Data Mesh-Driven Enhancements in Sign Language Recognition Using LRCN and 3DCNN Models

Sunil Pradhan Sharma*, **Elakkiya Daivam****
*Senior Lead Software Engineer, Capital One Services, LLC
**Lead Software Engineer, Capital One Services, LLC
* Corresponding author
doi: https://doi.org/10.21467/proceedings.7.6.1

***Abstract*— This study aims to narrow the communication gap between sign language users and non-users by developing a sign language recognition application. Employing machine learning techniques, particularly the LRCN and 3DCNN models, the study modifies and enhances its baselines to meet its objectives. Techniques such as the sliding window method, dropout layers, and L_2 regularization are employed to expand the dataset and reduce overfitting, resulting in significant improvements in model accuracy. Utilizing the WLASL dataset, renowned for its extensive vocabulary and diverse signers, the study focuses on 10 selected classes for implementation. Comparative analysis based on the F1-score is conducted between LRCN and 3DCNN models, with both achieving results above 80% in each class and 95% and 94% overall, respectively. Resource usage monitoring indicates that LRCN requires more memory due to higher epochs needed for accuracy.**

***Keywords*— 3DCNN, LRCN, Machine Learning, WLASL**

I. Introduction

Learning methods to bridge the communication gap between sign language users and non-sign language users [1], [2], [3], [4], [5], [6]. Despite the prevalent use of sign language, there exists a significant communication barrier between sign language users and individuals unfamiliar with it. Therefore, this study endeavours to tackle this issue by creating a system capable of accurately recognizing sign language gestures and translating them into words. The study will make use of the Word-Level American Sign Language (WLASL) dataset, which is one of the largest datasets available for sign language recognition. This dataset comprises videos of various signers performing different signs, and the study will employ machine-learning techniques to build a model capable of recognizing and classifying these signs. The study has several aims, including the development of a robust sign language recognition system, improving model performance, evaluating model performance, and analysing the strengths and limitations of two machine learning models. The significance of this study lies in its potential to enhance communication between sign language users and non-sign language users, which could positively impact the daily lives of millions of people worldwide. Additionally, the study will contribute to the development of more inclusive technologies that support diversity and accessibility.

The study is as follows; similar papers are shown in the following section. The materials and methods are provided in Section III. The experimental analysis is carried out in Section IV, and in Section V, we provide some conclusions and plans for future research.

II. Related Works

Several studies have investigated the application of machine learning methods to predict and translate human gestures, particularly focusing on sign language recognition. The survey of intelligent tools for sign language recognition emphasizes the need to consider both feature extraction and sequence processing when dealing with input videos. Commonly used methods for feature extraction include Convolutional Neural Networks (CNN) [7], while sequence processing methods include Gated Recurrent Units (GRU) [8], Recurrent Neural Networks (RNN) [9], and Long Short-Term Memory (LSTM) [10]. Additionally, [11] have explored combining two or more models to enhance performance and have utilized various input data structures for training, such as RGB videos, skeleton, thermal, depth, and flow of information. [12] have used CNN and RNN for predicting words from self-collected RGB videos and have implemented data augmentation to increase dataset size. Challenges related to skin tones, facial features, and clothing affecting model accuracy have been highlighted by [13]. In the context of Bangla Sign Language (BSL), [14] have employed CNN and LSTM, achieving 88.5% testing accuracy while also addressing challenges related to overfitting during training.

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

Furthermore, comparisons between CNN combined with LSTM and RNN have been made in Chinese Sign Language (CSL) recognition, showing that LSTM performed better than RNN in terms of storing and accessing information by [15]. Similar methodologies have been applied in Indian sign language, where an additional LSTM was incorporated to improve model performance by [16]. In the realm of dynamic gesture recognition, 3DCNN has proved to be effective, achieving satisfactory precision, recall, and F-measure scores. To accurately detect both hands, CNN combined with spatial transform layers has been utilized in a multi-label environment by [17]. Additionally, 3DCNN has been used for feature extraction in sign language recognition without requiring prior model experience by [18]. In addition to CNN, other methods, such as 3D hand key points and LSTM have been employed for real-time sign language prediction, achieving over 99% accuracy for all signs. Challenges related to misclassification and accurate detection of both hands have been discussed by [19]. The transformer method was also utilized for continuous sign language translation. Challenges encountered included limitations in labelled datasets and the need for a large amount of information to optimize model performance by [20]. Based on the literature reviewed, our study aims to integrate data from the WLASL dataset, claimed to be one of the largest ASL datasets, with the LRCN model, which combines CNN and LSTM. The objective is to enhance model performance through this integration and to develop a prototype of a sign language recognition application.

III. Materials and Methods

The study utilizes machine learning techniques, specifically LSTM and 3DCNN, to identify sign language gestures. The dataset employed is the WLASL video dataset, comprising 2,000 words, 21,083 videos, and 119 signers. Owing to resource limitations, the model prototype exclusively focuses on 10 classes as shown in Fig. 1. The study encompasses data processing, model training, model tuning, evaluation metrics, result analysis, and deployment. Data processing involves grouping, frame extraction, data augmentation, and data partitioning. The frames are resized to 64 x 64 and limited to 50 per sequence as shown in Fig. 2. Data augmentation is utilized to expand the input data, and the dataset is split into 75% for training and 25% for testing and validation. The primary goal of the study is to develop a robust sign language recognition system capable of accurately interpreting sign language gestures and translating them into words. The study's success hinges on the practical design of the software solution, with an emphasis on structured design based on established software requirements. A comparative analysis between CNN combined with LSTM and 3DCNN is conducted to determine the optimal approach for sign language recognition.

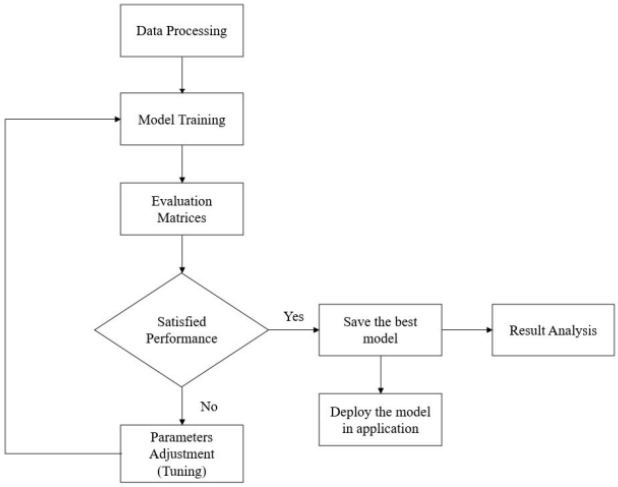

Fig. 1. Flowchart for overall experimental design

Fig. 2. Extracting frames

A. Model Analysis

The Long-Term Recurrent Convolutional Network (LRCN) [21], [22], [23], [24], [25], [26] method combines CNN and LSTM with time distributed layers. By leveraging time distributed layers, it encapsulates the input data, allowing it to be processed as a single entity. This approach aims to exploit both the spatial and temporal information embedded in sign language gestures for accurate recognition. The model structure, depicted in Fig. 3, feeds captured frames into the CNN model for feature extraction and then passes them to the LSTM mode for processing frame sequences. The baseline of the LRCN model structure includes a convolutional layer (Conv2D), a pooling layer (MaxPooling2D), a flatten layer, a dense layer, a time distributed layer, and an LSTM layer. Additionally, the 3D Convolutional Neural Network (3D-CNN) is suggested as a practical alternative for effectively classifying videos based on their temporal features. This model replaces 2D-CNN with 3D filters and pooling kernels, which account for

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

height, width, and time (or depth) as shown in Fig. 4. It is noted that the 3D model outperforms 2D-CNN in terms of accuracy and sensitivity, reducing false positives. The baseline structure of the 3DCNN model included a convolutional layer (Conv3D), a pooling layer (MaxPooling3D), a flatten layer, and a dense layer. Once the machine learning models were built, the process of compiling and training the models involved early stopping callback implementation to prevent overfitting, compilation using categorical cross-entropy as a loss function and Adam's method as an optimizer, and model training using training data to compare model performance with validation data during the training process, and adjusting parameters accordingly.

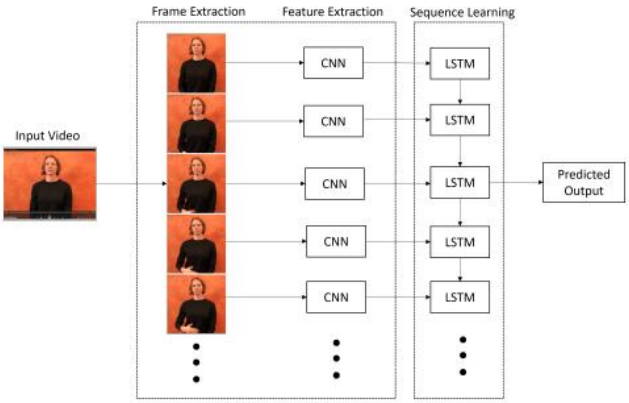

Fig. 3. Model LRCN

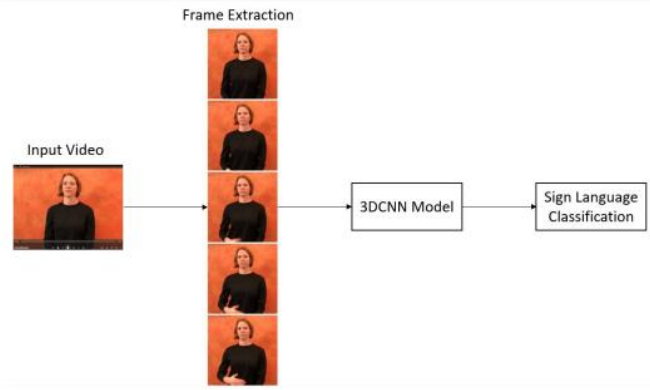

Fig. 4. Model 3DCNN

IV. Experimental Analysis

The outcomes of the baseline approach and the incorporation of the sliding window method, dropout, and L2 regularization with LRCN and 3DCNN models for sign language gesture recognition are shown in Figs. 5 to 14. The baseline models encountered challenges, including overfitting and low accuracy, which were mitigated by implementing data augmentation through the sliding window method. Subsequent enhancements were achieved by introducing dropout layers and L_2 regularization to improve the model's robustness and generalization performance. A comparison of the LRCN and 3DCNN models reveals that both models attained high F1-scores for each class, with the LRCN model outperforming in certain gestures and the 3DCNN model excelling in others. However, the 3DCNN model demands less memory and execution time compared to the LRCN model. The results suggest that the LRCN model is more adept at learning from and generalizing the dataset, while the 3DCNN model demonstrates greater efficiency in terms of resource usage. The selection of model architecture should be guided by specific task requirements, encompassing accuracy, memory usage, and execution time. This study offers valuable insights into the strengths and weaknesses of each model and underscores the significance of weighing resource usage alongside model performance. Overall, the study underscores the effectiveness of both LRCN and 3DCNN models for sign language gesture recognition and underscores the necessity for further research to enhance the accuracy and efficiency of these models. The results of this study can be leveraged to guide the development of more sophisticated models for sign language recognition and other computer vision tasks.

Fig. 5. Training loss and accuracy plot before applying sliding window of LRCN

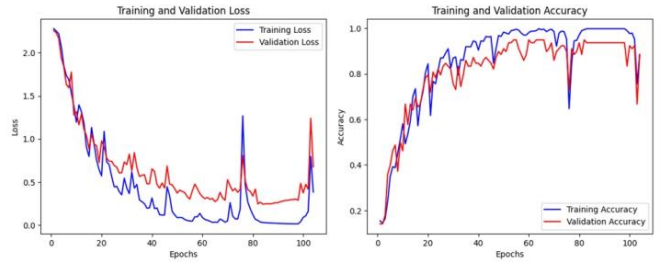

Fig. 6. Training loss and accuracy plot after applying sliding window of LRCN

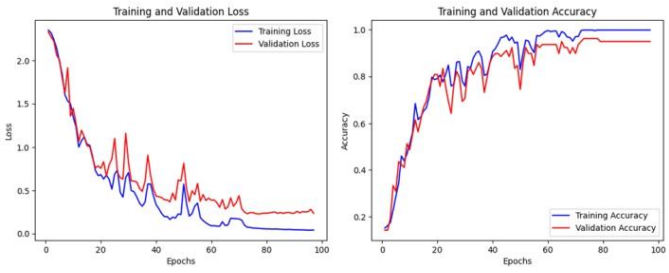

Fig. 7. Training loss and accuracy plot after applying dropout and l_2 of LRCN

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

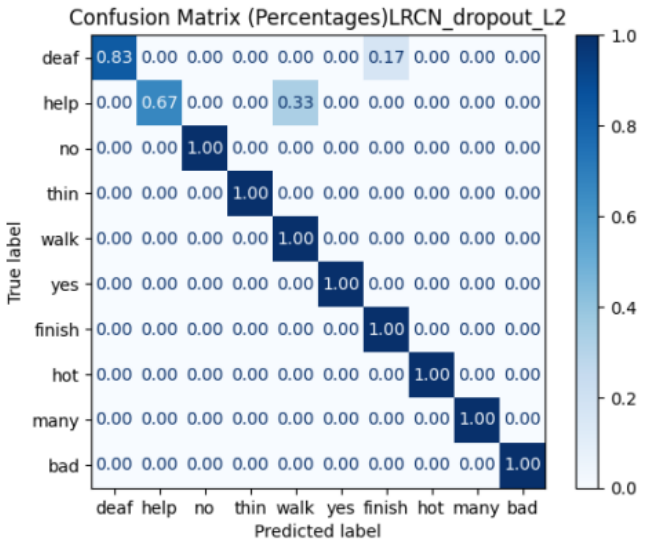

Fig. 8. Confusion matrix plot after applying dropout and l_2 of LRCN

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 525.105.17   Driver Version: 525.105.17   CUDA Version: 12.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   39C    P0    39W / 300W |   9398MiB / 16384MiB |      2%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
+-----------------------------------------------------------------------------+
```

Fig. 9. LRCN resource utilized

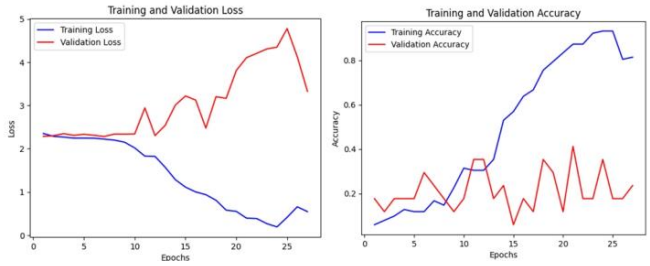

Fig. 10. Training loss and accuracy plot before applying sliding window of 3DCNN

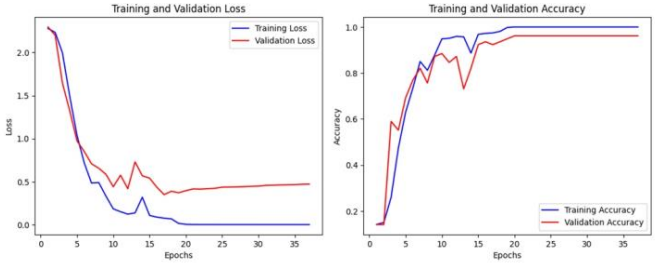

Fig. 11. Training loss and accuracy plot after applying sliding window of 3DCNN

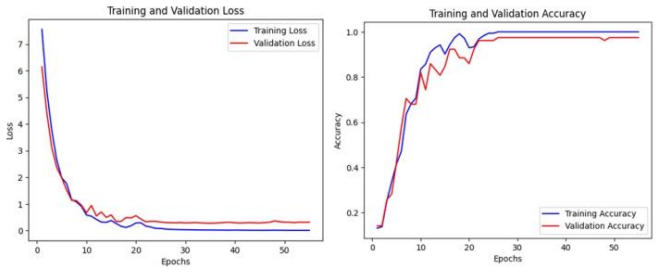

Fig. 12. Training loss and accuracy plot after applying dropout and l_2 of 3DCNN

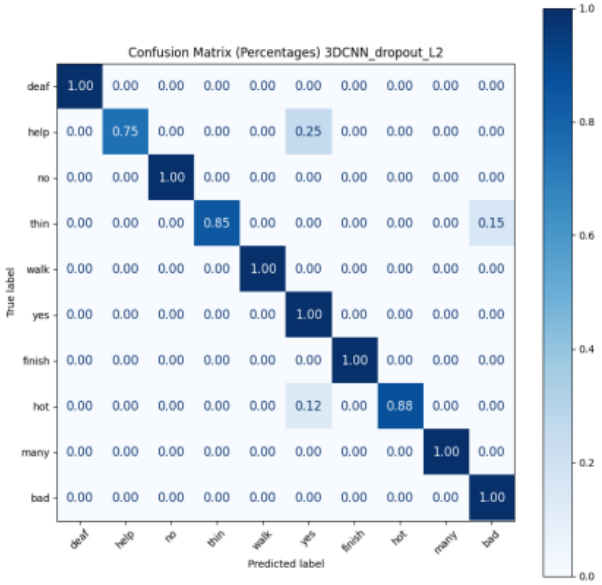

Fig. 13. Confusion matrix plot after applying dropout and l_2 of 3DCNN

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 525.105.17   Driver Version: 525.105.17   CUDA Version: 12.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla V100-SXM2...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   43C    P0    51W / 300W |   5166MiB / 16384MiB |     11%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
+-----------------------------------------------------------------------------+
```

Fig. 14. 3DCNN resource utilized

There are a few drawbacks to take into account, even with the LRCN and 3DCNN models' encouraging performance in sign language gesture detection. A significant constraint is the reliance on the caliber and volume of the training dataset. Even though both models show good F1-scores for gesture detection, their performance can deteriorate if they are exposed to other datasets, especially ones that differ in terms of backgrounds, lighting, or gesture styles. Since the models in this study were trained on a particular dataset, it's possible that the findings won't translate

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

well other datasets that are more varied or expansive. In order to assess the resilience and generalization ability of the models, future research should concentrate on enlarging the dataset by inserting gestures from various sign languages and contexts. The computational expense of training and implementing the models, especially the LRCN model, is another drawback. While the LRCN model outperformed the 3DCNN model in certain movements, it uses a lot more memory and takes a lot longer to execute. Because of this, it is less appropriate for situations with limited resources or real-time applications, like embedded systems or mobile devices. Although dropout and L_2 regularization aid in improving generalization and reducing overfitting, they do not fully address the difficulties in maximizing performance while minimizing processing costs. Subsequent studies ought to investigate more effective model architectures that can preserve accuracy while using less resources, like simplified versions of recurrent or convolutional models.

The sliding window method, which was used to enhance the data and enhance the temporal resolution of the input sequences, presents another difficulty. This approach creates a trade-off between the quality of gesture segmentation and model complexity, even though it helps reduce the problems caused by overfitting. The temporal dynamics of sign language movements may not always be faithfully captured by the sliding window method, particularly when the gestures involve complicated motions or are executed at different rates. This restriction may result in less accurate gesture segmentation, which would affect the recognition accuracy as a whole. Future research could look into more sophisticated temporal modeling methods, which are known to perform better than the sliding window approach at capturing temporal dependencies. Examples of these methods include attention mechanisms and transformer-based architectures. Furthermore, the work does not investigate other regularization procedures that can enhance the models' performance, even while it highlights the significance of regularization techniques like dropout and L_2 regularization. It may be possible to try strategies like early stopping, ensemble approaches, and batch normalization to improve the models' stability and convergence. Optimizing hyperparameters like regularization weights, dropout rates, and learning rates may result in better model performance. Finally, the study ignores the integration of other modalities that can enhance sign language recognition in favor of concentrating only on model design and regularization techniques. Facial expressions and lip movements are frequently used in conjunction with sign language motions to add context and greatly increase identification accuracy. A more complete approach to sign language identification may involve the integration of both visual and auditory cues using multimodal learning approaches. In order to create more reliable and accurate sign language recognition systems, future research should examine the use of multimodal techniques, including using 2D face landmark models or audio processing networks.

V. Conclusion and Future Works

The study focused on the development of a sign language recognition system aimed at facilitating communication for individuals with hearing impairments. Two machine learning models, LRCN and 3DCNN, were trained using the WLASL video dataset, achieving high F1-scores (above 80%) and weighted averages (95% and 94% respectively). However, LRCN necessitated more memory and time due to a higher number of epochs. Future enhancements are envisioned to include noise removal, dataset expansion, continuous learning, exploration of alternative approaches (e.g., vision transformer model), utilization of skeletal tracking data, and the development of real-time recognition applications. The ultimate goal of the study is to establish a more resilient, precise, and accessible sign language recognition system, thereby lessening communication barriers between sign and non-sign users. Subsequent research will be centred on devising more pragmatic solutions applicable in real-world scenarios.

VI. Declarations

A. ***Funding:*** No funds, grants, or other support was received.

B. ***Conflict of Interest:*** The authors declare that they have no known competing for financial interests or personal relationships that could have appeared to influence the work reported in this paper.

C. ***Data Availability:*** Data will be made on reasonable request.

D. Code Availability: Code will be made on reasonable request.

References

[1] G. S. Kashyap, K. Malik, S. Wazir, and R. Khan, "Using Machine Learning to Quantify the Multimedia Risk Due to Fuzzing," *Multimed. Tools Appl.*, vol. 81, no. 25, pp. 36685–36698, Oct. 2022, doi: 10.1007/s11042-021-11558-9.

[2] S. Wazir, G. S. Kashyap, and P. Saxena, "MLOps: A Review," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.10908v1

[3] G. S. Kashyap *et al.*, "Revolutionizing Agriculture: A Comprehensive Review of Artificial Intelligence Techniques in Farming," Feb. 2024, doi: 10.21203/RS.3.RS-3984385/V1.

[4] S. Naz and G. S. Kashyap, "Enhancing the predictive capability of a mathematical model for pseudomonas aeruginosa through artificial neural networks," *Int. J. Inf. Technol. 2024*, pp. 1–10, Feb. 2024, doi: 10.1007/S41870-023-01721-W.

[5] G. S. Kashyap *et al.*, "Detection of a facemask in real-time using deep learning methods: Prevention of Covid 19," Jan. 2024, Accessed: Feb. 04, 2024. [Online]. Available: https://arxiv.org/abs/2401.15675v1

[6] S. Wazir, G. S. Kashyap, K. Malik, and A. E. I. Brownlee, "Predicting the Infection Level of COVID-19 Virus Using Normal Distribution-Based Approximation Model and PSO," Springer, Cham, 2023, pp. 75–91. doi: 10.1007/978-3-031-33183-1_5.

[7] J. Carletta *et al.*, "The AMI Meeting Corpus: A pre-announcement," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Berlin, Heidelberg, 2006, pp. 28–39. doi: 10.1007/11677482_3.

[8] F. Abu-Amara, A. Bensefia, H. Mohammad, and H. Tamimi, "Robot and virtual reality-based intervention in autism: a comprehensive review," *Int. J. Inf. Technol.*, vol. 13, no. 5, pp. 1879–1891, Oct. 2021, doi: 10.1007/s41870-021-00740-9.

[9] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, Institute of Electrical and Electronics Engineers Inc., Jun. 2018, pp. 59–66. doi: 10.1109/FG.2018.00019.

[10] A. Anzulewicz, K. Sobota, and J. T. Delafield-Butt, "Toward the Autism Motor Signature: Gesture patterns during smart tablet gameplay identify children with autism," *Sci. Rep.*, vol. 6, Aug. 2016, doi: 10.1038/srep31107.

[11] A. Guarino, D. Malandrino, R. Zaccagnino, C. Capo, and N. Lettieri, "Touchscreen gestures as images. A transfer learning approach for soft biometric traits recognition," *Expert Syst. Appl.*, vol. 219, p. 119614, Jun. 2023, doi: 10.1016/j.eswa.2023.119614.

[12] M. A. Rilvan, J. Chao, and M. S. Hossain, "Capacitive Swipe Gesture Based Smartphone User Authentication and Identification," in *Proceedings - 2020 IEEE International Conference on Cognitive and Computational Aspects of Situation Management, CogSIMA 2020*, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 59–66. doi: 10.1109/CogSIMA49017.2020.9215998.

[13] A. R. Khan, T. Saba, M. Z. Khan, S. M. Fati, and M. U. G. Khan, "Classification of human's activities from gesture recognition in live videos using deep learning," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 10, p. e6825, May 2022, doi: 10.1002/cpe.6825.

[14] W. Guo, X. Liu, C. Lu, and L. Jing, "PIFall: A Pressure Insole-Based Fall Detection System for the Elderly Using ResNet3D," *Electron. 2024, Vol. 13, Page 1066*, vol. 13, no. 6, p. 1066, Mar. 2024, doi: 10.3390/ELECTRONICS13061066.

[15] D. Lee, A. Franchi, P. R. Giordano, H. Il Son, and H. H. Bülthoff, "Haptic teleoperation of multiple unmanned aerial vehicles over the internet," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2011, pp. 1341–1347. doi: 10.1109/ICRA.2011.5979993.

[16] T. L. Nguyen, S. K. Ro, and J. K. Park, "Study of ball screw system preload monitoring during operation based on the motor current and screw-nut vibration," *Mech. Syst. Signal Process.*, vol. 131, pp. 18–32, Sep. 2019, doi: 10.1016/j.ymssp.2019.05.036.

[17] S. D. Bansod and A. V. Nandedkar, "Crowd anomaly detection and localization using histogram of magnitude and momentum," *Vis. Comput.*, vol. 36, no. 3, pp. 609–620, Mar. 2020, doi: 10.1007/S00371-019-01647-0/FIGURES/9.

[18] D. Nishiyama, S. Arita, D. Fukui, M. Yamanaka, and H. Yamada, "Accurate fall risk classification in elderly using one gait cycle data and machine learning," *Clin. Biomech.*, vol. 115, p. 106262, May 2024, doi: 10.1016/j.clinbiomech.2024.106262.

[19] S. LaConte, S. Strother, V. Cherkassky, J. Anderson, and X. Hu, "Support vector machines for temporal classification of block design fMRI data," *Neuroimage*, vol. 26, no. 2, pp. 317–329, Jun. 2005, doi: 10.1016/j.neuroimage.2005.01.048.

[20] R. L. Hughes, "A continuum theory for the flow of pedestrians," *Transp. Res. Part B Methodol.*, vol. 36, no. 6, pp. 507–535, Jul. 2002, doi: 10.1016/S0191-2615(01)00015-7.

[21] G. S. Kashyap, A. Siddiqui, R. Siddiqui, K. Malik, S. Wazir, and A. E. I. Brownlee, "Prediction of Suicidal Risk Using Machine Learning Models," Dec. 25, 2021. Accessed: Feb. 04, 2024. [Online]. Available: https://papers.ssrn.com/abstract=4709789

[22] H. Habib, G. S. Kashyap, N. Tabassum, and T. Nafis, "Stock Price Prediction Using Artificial Intelligence Based on LSTM– Deep Learning Model," in *Artificial Intelligence & Blockchain in Cyber Physical Systems: Technologies & Applications*, CRC Press, 2023, pp. 93–99. doi: 10.1201/9781003190301-6.

[23] N. Marwah, V. K. Singh, G. S. Kashyap, and S. Wazir, "An analysis of the robustness of UAV agriculture field coverage using multi-agent reinforcement learning," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 2317–2327, May 2023, doi: 10.1007/s41870-023-01264-0.

[24] P. Kaur, G. S. Kashyap, A. Kumar, M. T. Nafis, S. Kumar, and V. Shokeen, "From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility," Feb. 2024, Accessed: Mar. 21, 2024. [Online]. Available: https://arxiv.org/abs/2402.16142v1

[25] F. Alharbi and G. S. Kashyap, "Empowering Network Security through Advanced Analysis of Malware Samples: Leveraging System Metrics and Network Log Data for Informed Decision-Making," *Int. J. Networked Distrib. Comput.*, pp. 1–15, Jun. 2024, doi: 10.1007/s44227-024-00032-1.

[26] M. Kanojia, P. Kamani, G. S. Kashyap, S. Naz, S. Wazir, and A. Chauhan, "Alternative Agriculture Land-Use Transformation Pathways by Partial-Equilibrium Agricultural Sector Model: A Mathematical Approach," Aug. 2023, Accessed: Sep. 16, 2023. [Online]. Available: https://arxiv.org/abs/2308.11632v1