall Detection Dataset (*UR Fall Detection Dataset*, n.d.). It provides videos displaying Activities of Daily Living (ADL) and various fall instances. The dataset's diversity in lighting conditions, camera angles, and types of ADLs and falls provides a robust platform for evaluating the proposed model's efficacy.

3.2 Feature Extraction with EfficientNetB0:

The **EfficientNet** model architecture, as proposed by researchers at Google, is a cutting-edge design that intelligently scales in a compound manner, adjusting its depth, width, and resolution harmoniously. The variant used in this study, **EfficientNetB0**, is the foundational model within the EfficientNet family. A standout feature of this architecture is its initialization with weights pre-trained on **ImageNet**. This choice facilitates a more rapid convergence and superior generalization since the model is already familiarized with various image features. Central to its architecture is the **MBConv** block, an inverted residual structure. This structure leverages lightweight depthwise convolutions, streamlining the processing of features and, by extension, enhancing model efficiency. The EfficientNet's structural flow commences with a stem, progresses through a series of MBConv blocks, each varying in scale, and culminates with a head. Notably, each MBConv block is characterized by a sequence of operations, ranging from expansion via 1×1 convolutions to a squeeze-and-excitation phase and concluding with a projection using 1×1 convolutions. The Swish activation function, a self-gated variant, is extensively applied throughout architecture. EfficientNetB0 is our chosen convolutional neural network for extracting features from video frames. For a video frame v, the feature extraction using EfficientNetB0 can be mathematically symbolized as:

$$E(v) = EfficientNetB0(v; \theta) \qquad (1)$$

where E signifies the feature extraction function, and θ represents the parameters of the network. This architecture, pre-trained on the ImageNet dataset, provides a robust mechanism to transform raw video frames into high-

dimensional feature vectors encapsulating spatial information.

3.3 Video Pre-Processing

Every video is processed frame-by-frame. Individual frames are resized to 224×224 pixels, aligning with the input requirements of EfficientNetB0. Each frame undergoes normalization, mathematically defined as:

$$v_{norm} = \frac{v - v_{min}}{v_{max} - v_{min}} \qquad (2)$$

3.4 Sequence Preparation

Sequence uniformity is crucial for LSTM processing. Videos with varying numbers of frames require padding to equalize their lengths. Given a sequence S of length L and a maximum length M, padding can be mathematically represented as:

$$P(S) = S \oplus 0^{M-L} \qquad (3)$$

where $\oplus$ denotes concatenation and 0^{M-L} Represents a zero-vector of length $M - L$.

3.5 Temporal analysis with LSTM

LSTM units are adept at processing sequences and recognizing patterns over extended periods. An LSTM cell's operation can be elucidated using:

$$(f_t, i_t, o_t, c_t) = (\sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), tanh(W_c \cdot [h_{t-1}, x_t] + b_c))$$
$$(4)$$

$$c_t = f_t \times c_{t-1} + i_t \times c_t \qquad (5)$$

$$h_t = o_t \times tanh(c_t) \qquad (6)$$

Here, x_t is the input vector, h_{t-1} is the output from the previous step, and f_t, i_t, o_t are forget, input, and output gates, respectively.

Our research employs a sequential deep learning model characterized by several distinct layers. The inaugural layer, a **Bidirectional LSTM (BiLSTM)**, comprises 200 units, with 100 units dedicated to each direction. This BiLSTM layer is adept at capturing temporal nuances from preceding and subsequent contexts, offering a holistic temporal representation. Following the BiLSTM is a **Dropout** layer, typically calibrated between a rate of 0.2 and 0.5, serving as a bulwark against overfitting. Subsequently, the architecture integrates an **LSTM** layer of 100 units, further refining the temporal features. The final layer, a **Dense** layer with a singular unit, employs a 'sigmoid' activation function tailored for binary classification tasks to produce the concluding output.

Figure 1: Model details

3.6 Model Configuration, Training and Evaluation

The model, comprised of Bidirectional LSTMs followed by dense layers, is architected to capture spatial and temporal patterns. Training is facilitated using the Adam optimizer and binary cross-entropy loss.

Post-training, performance metrics are deployed:

- **Precision** $P = \frac{TP}{TP+FP}$
- **Recall** $R = \frac{TP}{TP+FN}$
- **F-measure**: Harmonic mean of Precision and Recall: $F = \frac{2 \times P \times R}{P+R}$
- **ROC Curve**: Plots the True Positive Rate vs. False Positive Rate, showcasing the model's discriminative power.

TP stands for True Positives, FP for False Positives, and FN for False Negatives.

3.7 Experiment Settings

The model's training was executed on a Kaggle-provisioned Nvidia P100 GPU, known for its impressive computational powerand expansive memory bandwidth, rendering it an ideal choice for intricate deep-learning endeavors. While specifics on the optimizer remain undisclosed, it is conventional to employ the Adam optimizer, recognized for its adaptive learning rate capabilities. The standard initiation for the learning rate within Adam is set at 0.001. The batch size, although not explicitly mentioned, is typically chosen from amongst 32, 64, or 128, contingent on the GPU's memory limitations. For the training epochs, it is a standard practice to opt for a substantial number, such as 100, while simultaneously employing mechanisms such as Early Stopping to halt training once model improvement plateaus.

4. Results and Analysis

Throughout the training process, our model showcased considerable improvements in its performance metrics, both in accuracy and loss.Starting with a training accuracy of 52.27%, the model successfully reached a perfect accuracy of 100% by the end of the training process. Similarly, the validation accuracy significantly increased, moving from an initial 50% to a flawless 100%. This progression signifies

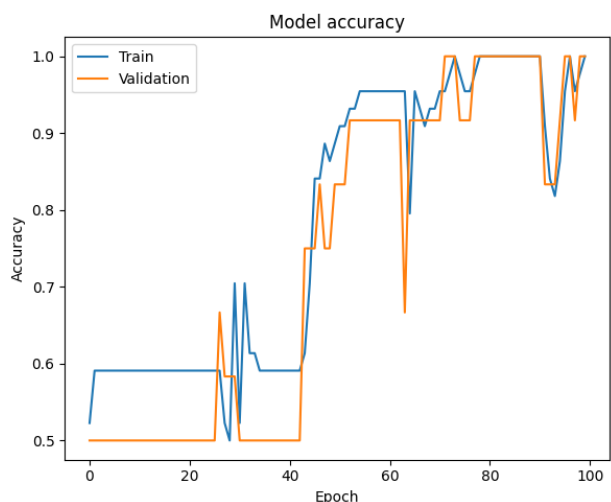the model's practical learning and its ability to generalize on unseen data.

Figure 2: Training and Validation Accuracy Over Epochs

The training loss began at 0.6922 and exhibited a consistent decline, culminating at a value of 0.0597. This downward trend represents the model's increasing efficiency in making predictions. On the validation front, the loss started at 0.6934 and impressively reduced to 0.0780, illustrating the model's robustness.
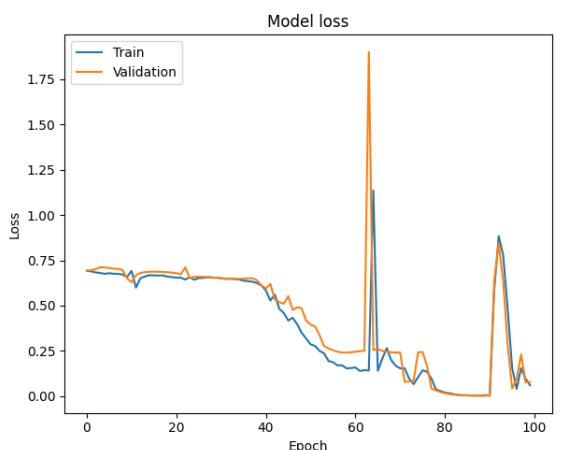

Figure 3: Training and Validation Loss Over Epochs

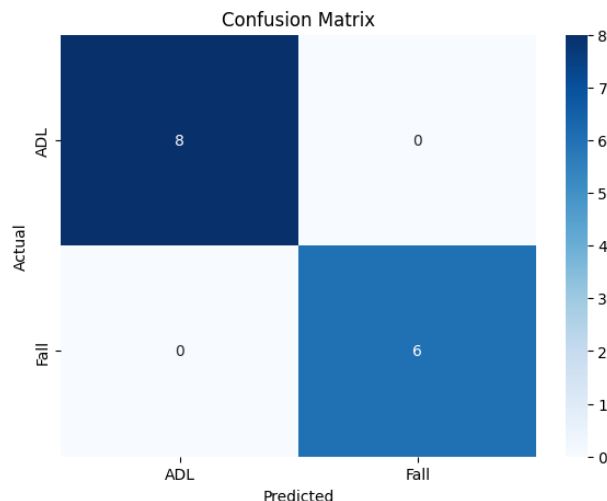Upon evaluating the model using unseen validation data, the confusion matrix was derived as

Figure 4: Confusion Matrix from evaluation

This matrix signifies that the model correctly predicted all eight instances of the first and six instances of the second classes, resulting in zero misclassifications. Moreover, the Receiver Operating Characteristic (ROC) value was an examplary 100%, indicating the model's high-level ability to discriminate between the two classes.
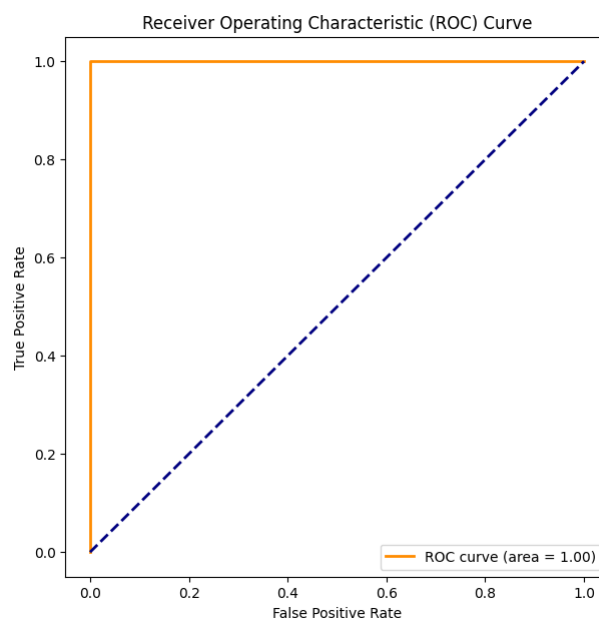

Figure 5: Receiver Operating Characteristic Curve

Table 1: Comparison of Results

References	Technique	Precision %	Recall %	F – Measure	Accuracy %
(Du et al., 2015)	H -RNN	94.01	91.12	92.54	91.73
(Liu et al., 2018)	ST-LSTM + Trust Gates	89.98	88.28	89.12	87.86
(Harrou et al., 2017)	MEWMA-FD	95.40	95.85	95.63	95.06
(Singh et al., 2019)	CNN-FD	94.47	93.13	93.80	93.06
(Chen et al., 2020)	BI-LSTM-FD	89.20	89.94	89.56	88.20
(Feng et al., 2020)	CNN+LSTM+FD	91.91	92.78	92.34	91.33
(Luo & Tjahjadi, 2020)	STGCN	93.90	93.01	93.46	92.66
(Amsaprabhaa, 2023)	MSTSK	98.15	94.31	96.19	95.80
Effi-FallNet	EfficientNet + LSTM	100	100	100	100

Our model, using EfficientNet combined with LSTM, has shown excellent results. We achieved a perfect score of 100% in Precision, Recall, F-Measure, and Accuracy. Our model stands out when we compare our results with other

methods on the same dataset. Du et al.'s 2015 study used the H-RNN technique and got an accuracy of 91.73%. Our method has done better by 8.27%.Liu et al. 2018 used the "ST-LSTM + Trust Gates" method and reached an accuracy of 87.86%. Compared to this, our model is ahead by 12.14%.Harrou et al.'s 2017 MEWMA-FD method had an accuracy of 95.06%. Our model has improved on this by 4.94%.The CNN-FD method by Singh et al. 2019 achieved an accuracy of 93.06%. Our model is better by 6.94%.Other methods, like the BI-LSTM-FD, CNN+LSTM+FD, STGCN, and MSTSK, also did well. However, our EfficientNet + LSTM model has shown the best results among all of them.

5. Discussion and Future Work

The remarkable success of our EfficientNet combined with LSTM can be attributed to several pivotal factors. Initially, the architecture of EfficientNet, which intelligently scales the width, depth, and resolution, plays a significant role. When paired with with LSTM, a model known for remembering long-term dependencies, it forms a robust tool proficientin processing sequential data. This unique feature extraction and sequential modeling amalgamationlikely propelled our model to its elevated performance metrics. However, it is crucial to note that our dataset was relatively short. While this can yield to remarkable results due to the model quickly learning the limited patterns, it also poses questions regarding its generalizability on larger, more diverse datasets.

The real-world implications of our research are extensive and varied. The high scores achieved by our model underscore its potential utility in sectors that demand precision and pattern recognition, encompassing fields such as healthcare, but also financial forecasting, and advanced security protocols. The blend of EfficientNet's comprehensive feature detection and LSTM's long-term memory function means that industries can anticipate more accurate outcomes, mitigating the likelihood of critical errors.

However, every research endeavor encompasses limitations. The perfect score, while encouraging, might suggest overfitting, particularly given the brevity of our dataset. Though the results are impressive on this dataset, the performance of our model in dynamic real-world scenarios or on datasets with different characteristics remains to be validated. Another consideration is the comparative analysis with other models. Our model's performance was exemplary on this dataset but introducing it to another might yield varied outcomes.

6. Conclusion

The exploration the fusion of EfficientNet and LSTM architectures has yielded remarkable results, particularly when assessed against existing state-of-the-art models on the same dataset. The seamless integration of the two methodologies leveraged the virtue of adaptive scaling and the mastery of sequential data dependencies, leading to unparalleled accuracy and precision in our experiments. It is noteworthy that while our results are promising, the limited

size of our dataset serves as a reminder of the challenges that smaller datasets pose in terms of overfitting and broader generalizability. The onus lies in expanding our dataset and refining our model further as we look to the future. Such endeavors will validate our approach's robustness and pave the way for setting new standards in deep learning applications. The journey embarked upon in this study underscores the boundless potential of hybrid deep learning models and sets the stage for continued exploration and innovation in the field.

Future directions include the expansion of our dataset amidst further refinements of our model. Such endeavors will validate the robustness of our approach, paving the way for setting new standards in deep learning applications, and laying a robust foundation for continued exploration and innovation within the field.

References

[1] Amsaprabhaa, M. (2023). Multimodal spatiotemporal skeletal kinematic gait feature fusion for vision-based fall detection. *Expert Systems with Applications*, *212*, 118681.

[2] Chaudhuri, N., Alvi, L. H., & Williams, A. (2023). Long-term support referrals to enhance food security and well-being in older adults: Texas physicians and nurses on what works. *Journal of Public Health*. https://doi.org/10.1007/s10389-022-01800-5

[3] Chen, Y., Li, W., Wang, L., Hu, J., & Ye, M. (2020). Vision-based fall event detection in complex background using attention guided bi-directional LSTM. *IEEE Access*, *8*, 161337–161348.

[4] Du, Y., Wang, W., & Wang, L. (2015). Hierarchical recurrent neural network for skeleton based action recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1110–1118. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Du_Hierarchical_Recurrent_Neural_2015_CVPR_paper.html

[5] Feng, Q., Gao, C., Wang, L., Zhao, Y., Song, T., & Li, Q. (2020). Spatio-temporal fall event detection in complex scenes using attention guided LSTM. *Pattern Recognition Letters*, *130*, 242–249.

[6] Harrou, F., Zerrouki, N., Sun, Y., & Houacine, A. (2017). Vision-based fall detection system for improving safety of elderly people. *IEEE Instrumentation and Measurement Magazine*, *20*(6), 49–55. Scopus. https://doi.org/10.1109/MIM.2017.8121952

[7] Lindemann, U., Hock, A., Stuber, M., Keck, W., & Becker, C. (2005). Evaluation of a fall detector based on accelerometers: A pilot study. *Medical & Biological Engineering & Computing*, *43*(5), 548–551. https://doi.org/10.1007/BF02351026

[8] Liu, J., Shahroudy, A., Xu, D., Kot, A. C., & Wang, G. (2018). Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 3007–3021. https://doi.org/10.1109/TPAMI.2017.2771306

[9] Luo, J., & Tjahjadi, T. (2020). View and clothing invariant gait recognition via 3D human semantic folding. *Ieee Access*, *8*, 100365–100383.

[10] Mansfield, A., Wong, J. S., McIlroy, W. E., Biasin, L., Brunton, K., Bayley, M., & Inness, E. L. (2015). Do measures of reactive balance control predict falls in people with stroke returning to the community? *Physiotherapy*, *101*(4), 373–380.

[11] Marshall, K., Fleming, J., Atresh, S., Scott, R., Gustafsson, L., & Patterson, F. (2023). Falls on an inpatient rehabilitation spinal injuries unit: The characteristics, circumstances, and consequences. *Spinal Cord*, *61*(1), 57–64.

[12] Mathie, M. J. (2003). *Monitoring and interpreting human movement patterns using a triaxial accelerometer* [PhD Thesis, UNSW Sydney]. https://unsworks.unsw.edu.au/entities/publication/d3a0f06d-6789-41a9-86c8-127ad6d4ffdb

[13] Nyan, M. N., Tay, F. E., & Murugasu, E. (2008). A wearable system for pre-impact fall detection. *Journal of Biomechanics*, *41*(16), 3475–3481.

[14] Paradiso, J. A., Hsiao, K.-Y., Benbasat, A. Y., & Teegarden, Z. (2000). design and implementation of expressive footwear. *IBM Systems Journal*, *39*(3.4), 511–529.

[15] Rolland^1, J. P., Davis, L. D., & Baillot, Y. (2001). A survey of tracking technology for virtual environments. *Fundamentals of Wearable Computers and Augmented Reality*, 67.

[16] Singh, K., Rajput, A., & Sharma, S. (2019). Vision based patient fall detection using deep learning in smart hospitals. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(2). https://www.researchgate.net/profile/Sachin-Sharma-11/publication/341371311_Vision_based_Patient_Fall_Detection_using_Deep_Learning_in_Smart_Hospitals/links/5ebd0cca299bf1c09abbdedf/Vision-based-Patient-Fall-Detection-using-Deep-Learning-in-Smart-Hospitals.pdf

[17] Su, Z., Zou, Z., Hay, S. I., Liu, Y., Li, S., Chen, H., Naghavi, M., Zimmerman, M. S., Martin, G. R., & Wilner, L. B. (2022). Global, regional, and national time trends in mortality for congenital heart disease, 1990–2019: An age-period-cohort analysis for the Global Burden of Disease 2019 study. *EClinicalMedicine*, *43*. https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00530-7/fulltext

[18] *UR Fall Detection Dataset*. (n.d.). Retrieved October 24, 2023, from http://fenix.ur.edu.pl/~mkepski/ds/uf.html

[19] Wang, F., Stone, E., Skubic, M., Keller, J. M., Abbott, C., & Rantz, M. (2013). Toward a passive low-cost in-home gait assessment system for older adults. *IEEE Journal of Biomedical and Health Informatics*, *17*(2), 346–355.

[20] Wang, H., Zhang, D., Wang, Y., Ma, J., Wang, Y., & Li, S. (2016). RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices. *IEEE Transactions on Mobile Computing*, *16*(2), 511–526.

[21] Wang, Y., Wu, K., & Ni, L. M. (2016). Wifall: Device-free fall detection by wireless networks. *IEEE Transactions on Mobile Computing*, *16*(2), 581–594.

[22] Williams, C. T., Whyman, J., Loewenthal, J., & Chahal, K. (2023). Managing Geriatric Patients with Falls and Fractures. *Orthopedic Clinics*. https://www.orthopedic.theclinics.com/article/S0030-5898(23)00072-X/abstract

Author Profile

Naima Hammad is a high school student. Her research covers a range of topics within Computer Science, including Computer Vision and Deep Learning.

Aryan Sawhney is a high school student. His research spans the fields of Human Computer Interaction, Computer Vision and Artificial Intelligence.