Publication Series:
Joint DFH/UFA workshop
on AI in Medicine

Jürgen Hesser, Xavier Fresquet (eds.)

2nd Sorbonne-Heidelberg Workshop on AI in Medicine: Machine Learning for Multi-modal Data

June 25-27, 2025

Abstract

Machine Learning is transforming science, especially the way we do research in medicine. It can analyze non-linear dependencies of structured clinical data, and it is starting to support in the huge amount of existing text and other unstructured information to extract useful information using recent techniques based on large language models. There is also an increasing amount of specific omics data for each patient, which makes it hard to manually inspect all the details. This is where multimodal data analysis comes in, which is the focus of this year's AI in Medicine workshop. Researchers from Sorbonne and Heidelberg will give keynote speeches to provide insight into their research field, which will fuel discussions.

It brings together junior and senior researchers from Sorbonne University, Heidelberg University, and their partner universities in 4EU+. Scientific exchange takes center stage through participants' presentations & posters, keynotes by invited speakers, and discussions. Key techniques are trained during hands-on sessions, and social events invite you to network while experiencing the unique setting of the oldest German university and the environment of a vibrant student city.

Acknowledgements

This workshop is funded by Université franco-allemande (UFA) / Deutsch-Französische Hochschule (DFH) and co-funded through the 4EU+ 1CORE Project. This event is jointly organised by the Mannheim Institute for Intelligent Systems in Medicine (MIISM) and the Sorbonne Center for artificial intelligence (SCAI), and supported by the Flagship 3 of the 4EU+ European University Alliance.

Cite as

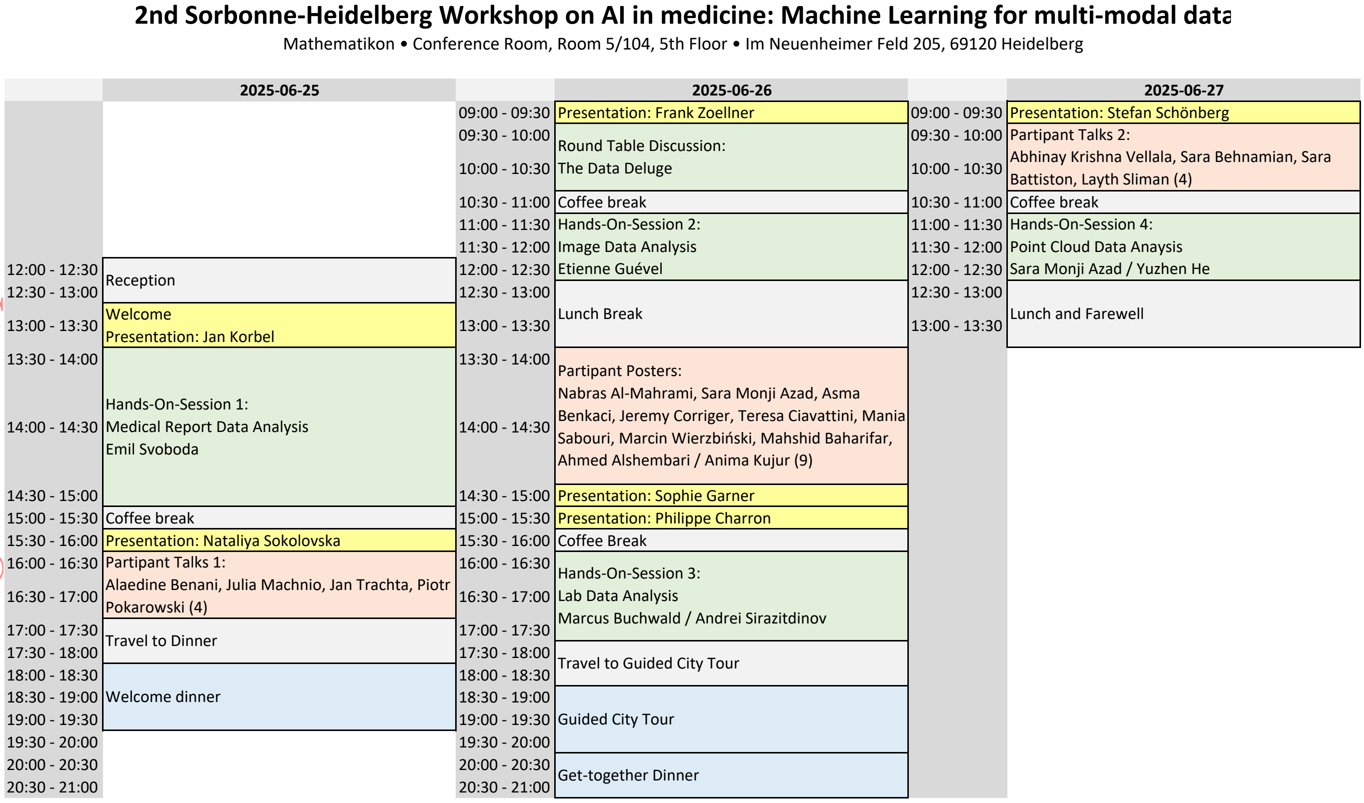
Jürgen Hesser, Xavier Fresquet (2025): 2nd Sorbonne-Heidelberg Workshop on AI in Medicine: Machine Learning for Multi-modal Data. Heidelberg University Library.
http://doi.org/10.11588/heidok.00036956

Published by Heidelberg University Library, 2025
The electronic version of this work is permanently available on:
https://archiv.ub.uni-heidelberg.de/volltextserver/

doi: http://doi.org/10.11588/heidok.00036956

Contents

,



2nd Sorbonne-Heidelberg Workshop on AI in medicine: Machine Learning for multi-modal data

Mathematikon • Conference Room, Room 5/104, 5th Floor • Im Neuenheimer Feld 205, 69120 Heidelberg

	2025-06-25		2025-06-26		2025-06-27
		09:00 - 09:30	Presentation: Frank Zoellner	09:00 - 09:30	Presentation: Stefan Schönberg
		09:30 - 10:00	Round Table Discussion: The Data Deluge	09:30 - 10:00	Partipant Talks 2: Abhinay Krishna Vellala, Sara Behnamian, Sara Battiston, Layth Sliman (4)
		10:00 - 10:30		10:00 - 10:30	
		10:30 - 11:00	Coffee break	10:30 - 11:00	Coffee break
		11:00 - 11:30	Hands-On-Session 2: Image Data Analysis Etienne Guével	11:00 - 11:30	Hands-On-Session 4: Point Cloud Data Anaysis Sara Monji Azad / Yuzhen He
		11:30 - 12:00		11:30 - 12:00	
12:00 - 12:30	Reception	12:00 - 12:30		12:00 - 12:30	
12:30 - 13:00		12:30 - 13:00	Lunch Break	12:30 - 13:00	Lunch and Farewell
13:00 - 13:30	Welcome Presentation: Jan Korbel	13:00 - 13:30		13:00 - 13:30	
13:30 - 14:00	Hands-On-Session 1: Medical Report Data Analysis Emil Svoboda	13:30 - 14:00	Partipant Posters: Nabras Al-Mahrami, Sara Monji Azad, Asma Benkaci, Jeremy Corriger, Teresa Ciavattini, Mania Sabouri, Marcin Wierzbiński, Mahshid Baharifar, Ahmed Alshembari / Anima Kujur (9)		
14:00 - 14:30		14:00 - 14:30			
14:30 - 15:00		14:30 - 15:00	Presentation: Sophie Garner		
15:00 - 15:30	Coffee break	15:00 - 15:30	Presentation: Philippe Charron		
15:30 - 16:00	Presentation: Nataliya Sokolovska	15:30 - 16:00	Coffee Break		
16:00 - 16:30	Partipant Talks 1: Alaedine Benani, Julia Machnio, Jan Trachta, Piotr Pokarowski (4)	16:00 - 16:30	Hands-On-Session 3: Lab Data Analysis Marcus Buchwald / Andrei Sirazitdinov		
16:30 - 17:00		16:30 - 17:00			
17:00 - 17:30	Travel to Dinner	17:00 - 17:30			
17:30 - 18:00		17:30 - 18:00	Travel to Guided City Tour		
18:00 - 18:30	Welcome dinner	18:00 - 18:30			
18:30 - 19:00		18:30 - 19:00	Guided City Tour		
19:00 - 19:30		19:00 - 19:30			
19:30 - 20:00		19:30 - 20:00			
20:00 - 20:30		20:00 - 20:30	Get-together Dinner		
20:30 - 21:00		20:30 - 21:00			

,

Towards Scalable and Robust White Matter Lesion Localization via Multimodal Deep Learning

Julia Machnio
Pioneer Centre for AI
University of Copenhagen
juma@di.ku.dk

Sebastian Nørgaard Llambias
Pioneer Centre for AI
University of Copenhagen
snl@di.ku.dk

Mads Nielsen
Pioneer Centre for AI
University of Copenhagen
madsn@di.ku.dk

Mostafa Mehdipour Ghazi
Pioneer Centre for AI
University of Copenhagen
ghazi@di.ku.dk

Abstract

White matter hyperintensities (WMH) are radiological markers of small vessel disease and neurodegeneration, whose accurate segmentation and spatial localization are crucial for diagnosis and monitoring. While multimodal MRI offers complementary contrasts for detecting and contextualizing WM lesions, existing approaches often lack flexibility in handling missing modalities and fail to integrate anatomical localization efficiently. We propose a deep learning framework for WM lesion segmentation and localization that operates directly in native space using single- and multi-modal MRI inputs. Our study evaluates four input configurations: FLAIR-only, T1-only, concatenated FLAIR and T1, and a modality-interchangeable setup. It further introduces a multi-task model for jointly predicting lesion and anatomical region masks to estimate region-wise lesion burden. Experiments conducted on the MICCAI WMH Segmentation Challenge dataset demonstrate that multimodal input significantly improves the segmentation performance, outperforming unimodal models. While the modality-interchangeable setting trades accuracy for robustness, it enables inference in cases with missing modalities. Joint lesion-region segmentation using multi-task learning was less effective than separate models, suggesting representational conflict between tasks. Our findings highlight the utility of multimodal fusion for accurate and robust WMH analysis, and the potential of joint modeling for integrated predictions.

Keywords: *Multimodal deep learning, segmentation, localization, white matter hyperintensity, magnetic resonance imaging*

1. Introduction

White matter hyperintensities (WMH) are pathological abnormalities of the brain's white matter that commonly present as hyperintense areas on FLAIR images and hypointensities on T1-weighted MRIs [26]. The total WMH burden, typically measured as lesion volume, increases with age and is recognized as a marker of early neurodegeneration. It is associated with elevated risk of Alzheimer's disease [3], dementia [6], and ischemic stroke [2], among other conditions. An accurate diagnosis requires not only the detection of the lesion, but also a detailed assessment of its volume and spatial characteristics, considered alongside the clinical context [7, 10, 21, 25]. Manual annotation remains the clinical gold standard; however, it is labor-intensive and not easily scalable, underscoring the need for automated tools for WMH segmentation and localization.

Multimodal MRI has become an essential tool in clinical neuroimaging by leveraging complementary information from multiple sequences, such as T1-weighted and FLAIR scans [15]. FLAIR images enhance lesion visibility due to cerebrospinal fluid suppression [23], while T1-weighted scans provide superior anatomical contrast and more precise delineation of brain structures [11]. Although numerous studies have utilized multimodal inputs to enhance segmentation accuracy [14, 24], their potential for improving lesion localization or handling missing modalities remains relatively underexplored.

Recent findings have highlighted that the spatial distribution of lesions carries significant diagnostic and prognostic value [1]. However, deriving such insights often relies on resource-intensive pipelines. For instance, Coenen et al. [4, 5] manually harmonized MRI data across cohorts to perform voxel-wise and region-of-interest analyses. While informative, these workflows are limited by their constrained scalability and reproducibility. This points to a need for automated, registration-free methods that enable accurate lesion localization directly in subject space.

Deep learning methods have demonstrated strong performance in automating WMH segmentation [12, 14, 24]. Some approaches incorporate anatomical priors, such as distances to known brain landmarks [8, 9], or rely on registering lesions in a common template space [12]. In our previous work [18], we developed a deep learning method for segmenting anatomical regions in native space, removing the need for spatial alignment. However, that approach did not fully address multimodal integration or the tradeoffs in joint lesion-region prediction for WMH localization.

In this study, we present a deep learning framework for WM lesion segmentation and localization that supports both single- and multi-modal MRI inputs. Our method performs voxel-wise segmentation of WMH and anatomical regions directly in native space. We examine four input configurations: (1) FLAIR-only, (2) T1-only, (3) concatenated FLAIR and T1, and (4) interchangeable modality training using either FLAIR or T1. Furthermore, we train unified models that jointly predict regional lesion labels, enabling direct estimation of region-wise lesion burden.

Our experiments show that while multimodal inputs improve segmentation accuracy, multi-task learning introduces a trade-off, with reduced multimodal performance compared to task-specific models. Nevertheless, multi-task models offer practical benefits, such as reduced inference time and integrated anatomical insights. Overall, our findings suggest that carefully optimized multimodal and multitask models can provide a scalable, robust, and anatomically informed solution for WM lesion analysis in both clinical and research settings.

2. Methods

2.1. Multimodal Configurations

We present a deep learning approach for WM lesion segmentation and localization, utilizing both single- and multimodal MRI inputs. As illustrated in Figure 1, we examine four input configurations: (A) FLAIR only, (B) T1 only, (C) FLAIR and T1 concatenated as separate input channels, and (D) FLAIR and T1 treated as interchangeable modalities during training. Configurations (A) and (B) represent unimodal input settings, while (C) and (D) implement alternative strategies for multimodal integration. In configuration (D), the two modalities are considered interchangeable variants, effectively augmenting the training set and encouraging robustness to missing modality scenarios.

The same modeling strategy is applied for both WM lesion and anatomical region segmentation. Specifically, WM region labels are used in place of lesion masks in the architecture shown in Figure 1. Beyond multimodal input strategies, we also investigate joint WM localization, as illustrated in Figure 2, where WM lesions and anatomical regions are segmented simultaneously by a shared network. These models are trained with multimodal inputs to predict lesion masks restricted to white matter regions, using masked regional labels (element-wise multiplication of the binary lesion masks with the corresponding WM region labels) as supervision. This design enables the simultaneous estimation of lesion burden and anatomical localization in a single forward pass.

2.2. White Matter Labels

Ground truth labels for training the WM region segmentation models were derived from the refined reference labels provided by the JHU MNI White Matter Atlas Type II [20]. The refined version of the atlas delineates 34 white matter subregions, selected based on ontological hierarchies and clinical relevance. To generate subject-specific label maps, the atlas T1 image was affinely registered to each subject's T1 scan using the extracted WM region [19]. The resulting transformation was then applied to the atlas region labels, yielding anatomically aligned WM labels in the subject's native space. The complete preprocessing pipeline is described in detail in [18].

2.3. Training and Inference

All models are based on the 3D U-Net architecture described in [22]. To improve robustness, we apply extensive MRI-specific data augmentation during training [16]. These augmentations include additive and multiplicative noise, bias field distortion, elastic deformations, random rotations, and simulated motion artifacts. Model optimization is performed using a composite loss function that combines cross-entropy (CE) loss with the Dice-Sørensen (DS) loss. During inference, configuration (C) produces a single prediction directly from the concatenated multimodal input. For input configurations (A), (B), and (D), predictions from T1 and FLAIR scans are fused by averaging their softmax outputs, followed by voxel-wise `ArgMax`.

3. Experiments and Results

3.1. Data

We conducted all experiments using the MICCAI 2017 White Matter Hyperintensity (WMH) Segmentation Challenge dataset [13], which contains co-registered 3D FLAIR and T1 MRI scans from 170 subjects across three clinical sites: Utrecht, Amsterdam, and Singapore. The dataset provides expert-annotated lesion masks that differentiate WM lesions from healthy tissue and other pathologies.

To increase the number of samples for training lesion localization models, we inverted the original challenge-

,

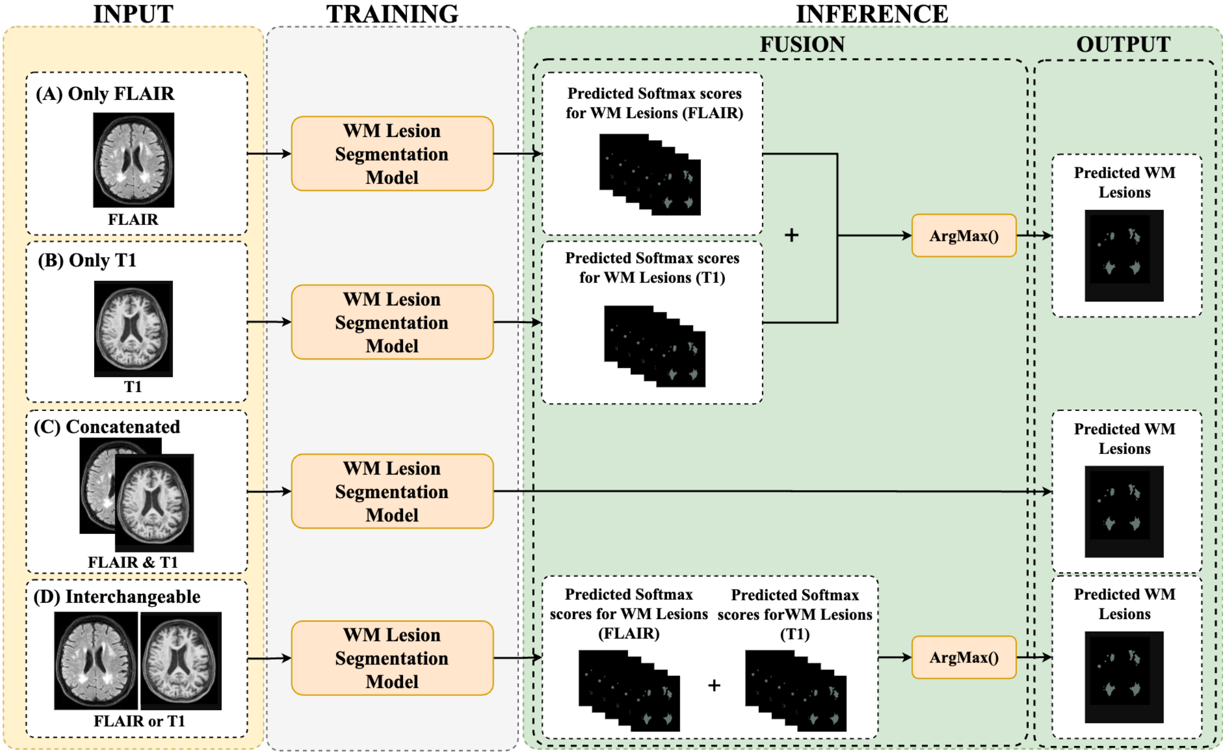

INPUT
TRAINING
INFERENCE
FUSION
OUTPUT
(A) Only FLAIR
FLAIR
(B) Only T1
T1
(C) Concatenated
FLAIR & T1
(D) Interchangeable
FLAIR or T1
WM Lesion Segmentation Model
WM Lesion Segmentation Model
WM Lesion Segmentation Model
WM Lesion Segmentation Model
Predicted Softmax scores for WM Lesions (FLAIR)
Predicted Softmax scores for WM Lesions (T1)
+
ArgMax()
Predicted Softmax scores for WM Lesions (FLAIR)
Predicted Softmax scores forWM Lesions (T1)
+
ArgMax()
Predicted WM Lesions
Predicted WM Lesions
Predicted WM Lesions

Figure 1. Overview of the proposed method for WM lesion segmentation. The pipeline illustrates four input configurations used during training: (A) FLAIR only, (B) T1 only, (C) FLAIR and T1 concatenated as separate input channels, and (D) sequential training where FLAIR and T1 are treated as interchangeable modalities and passed independently through the model. For configurations (A), (B), and (D), the final prediction is obtained by voxel-wise `ArgMax` fusion across the individual softmax outputs. The same pipeline is also used for WM region segmentation.

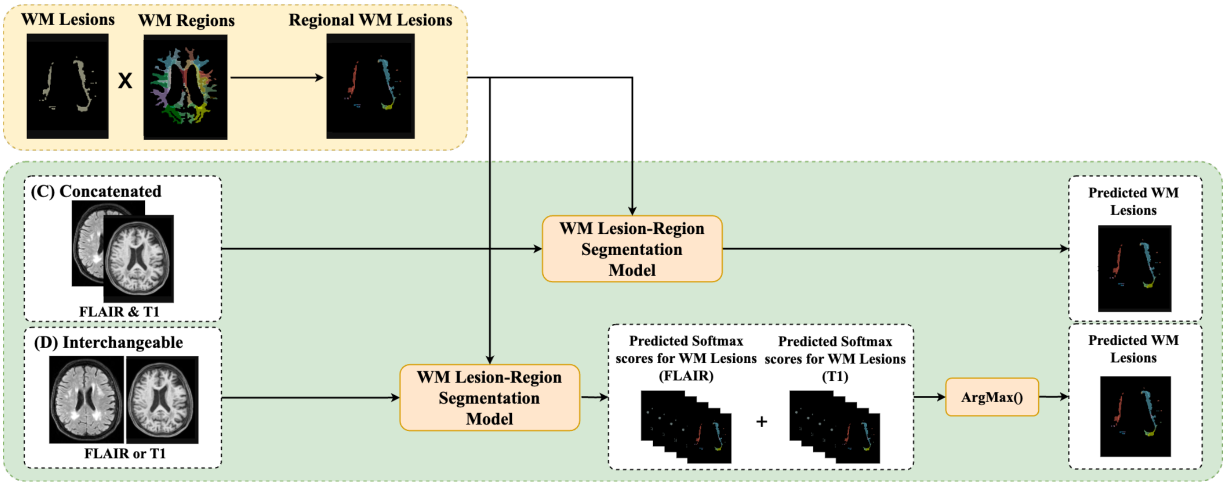

WM Lesions
X
WM Regions
Regional WM Lesions
(C) Concatenated
FLAIR & T1
(D) Interchangeable
FLAIR or T1
WM Lesion-Region Segmentation Model
WM Lesion-Region Segmentation Model
Predicted Softmax scores for WM Lesions (FLAIR)
+
Predicted Softmax scores for WM Lesions (T1)
ArgMax()
Predicted WM Lesions
Predicted WM Lesions

Figure 2. Overview of the proposed multi-task framework for multimodal regional WMH segmentation. The pipeline adopts the same input configurations as in Figure 1 (C–D). In this setting, WM lesions and anatomical regions are jointly segmented using a unified model. Region-specific lesion labels are generated by intersecting lesion masks with WM region annotations.

Table 1. Overview of the WMH data used in this study. The vendor abbreviations refer to GE (G), Philips (P), and Siemens (S).

Dataset Split	#Subjects	Dimensions	Resolution	Strength	Vendor
Train	110	181×251×81	1.1×1×3mm^3	1.5T,3T	G, P, S
Test	60	202×250×60	1.1×0.98×3mm^3	3T	G, P, S

defined splits, repurposing the original test set for training. The resulting dataset was then used for 5-fold cross-validation. Summary statistics are provided in Table 1.

3.2. Settings

For all models, the composite loss function was computed as an equal-weighted sum of CE and DS losses. The 3D U-Net models were implemented in PyTorch [17] and trained on NVIDIA A100 GPUs with 80 GB of VRAM. Optimization was performed using stochastic gradient descent (SGD) with Nesterov momentum set to 0.9 and an initial learning rate of 0.001. Each model was trained for 1,000 iterations, with up to 250 mini-batches per epoch. In the multi-task setup, we used a batch size of 12 and 3D input patches of size 32 × 128 × 128 voxels.

3.3. Results

WM Lesion Segmentation We first evaluated the performance of the WM lesion segmentation models on the test set. Table 2 presents the results across different training and inference configurations. The model trained with concatenated T1 and FLAIR inputs achieved the highest Dice score of 0.74, highlighting the advantage of leveraging complementary multimodal information where T1 provides anatomical detail, while FLAIR emphasizes contrast between healthy and pathological tissue.

Models trained on a single modality showed slightly reduced performance: the FLAIR-only model reached a DSC of 0.72, while the T1-only model scored 0.59. The model trained with T1 and FLAIR as interchangeable modalities yielded a lower overall accuracy of 0.67, but it offers unique robustness. This configuration supports inference with just one modality available, making it particularly valuable in clinical contexts where one sequence may be missing or degraded. While slightly less accurate, it provides increased flexibility for real-world deployment.

WM Region Segmentation We next evaluated the performance of anatomical WM region segmentation. As shown in Table 3, all input configurations achieved similarly high Dice scores, averaging around 0.75, indicating that regional white matter structures can be reliably segmented regardless of modality. Models evaluated using only FLAIR showed a slight reduction in accuracy compared to those using T1 or both modalities, reflecting the lower anatomical contrast inherent in FLAIR images. The multimodal ensemble model produced consistent results across all evaluation settings.

Table 2. Test DSC (mean±SD) for WM lesion segmentation using 3D U-Net models with different training inputs (rows) and inference strategies (columns). The “T1 & FLAIR” column reports voxel-wise predictions obtained by using softmax outputs from processed T1 and FLAIR inputs.

Training / Inference	**T1**	**FLAIR**	**T1 & FLAIR**
FLAIR	-	0.72 ± 0.12	0.68 ± 0.14
T1	0.59 ± 0.16	-	
T1 and FLAIR	-	-	0.74 ± 0.11
T1 or FLAIR	0.58 ± 0.17	0.73 ± 0.11	0.67 ± 0.15

Table 3. Test DSC (mean±SD) for WM region segmentation using 3D U-Net models with different training inputs (rows) and inference strategies (columns). The “T1 & FLAIR” column reports voxel-wise predictions obtained by using softmax outputs from processed T1 and FLAIR inputs.

Training / Inference	**T1**	**FLAIR**	**T1 & FLAIR**
FLAIR	-	0.75 ± 0.05	0.75 ± 0.05
T1	0.75 ± 0.05	-	
T1 and FLAIR	-	-	0.75 ± 0.05
T1 or FLAIR	0.75 ± 0.05	0.70 ± 0.06	0.74 ± 0.06

Table 4. Test DSC (mean±SD) for WM lesion segmentation using 3D U-Nets trained for regional WMH label segmentation with different training inputs (rows) and inference strategies (columns). The “T1 & FLAIR” column reports voxel-wise predictions obtained by using softmax outputs from T1 and FLAIR inputs.

Training / Inference	**T1**	**FLAIR**	**T1 & FLAIR**
T1 and FLAIR	-	-	0.43 ± 0.20
T1 or FLAIR	0.27 ± 0.18	0.36 ± 0.19	0.26 ± 0.19

Table 5. Test DSC (mean±SD) for WM region segmentation using 3D U-Nets trained for regional WMH label segmentation with different training inputs (rows) and inference strategies (columns). The “T1 & FLAIR” column reports voxel-wise predictions obtained by using softmax outputs from T1 and FLAIR inputs.

Training / Inference	**T1**	**FLAIR**	**T1 & FLAIR**
T1 and FLAIR	-	-	0.29 ± 0.12
T1 or FLAIR	0.17 ± 0.11	0.25 ± 0.11	0.17 ± 0.11

WM Lesion Localization Finally, we trained unified models to jointly segment regional WM lesions within a single network. Tables 4 and 5 summarize the corresponding lesion and region segmentation results. Although this approach offers a compact framework for simultaneously predicting regional lesion labels, it exhibited notably reduced performance compared to separate single-task models. In the multimodal configuration, the lesion segmentation Dice score declined from 0.74 to 0.43, while region segmentation dropped from 0.75 to 0.29.

3.4. Discussion

Our results demonstrate the advantage of multimodal input for WM lesion segmentation, with the highest performance achieved when T1 and FLAIR images were concatenated and jointly processed. This finding reinforces prior work showing that multimodal MRI leverages complementary contrasts: FLAIR enhances lesion visibility due to CSF suppression, while T1 provides clearer anatomical context. Although FLAIR-only models outperformed T1-only models, consistent with FLAIR's superior lesion contrast, the combined input configuration offered improved spatial precision and generalization.

The modality-interchangeable configuration, in which T1 and FLAIR were treated as alternative inputs, yielded lower segmentation performance. Nevertheless, this approach offers a practical advantage: the ability to operate when only one modality is available. This robustness is particularly valuable in real-world clinical scenarios, where incomplete or corrupted data are common. In such settings, the flexibility of this configuration may outweigh the modest reduction in accuracy, especially for large-scale studies or multi-site applications with variable imaging protocols.

For WM region segmentation, performance was more consistent across input types. Most configurations achieved comparable accuracy, with FLAIR-only predictions showing slightly lower performance, likely due to the reduced anatomical contrast in FLAIR images. These results indicate that while FLAIR is well-suited for lesion detection, T1-weighted images remain more informative for anatomical delineation of WM subregions.

In our final set of experiments, we explored a multi-task learning setup where the model jointly segmented lesions and anatomical regions. This configuration resulted in a marked performance drop relative to the single-task models. The reduced accuracy may reflect optimization conflicts or representational interference between the two tasks. To fairly compare the multi-task models with the single-task baselines, we evaluated lesion and region segmentation separately. For lesion assessment, we combined all predicted lesion subregion labels into a single binary mask. For region segmentation, we evaluated predictions only for WM subregions present in each scan. This ensured consistent and representative comparison across all settings.

4. Conclusion

We presented a systematic study of deep learning strategies for WM lesion segmentation and localization using single- and multimodal MRI inputs. Our framework evaluated multiple input configurations, including unimodal (FLAIR or T1), concatenated multimodal, and modality-interchangeable training. We further extended this setup to jointly segment WM lesions and anatomical subregions via multi-task learning. Experiments on the WMH segmentation dataset demonstrated that combining T1 and FLAIR inputs in a shared model yields the highest segmentation performance, outperforming unimodal baselines. While the modality-interchangeable setup underperformed slightly, it offers robustness in scenarios with missing or incomplete modalities, which is critical for clinical deployment.

Compared to state-of-the-art WMH segmentation approaches, our results reaffirm the importance of multimodal fusion for accurate lesion delineation, and highlight the limitations of single-task models in capturing spatial lesion distribution across anatomical regions. Our multi-task model, designed to jointly segment lesions and WM regions, led to performance degradation, indicating possible interference between task objectives. These findings suggest that while joint learning is promising for efficient inference and spatial lesion quantification, careful architectural and training considerations are necessary. Further investigation is warranted before drawing definitive conclusions. For instance, future work could explore alternative training strategies in which lesion and region labels are treated as separate binary outputs, reducing task entanglement during optimization.

Acknowledgments

This project has received funding from the Pioneer Centre for AI, Danish National Research Foundation (DNRF), grant number P1.

References

[1] J Matthijs Biesbroek, Nick A Weaver, and Geert Jan Biessels. Lesion location and cognitive impact of cerebral small vessel disease. *Clinical Science*, 131(8):715–728, 2017. 1

[2] Anna K Bonkhoff, Sungmin Hong, Martin Bretzner, Markus D Schirmer, Robert W Regenhardt, E Murat Arsava, Kathleen Donahue, Marco Nardin, Adrian Dalca, Anne-Katrin Giese, et al. Association of stroke lesion pattern and white matter hyperintensity burden with stroke severity and outcome. *Neurology*, 99(13):e1364–e1379, 2022. 1

[3] Adam M Brickman, Laura B Zahodne, Vanessa A Guzman, Atul Narkhede, Irene B Meier, Erica Y Griffith, Frank A Provenzano, Nicole Schupf, Jennifer J Manly, Yaakov Stern, et al. Reconsidering harbingers of dementia: progression of parietal lobe white matter hyperintensities predicts alzheimer's disease incidence. *Neurobiology of aging*, 36 (1):27–32, 2015. 1

[4] Mirthe Coenen, Hugo J Kuijf, Irene MC Huenges Wajer, Marco Duering, Frank J Wolters, Evan F Fletcher, Pauline M Maillard, Alzheimer's Disease Neuroimaging Initiative, Frederik Barkhof, Josephine Barnes, et al. Strategic white matter hyperintensity locations for cognitive impairment: a multicenter lesion-symptom mapping study in 3525 memory clinic patients. *Alzheimer's & Dementia*, 19 (6):2420–2432, 2023. 1

[5] Mirthe Coenen, Floor AS de Kort, Nick A Weaver, Hugo J

Kuijf, Hugo P Aben, Hee-Joon Bae, Régis Bordet, Christopher PLH Chen, Anna Dewenter, Thomas Doeven, et al. Strategic white matter hyperintensity locations associated with post-stroke cognitive impairment: A multicenter study in 1568 stroke patients. *International Journal of Stroke*, page 17474930241252530, 2024. 1

[6] Stéphanie Debette and HS Markus. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *Bmj*, 341, 2010. 1

[7] Tracy d'Arbeloff, Maxwell L Elliott, Annchen R Knodt, Tracy R Melzer, Ross Keenan, David Ireland, Sandhya Ramrakha, Richie Poulton, Tim Anderson, Avshalom Caspi, et al. White matter hyperintensities are common in midlife and already associated with cognitive decline. *Brain Communications*, 1(1):fcz041, 2019. 1

[8] Mohsen Ghafoorian, Nico Karssemeijer, Inge van Uden, Frank Erik de Leeuw, Tom Heskes, Elena Marchiori, and Bram Platel. Small white matter lesion detection in cerebral small vessel disease. In *Medical Imaging 2015: Computer-Aided Diagnosis*, pages 265–270. SPIE, 2015. 2

[9] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Inge WM van Uden, Clara I Sanchez, Geert Litjens, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, and Bram Platel. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):5110, 2017. 2

[10] Saima Hilal, J Matthijs Biesbroek, Henri Vrooman, Eddie Chong, Hugo J Kuijf, Narayanaswamy Venketasubramanian, Ching-Yu Cheng, Tien Yin Wong, Geert Jan Biessels, and Christopher Chen. The impact of strategic white matter hyperintensity lesion location on language. *The American Journal of Geriatric Psychiatry*, 29(2):156–165, 2021. 1

[11] Clare Howarth, Chloe Hutton, and Ralf Deichmann. Improvement of the image quality of t1-weighted anatomical brain scans. *Neuroimage*, 29(3):930–937, 2006. 1

[12] Wenhao Jiang, Fengyu Lin, Jian Zhang, Taowei Zhan, Peng Cao, and Silun Wang. Deep-learning-based segmentation and localization of white matter hyperintensities on magnetic resonance images. *Interdisciplinary Sciences: Computational Life Sciences*, 12(4):438–446, 2020. 2

[13] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11): 2556–2568, 2019. 2

[14] Li Liang, Pengzheng Zhou, Wanxin Lu, Xutao Guo, Chenfei Ye, Haiyan Lv, Tong Wang, and Ting Ma. An anatomical knowledge-based mri deep learning pipeline for white matter hyperintensity quantification associated with cognitive impairment. *Computerized Medical Imaging and Graphics*, 89: 101873, 2021. 1, 2

[15] Heidi Lindroth, Veena A Nair, Casandra Stanfield, Cameron Casey, Rosaleena Mohanty, Daniel Wayer, Paul Rowley, Roger Brown, Vivek Prabhakaran, and Robert D Sanders. Examining the identification of age-related atrophy between t1 and t1+ t2-flair cortical thickness measurements. *Scientific reports*, 9(1):11288, 2019. 1

[16] Sebastian Nørgaard Llambias, Mads Nielsen, and Mostafa Mehdipour Ghazi. Data augmentation-based unsupervised domain adaptation in medical imaging. *arXiv preprint arXiv:2308.04395*, 2023. 2

[17] Sebastian Nørgaard Llambias, Julia Machnio, Asbjørn Munk, Jakob Ambsdorf, Mads Nielsen, and Mostafa Mehdipour Ghazi. Yucca: A deep learning framework for medical image analysis. *arXiv preprint arXiv:2407.19888*, 2024. 4

[18] Julia Machnio, Mads Nielsen, and Mostafa Mehdipour Ghazi. Deep learning for localization of white matter lesions in neurological diseases. In *Northern Lights Deep Learning Conference 2025*, 2024. 2

[19] Mostafa Mehdipour Ghazi and Mads Nielsen. Fast-aid brain: Fast and accurate segmentation tool using artificial intelligence developed for brain. *arXiv preprint arXiv:2208.14360*, 2022. 2

[20] Kenichi Oishi, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Jiangyang Zhang, John T Hsu, Michael I Miller, Peter CM van Zijl, Marilyn Albert, et al. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and alzheimer's disease participants. *Neuroimage*, 46(2):486–499, 2009. 2

[21] Jakob Rath, Olivia Foesleitner, Lukas Haider, Hubert Bickel, Fritz Leutmezer, Stephan Polanec, Michael A Arnoldner, Gere Sunder-Plassmann, Daniela Prayer, Thomas Berger, et al. Neuroradiological differentiation of white matter lesions in patients with multiple sclerosis and fabry disease. *Orphanet Journal of Rare Diseases*, 17(1):37, 2022. 1

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[23] Pascal Sati, Ilena C George, Colin D Shea, María I Gaitán, and Daniel S Reich. Flair*: a combined mr contrast technique for visualizing white matter lesions and parenchymal veins. *Radiology*, 265(3):926–932, 2012. 1

[24] Dan Wu, Marilyn Albert, Anja Soldan, Corinne Pettigrew, Kenichi Oishi, Yusuke Tomogane, Chenfei Ye, Ting Ma, Michael I Miller, and Susumu Mori. Multi-atlas based detection and localization (madl) for location-dependent quantification of white matter hyperintensities. *NeuroImage: Clinical*, 22:101772, 2019. 1, 2

[25] H Yamauchi, H Fukuda, and C Oyanagi. Significance of white matter high intensity lesions as a predictor of stroke from arteriolosclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 72(5):576–582, 2002. 1

[26] Kai Zheng, Zheng Wang, Xi Chen, Jiajie Chen, Yu Fu, and Qin Chen. Analysis of risk factors for white matter hyperintensity in older adults without stroke. *Brain Sciences*, 13(5): 835, 2023. 1

Title: Use of ChatGPT-4o in Pediatric Surgery and Pediatric Urology: A Narrative Review

Authors: Trachta Jan, Poš Lucie

Department of Paediatric Surgery, Second Faculty of Medicine, Charles University and Motol University Hospital

Abstract

The rapid development of large language models (LLMs) such as ChatGPT-4o has opened up new opportunities for the integration of artificial intelligence (AI) into pediatric surgical and urological care. As a multimodal, conversational AI tool, ChatGPT-4o can generate human-like text, interpret images, and communicate in real time. This review article discusses the potential clinical, educational, and administrative applications of ChatGPT-4o in pediatric surgery and pediatric urology, while addressing its limitations, ethical concerns, and future possibilities for responsible implementation.

In a clinical setting, ChatGPT-4o is able to help pediatric surgeons and pediatric urologists in triaging patients at the reception of the Accident and Emergency Department and serve as a tool for brainstorming of preliminary diagnostic options and treatment planning. It can provide structured summaries of differential diagnoses based on described symptoms in combination with laboratory results and descriptions of imaging performed in out-patient clinics. Although ChatGPT-4o does not have formal regulatory approval for clinical decision making, its use as a decision support tool may be valuable in everyday practice and particularly in resource-limited settings or during consultations outside of regular working hours.

In addition to assisting in clinical diagnostics, ChatGPT-4o can be very useful in automating time-consuming documentation tasks. It could generate draft pre-operative instructions, admission and discharge reports based on brief input from the physician or edit and clean up the text of dictated operative protocols. Integration into the electronic health record (EHR) could reduce administrative burden, improve workflow efficiency, and increase the consistency of the clinical record. However, all AI-generated documentation must be reviewed and verified by licensed professionals to ensure accuracy and compliance.

Another promising area is communication with patients and families. Pediatric patients and their parents often need simplified explanations of their conditions and planned diagnostic and treatment procedures that are age-appropriate and lay-understanding, which ChatGPT-4o can provide upon request. This is particularly useful in pre-operative counseling or in detailed explaining of an informed consent. ChatGPT-4o can also serve as a support tool in chatbot-based platforms, answering frequently asked questions and helping to reduce anxiety for caregivers navigating complex diagnostic and treatment plans.

In the field of education, ChatGPT-4o can serve as an interactive learning program for medical students, residents and interns in pediatric surgery and urology. It can create customized quizzes, explain surgical procedures at different levels of complexity, simulate patient interviews, and help trainees practice clinical reasoning. The model's adaptability and ability for instant feedback offer a new way of personalized learning and skill reinforcement.

Despite its broad potential, ChatGPT-4o has several limitations and risks. These include factual inaccuracies ("hallucinations") or unreliability in reading imaging as X ray or CT scans (and so far, must therefore be replaced by other AI software integrated into radiology viewers.) A major obstacle to practical implementation is non-compliance with GDPR data protection laws. There are also

unresolved ethical issues regarding potential biases arising from ChatGPT training, its transparency, accountability or authorship when used as support in research papers and presentations. Therefore, caution should be exercised in its clinical deployment, supported by institutional guidelines, user education, and strict oversight.

In conclusion, ChatGPT-4o holds significant promise for improving the efficiency, accessibility, and clarity of pediatric surgical and urologic care. When used responsibly and ethically, it can stimulate or supplement - but not replace - human expertise. Further research, policy development, and model refinement are necessary for its safe and meaningful incorporation into pediatric practice.

Disclaimer

The ChatGPT language model (paid version GPT-4o Plus, OpenAI, 2025) was used as an auxiliary tool in the preparation of this manuscript. The model was used for linguistic editing and the creation of drafts of some paragraphs. All content was subsequently reviewed, edited, and supplemented by the authors.

Introduction

The rapid development of large language models (LLMs) such as ChatGPT and similar chatbots is very likely to represent a breakthrough moment in the application of artificial intelligence (AI) in many fields of modern medicine, including pediatric surgery and urology. ChatGPT-4o is a multimodal model capable of generating coherent text based on context and previous prompts, processing simple visual data, and conducting near-human-perfect conversations in real time.(1–3) This makes it a versatile support tool for clinical practice, especially when we consider the enormous theoretical background and amount of data that a chatbot is capable of processing compared to the human brain.(4)

In pediatric surgery, where communication, decision-making, and accuracy are critically important, LLMs have the potential to support physicians in triage or in generating differential diagnoses through interactive brainstorming and proposing investigative and therapeutic algorithms.(5–7) They can significantly reduce the administrative burden on doctors in hospitals by generating admission and discharge reports based on electronic and continuously updated outpatient records, available online to both general practitioners and all specialists who encounter the patient.(8–11) Chatbots can also be used to write and edit surgical protocols, as well as to analyze them in predicting possible postoperative complications or correlations in groups of patients operated on for the same diagnosis.(11–13) ChatGPT-4o can facilitate communication between parents and doctors in emotionally charged situations.(14,15) ChatGPT-4o can improve medical education by acting as an interactive tutor for interns or simulating clinical scenarios.(10,16–19)

However, there are still unresolved challenges in its real-world application, such as the security of the data entered, chatbot hallucinations, and the need for constant human supervision to ensure patient safety.(11,20–23) This article explores the scope and limitations of ChatGPT-4o's use in pediatric surgery and urology. For a more effective, safer, and ethically acceptable use of GPT-4o or similar chatbots soon, multidisciplinary collaboration between IT specialists, doctors, nurses, hospital managers, and even politicians at the European and national levels will be necessary in hospitals and outpatient practices.

Results

If we delve into the details of the potential use of ChatGPT-4o or its future versions when a patient is admitted to a hospital, its use in patient triage comes to mind. Real-world use is limited by the lack of personal data protection and the provision of data on the health of child patients to a chatbot, where we cannot be sure that companies such as Open AI will not use this data for their own benefit and whether they can protect it from misuse. For research purposes, the input data can, of course, be anonymized. If data protection could be set up in the hospital version of ChatGPT-4o, for example through automatic patient anonymization, the chatbot could help frontline staff evaluate reported symptoms and prioritize patients using structured input data classified according to evidence-based medicine and guidelines.(5–7) For example, in the case of a child with acute abdominal pain, the chatbot could suggest probable diagnoses such as constipation, acute enteritis, lymphadenitis, appendicitis, intussusception, or even testicular torsion with pain radiating to the groin. Several such pilot studies have already been conducted in adult patients, and some of them have shown better triage results than doctors specializing in emergency medicine, while others have not.(24–26) Even after the chatbot is integrated into triage, however, the patient will still ultimately have to be examined by a doctor, who will continue to bear all law liability for the examination and treatment of the child patient. On top of that the current prevailing belief that no machine with superhuman intelligence can fully replace an empathetic flesh-and-blood professional in normal interpersonal contact will not change for a long time, if ever.(23)

In addition to triaging patients, ChatGPT-4o can significantly contribute to diagnostic brainstorming by generating structured lists of differential diagnoses. It can synthesize loosely formulated clinical notes, laboratory findings, and outpatient ultrasound, X-ray or CT reports into coherent summaries that support surgical decision-making.(10,11) This ability can help young doctors in outpatient clinics or during night shifts, or in resource-limited settings where older and more experienced specialists may not be available, such as in many regional hospitals in sub-Saharan Africa and other low- and middle-income countries.(27) Of course, despite its potential, ChatGPT-4o remains only a supportive tool in decision-making, complementing rather than replacing the physician's judgment.

In addition to its role in diagnostic support, ChatGPT-4o offers considerable potential in addressing one of the most persistent inefficiencies in modern healthcare: the administrative burden associated with clinical documentation. Pediatric surgery departments, like most modern hospital departments, are burdened with the need to produce a large volume of medical reports, including preoperative examinations, informed consent forms, admission and discharge reports, and surgical protocols. ChatGPT-4o can help clinicians by generating structured drafts of these documents out of short inputs from the physician, such as voice dictations, bullet points, or electronic health record (EHR) entries.(8) The challenge remains not only GDPR and data protection, but also the incorporation of the chatbot into complex hospital software, which is different for each hospital and on which the entire administration relies. However, the chatbot can already be used for individual sub-steps and the result can be copied into hospital software. For example, a pediatric urologist can dictate only the key intraoperative steps during hypospadias surgery and ask ChatGPT-4 to transcribe them into a standardized surgical protocol suitable for medical documentation.(9)

Integrating the chatbot into the hospital's electronic health records (EHR) could streamline workflows by automatically filling in patient demographic data upon admission or entering laboratory values, other specialists' examinations, and descriptions of imaging methods performed into template documents upon admission and discharge.(9) Preoperative instructions for parents or discharge reports with further follow-up and care plans could ultimately be automatically summarized in plain language that is understandable to the family without the need to google what each abbreviation and term in the discharge report means. In addition, such an understandable summary can be generated in a report in any native language of the patient and their parents if the family does not have a good command of the language of the country in which they receive healthcare. Such use of a chatbot could not only save time but also reduce variability among care providers, ensuring greater clarity and completeness of documentation.(10,11)

However, the use of LLM in clinical documentation raises ethical and legal issues. Errors in summaries or inappropriate wording—such as omitting certain diagnostic and treatment steps or important laboratory values—pose a potential risk to patient safety and regulatory compliance.(21) Therefore, every document generated by artificial intelligence must undergo review, verification, and final approval by an authorized healthcare provider. In a pediatric setting, where documentation accuracy is critical for medical-legal clarity and communication with families, the stakes are particularly high.

A particularly promising use of ChatGPT-4o in pediatric surgical and urological practice is to improve communication with patients and families. Effective communication is critically important in pediatrics, where emotional vulnerability is high and caregivers are often overwhelmed by complex medical terminology and treatment decisions. ChatGPT-4o can support physicians by generating age-appropriate, layman-friendly explanations of diseases, diagnostic procedures, and surgical interventions.(11) For example, during an outpatient consultation with a family about vesicoureteral reflux or bladder augmentation surgery, the model can translate technical information into simplified language tailored to the developmental level of the child and their caregivers.

In a preoperative setting, ChatGPT-4o can help reduce parental anxiety by providing consistent and accessible answers to frequently asked questions (e.g., "Will my child feel pain after surgery?" or "What are the risks of anesthesia?").(14,28) When integrated into chatbot-based interfaces, it can function as a 24/7 digital assistant that reinforces key messages provided during clinical encounters. This can be particularly valuable for families undergoing multi-phase treatment, such as sequential surgeries or long-term follow-up. In a study by Ayers et al., it was even found that a chatbot in online forums dedicated to patient queries was more empathetic and understandable than real doctors.(29)

In addition, ChatGPT-4o can assist in the preparation of informed consent documents and offer standardized proposals that explain risks, benefits, and alternatives in a format that is understandable to patients. Such a feature supports both legal compliance and ethical standards of informed decision-making, especially in sensitive or risky procedures. Although AI cannot replace human empathy or nuanced interpretations of emotional cues, it can serve as a complement that expands the availability of medical information and reinforces understanding through repetition in understandable language. Quality control, culturally sensitive wording, and real-time clinical oversight remain essential for safe implementation.

In the field of medical education, ChatGPT-4o represents a significant advance in providing personalized interactive training for medical students, residents, and young physicians in the fields of pediatric surgery and urology. Traditional educational methods often fail to meet the diverse educational needs and schedules of surgical trainees. ChatGPT-4o, on the other hand, offers an adaptive, on-demand learning environment that can tailor content to the individual knowledge level and goals of each student.(11,16,17) It can converse on a technical topic or generate customized quizzes, from the basics of anatomy and embryology to advanced questions based on cases of congenital anomalies such as posterior urethral valves or bladder exstrophy, helping students consolidate factual knowledge and identify gaps in their knowledge.

In addition to teaching, the model can simulate virtual encounters with patients, guide interns through structured interviews, and encourage them to ask appropriate diagnostic questions based on the symptoms presented. For example, it can simulate a consultation with the parent of a child with undescended testicles and encourage the student to gather relevant medical history, propose a differential diagnosis, and formulate an initial treatment plan. It can also describe surgical procedures step by step, such as laparoscopic pyeloplasty, at varying levels of complexity depending on the user's background—whether they are a medical student looking for basic guidance or a resident who needs detailed information about the operation and what to look out for during surgery.

The platform's ability to provide immediate feedback and explanations supports a formative approach to learning and enables continuous reflection and correction of misconceptions. It can also be used to prepare for exams or to independently review procedures prior to live operations. This flexibility is a powerful complement to traditional bedside teaching and operating room practice, especially in programs where the number of surgical cases may be limited. When the content is verified and used under expert supervision, ChatGPT-4o can help standardize teaching and accelerate skill acquisition in pediatric surgical education.(17)

Despite its remarkable versatility and growing role in healthcare, the deployment of ChatGPT-4o in pediatric surgery and urology should be approached with caution given its current technical, ethical, and regulatory limitations. One significant concern is the phenomenon of AI hallucinations, in which the model generates incorrect or fabricated information while maintaining a confident tone. This can be particularly dangerous in a clinical setting, where factual accuracy is critical for diagnosis and treatment planning.(11)

Furthermore, while ChatGPT-4o can interpret some text-based radiology reports, it is not yet capable of reliably analyzing medical images such as X-rays, ultrasounds, or CT scans. Specialized AI models trained on large, annotated image datasets and integrated directly into radiology viewers (e.g., Aidoc, Arterys) and approved for clinical use by the FDA (Food and Drug Administration) or CE (Conformité Européenne). In practice, ChatGPT-4o should only be used in conjunction with expert radiological input, not in place of it.

A significant barrier to real-world deployment is insufficient compliance with the GDPR. Current data processing protocols for many LLMs, especially cloud-based ones, may not meet European privacy and consent requirements, posing legal and ethical barriers in clinical settings. In addition, concerns remain unresolved regarding algorithmic bias arising from the model's training data. These biases can lead to persistent inequalities in care if left unchecked, particularly in pediatric populations with diverse ethnic, linguistic, or socioeconomic backgrounds.(21)

Other ethical issues relate to transparency, accountability, and authorship, particularly in academic settings where AI-assisted writing is becoming increasingly common.(17,21,23,30) Institutions must define clear policies on how content generated by ChatGPT should be published and acknowledged. To ensure the safe and ethical use of ChatGPT in academic work, leading publishing houses and journals already have a clearly defined and permitted role for ChatGPT. Where ChatGPT fails almost completely, when used as an aid in formulating certain ideas or sentences, is in finding relevant and current sources and citing them. ChatGPT almost always invents references for scientific work, and they cannot be used at all. We therefore believe that the role of ChatGPT in writing professional articles and books is again only as a support tool in brainstorming ideas and providing inspiration as to which part of the studied topic can still be explored after finding relevant and current sources.

Conclusion

It can be said that ChatGPT-4o represents a significant advance in the integration of artificial intelligence into pediatric surgical and urological care, but most doctors do not yet use it due to mistrust and unfamiliarity, a certain unpredictability, the possibility of hallucinations, and a lack of personal data protection. However, its potential to increase the efficiency of workflows, improve access to information, and facilitate communication between patients and doctors is becoming increasingly apparent. From clinical documentation and diagnostic reasoning to patient education and surgical training, ChatGPT-4o offers a wide range of applications that can address logistical and cognitive challenges in pediatric healthcare.

However, its usefulness must be contextualized within the framework of ethical responsibility and clinical judgment. As a tool without perception, ChatGPT-4o cannot replace the nuanced, empathetic, and morally grounded decision-making of pediatric surgeons and urologists. Rather, it should be considered a complementary support system—a cognitive and administrative amplifier that can stimulate human thinking, reduce administrative burden, and promote standardization, especially in resource-constrained settings.

Interdisciplinary collaboration between clinicians, IT specialists, ethicists, and hospital and government policymakers is essential for meaningful and safe adoption. Current key priorities include improving model performance, ensuring transparency and interpretability, and developing institutional policies regarding authorship, consent, and privacy. In addition, robust validation studies are needed to quantify the clinical impact of ChatGPT-4o in pediatric contexts, including long-term outcomes and user satisfaction.

The integration of ChatGPT-4o into pediatric surgical practice ultimately reflects a broader paradigm shift toward AI-assisted medicine. With careful deployment, transparency, and regulatory oversight, this technology has the potential to improve the quality of care, reduce disparities, and enhance the human dimension of clinical work by freeing up time for what matters most: the child and their family.

Literature:

1. Miyake Y, Retrosi G, Keijzer R. Artificial intelligence and pediatric surgery: where are we? Pediatr Surg Int. 2025 Dec 1;41(1).
2. González R, Poenaru D, Woo R, Trappey AF, Carter S, Darcy D, et al. ChatGPT: What Every Pediatric Surgeon Should Know About Its Potential Uses and Pitfalls. Vol. 59, Journal of Pediatric Surgery. W.B. Saunders; 2024. p. 941–7.
3. Xiao D, Meyers P, Upperman JS, Robinson JR. Revolutionizing Healthcare with ChatGPT: An Early Exploration of an AI Language Model's Impact on Medicine at Large and its Role in Pediatric Surgery. Vol. 58, Journal of Pediatric Surgery. W.B. Saunders; 2023. p. 2410–5.
4. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med. 2023 Dec 1;6(1).
5. Kaboudi N, Firouzbakht S, Eftekhar MS, Fayazbakhsh F, Joharivarnoosfaderani N, Ghaderi S, et al. Diagnostic Accuracy of ChatGPT for Patients' Triage; a Systematic Review and Meta-Analysis. Arch Acad Emerg Med. 2024;12(1).
6. Colakca C, Ergın M, Ozensoy HS, Sener A, Guru S, Ozhasenekler A. Emergency department triaging using ChatGPT based on emergency severity index principles: a cross-sectional study. Sci Rep [Internet]. 2024 Sep 27;14(1):22106. Available from: http://www.ncbi.nlm.nih.gov/pubmed/39333599

7. , Masanneck L, Schmidt L, Seifert A, Kölsche T, Huntemann N, Jansen R, et al. Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study. J Med Internet Res. 2024;26(1).

8. Seth P, Carretas R, Rudzicz F. The Utility and Implications of Ambient Scribe in Primary Care (Preprint). JMIR AI. 2024 Oct 4;

9. Liu F, Zhou H, Wang K, Yu Y, Gao Y, Sun Z, et al. MetaGP: A generative foundation model integrating electronic health records and multimodal imaging for addressing unmet clinical needs. Cell Rep Med. 2025 Apr 15;6(4).

10. Miyake Y, Retrosi G, Keijzer R. Artificial intelligence and pediatric surgery: where are we? Pediatr Surg Int. 2025 Dec 1;41(1).

11. González R, Poenaru D, Woo R, Trappey AF, Carter S, Darcy D, et al. ChatGPT: What Every Pediatric Surgeon Should Know About Its Potential Uses and Pitfalls. Vol. 59, Journal of Pediatric Surgery. W.B. Saunders; 2024. p. 941–7.

12. Mao C, Bao Y, Yang Y, Cao Y. Application of ChatGPT in pediatric surgery: opportunities and challenges. Vol. 110, International journal of surgery (London, England). 2024. p. 2513–4.

13. McGee LM, Soo E, Seideman CA. Integration of novel artificial intelligence tools in pediatric urologic practice. Curr Opin Urol. 2025 May;35(3):230–5.

14. Zeltzer D, Herzog L, Pickman Y, Steuerman Y, Ber RI, Kugler Z, et al. Diagnostic Accuracy of Artificial Intelligence in Virtual Primary Care. Mayo Clinic Proceedings: Digital Health. 2023 Dec;1(4):480–9.

15. Chutia T, Baruah N. A review on emotion detection by using deep learning techniques. Artif Intell Rev. 2024 Aug 1;57(8).

16. Breeding T, Martinez B, Patel H, Nasef H, Arif H, Nakayama D, et al. The Utilization of ChatGPT in Reshaping Future Medical Education and Learning Perspectives: A Curse or a Blessing? American Surgeon. 2024 Apr 1;90(4):560–6.

17. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Medical Education. John Wiley and Sons Inc; 2024.

18. Gravina AG, Pellegrino R, Palladino G, Imperio G, Ventura A, Federico A. Charting new AI education in gastroenterology: Cross-sectional evaluation of ChatGPT and perplexity AI in medical residency exam. Digestive and Liver Disease. 2024 Aug 1;56(8):1304–11.

19. Yudovich MS, Makarova E, Hague CM, Raman JD. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. J Educ Eval Health Prof. 2024;21.

20. Robinson JR, Stey A, Schneider DF, Kothari AN, Lindeman B, Kaafarani HM, et al. Generative Artificial Intelligence in Academic Surgery: Ethical Implications and Transformative Potential. Journal of Surgical Research. Academic Press Inc.; 2025.

21. Jeyaraman M, Ramasubramanian S, Balaji S, Jeyaraman N, Nallakumarasamy A, Sharma S. ChatGPT in action: Harnessing artificial intelligence potential and addressing ethical challenges in medicine, education, and scientific research. World J Methodol. 2023 Sep 20;13(4):170–8.

22. , Resnik DB, Hosseini M. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. AI and Ethics. 2024 Apr 27;

23. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). Vol. 7, npj Digital Medicine. Nature Research; 2024.

24. Wang C, Wang F, Li S, Ren Q wen, Tan X, Fu Y, et al. Patient Triage and Guidance in Emergency Departments Using Large Language Models: Multimetric Study. J Med Internet Res. 2025 May 15;27:e71613.

25. Pasli S, Yadigaroğlu M, Kirimli EN, Beşer MF, Unutmaz İ, Ayhan AÖ, et al. ChatGPT-supported patient triage with voice commands in the emergency department: A prospective multicenter study. Am J Emerg Med. 2025 Aug;94:63–70.

26. Zaboli A, Brigo F, Brigiari G, Massar M, Parodi M, Pfeifer N, et al. Chat-GPT in triage: Still far from surpassing human expertise - An observational study. Am J Emerg Med. 2025 Jun;92:165–71.

27. Botelho F, Tshimula JM, Poenaru D. Leveraging ChatGPT to Democratize and Decolonize Global Surgery: Large Language Models for Small Healthcare Budgets. Vol. 47, World Journal of Surgery. Springer Science and Business Media Deutschland GmbH; 2023. p. 2626–7.

28. van Eerde AM, Teixeira A, Galletti F, Maternik M, Capone V, Westland R, et al. Risks and benefits of ChatGPT in informing patients and families with rare kidney diseases: an explorative assessment by the European Rare Kidney Disease Reference Network (ERKNet). Pediatric Nephrology. 2025;

29. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med. 2023 Jun 1;183(6):589.

30. Lee JM. Strategies for integrating ChatGPT and generative AI into clinical studies. Vol. 59, Blood Research. Springer; 2024.

2nd Sorbonne-Heidelberg Workshop on AI in Medicine

,

Multimodal AI approach for pain detection: combining facial gesture, vocal cues, and head movements and postures

Lina Arab[1], Asma Benkaci[2], Layth Sliman[2], and Hachemi Nabil Dellys[1]

[1]Higher National School of Computer Science (ESI ex INI), Algiers, Algeria
[2]Efrei Research Lab. Panthéon Assas University, 30-32 avenue de la République, Villejuif, Paris, France.

June 14, 2025

Abstract

Effective pain detection is crucial in clinical settings, especially for patients who cannot reliably self-report their pain. However, although automated pain-detection systems already exist, they typically rely on a *single* behavioural or physiological channel, even though pain can manifest through multiple modalities. We introduce a multimodal artificial intelligence (AI) framework that unifies facial-expression analysis, vocal-cue extraction, and head-pose dynamics. Each modality is trained on a *specific benchmark dataset* using dedicated deep-learning architectures: a Convolutional Neural Network (CNN) based on InceptionV3 combined with a Bidirectional Long Short-Term Memory (BiLSTM) network with an attention mechanism for facial-expression analysis, a seven-layer Multilayer Perceptron (MLP) driven by Mel-Frequency Cepstral Coefficient (MFCC) features for vocal-cue extraction, and a two-layer BiLSTM network for head-pose dynamics. Final pain-detection decisions are obtained via majority voting across modalities. The facial, audio, and head-pose models achieved accuracies of **85.5%**, **90.5%**, and **66.7%** with corresponding F1 scores of **0.90**, **0.906**, and **0.775**, respectively. When fused, the overall system reached an accuracy of **96.7%** and an F1 score of **0.929**. Our findings suggest that integrating facial expressions, speech characteristics, and head movements can significantly improve the detection of pain, particularly in situations where patients are unable to communicate clearly.

Keywords: Pain detection; multimodal fusion; facial-expression analysis; vocal cues; head-pose analysis; artificial intelligence; healthcare.

I INTRODUCTION

The world's population is rapidly aging, creating new challenges in healthcare for older adults. In 2019, about 1 billion people were aged 60 or above, a figure projected to reach 2.1 billion by 2050 [1]. A strong majority of seniors wish to "age in place" (remain in their own homes) rather than move to institutions [2], which increases the need for home-based health monitoring. Chronic pain is highly prevalent in this demographic, with over half of adults 65+ reporting ongoing pain and around one-third suffering chronic pain [3]. Managing pain in elders is difficult since many have cognitive impairments or communication barriers – for example, older adults with dementia often cannot reliably self-report pain, leading to under-diagnosis [4]. When pain goes unnoticed or untreated, it can seriously impact a person's well-being and may accelerate both physical and cognitive decline. This underscores the need for non-intrusive, automated, real-time pain monitoring systems to support aging in place.

Recent advances in artificial intelligence (AI) enable the development of such automated pain assessment tools. Unlike traditional nurse observations or self-reports, an AI-driven system can continuously and objectively detect pain behaviors without burdening the individual. Vision and audio modalities are especially promising for unobtrusive monitoring: cameras and microphones are passive, widely available sensors that can capture pain-related facial expressions, vocalizations, and body movements in real time [5]. Prior studies have shown that analyzing facial cues with AI can successfully indicate pain – for instance, a smartphone app has been used to detect pain from facial expressions in non-communicative patients [4]. Facial expressions are a well-established pain indicator: pain-related facial movements (e.g. brow lowering, eye tightening) carry specific information distinct from general expressions [6]. Similarly, vocal pain indicators such as moaning, crying, or changes in speech have been linked to painful episodes, and AI algorithms can mine vocal features for pain "biomarkers." Head pose and gesture changes (like guarding or restlessness) may further signal discomfort. A multimodal AI approach that fuses these cues is therefore a logical focus, as it can leverage complementary channels – e.g. detecting pain even when a patient is silent or has a stoic facial expression. In summary, there is a critical need for an automated pain monitoring system for older adults that integrates facial, vocal, and head movement cues to achieve non-invasive, real-time pain detection. This work aims to fulfill that need. Notably, most existing automated pain detection research has not been validated on older adult populations or those with dementia [7], highlighting the importance of our focus on an elderly monitoring context.

II RELATED WORK

A study by Lucey et al. introduced the UNBC-McMaster Shoulder Pain Expression Archive, establishing a benchmark for facial expression-based pain detection and enabling the development of numerous algorithms using the Facial Action Coding System (FACS) [8]. Werner et al. developed the BioVid Heat Pain database, which extended analysis beyond facial cues to include head pose and physiological signals [9, 10]. More recent works such as that by Fang et al. have documented the evolution from handcrafted features to deep learning methods like CNNs and spatio-temporal models, significantly boosting recognition performance [6]. However, facial-only systems face key limitations: pain expressions are highly individual, can be absent in stoic or cognitively impaired patients, and most datasets were collected from younger adults, limiting generalizability to elderly or dementia populations [7].

Vocal cues offer another perspective. Borna et al. reviewed AI-based voice analysis, showing that vocal features such as pitch and loudness can reliably indicate pain when analyzed with modern deep learning techniques [11]. Icht et al. and Lautenbacher et al. further demonstrated that pain is associated with characteristic vocalizations, such as moaning and changes in speech quality, which can be detected with signal processing methods [12, 13]. Yet, unimodal audio approaches struggle with background noise, speaker variability, and the lack of large, standardized datasets [11]. Most importantly, pain may not always be vocalized—especially in certain medical or cultural contexts—leaving these systems blind in such situations.

Head pose and body movement have recently been leveraged for pain detection. Werner et al. analyzed head orientation and observed systematic downward or averted movements during pain [14], while Walsh et al. found that "head averted" and "gaze downward" are prototypical pain postures [15]. Egede et al., through the EmoPain Challenge, showed that integrating body movement features with facial cues improves recognition of pain-related behaviors, particularly in chronic pain patients [16]. Nevertheless, relying solely on posture or movement can overlook cases where pain does not significantly alter body language.

These limitations have driven a shift toward multimodal fusion approaches. Thiam et al. demonstrated that combining facial, vocal, and physiological data using late fusion strategies produces more robust and reliable pain detection than any single modality alone [5]. Gutierrez et al. highlighted the benefits of integrating facial gesture and paralanguage analysis for real-world scenarios [17]. Borna et al. also found that multimodal systems can compensate for missing or ambiguous signals in one channel by leveraging complementary cues from others [11]. However, as Rezaei et al. pointed out, most automated pain detection models are still rarely validated on older adults or dementia patients, emphasizing the need for research focused on these vulnerable populations [7].

III Methods

In this work, we developed a multimodal pain detection framework that integrates three complementary approaches—facial expression analysis, vocal cue analysis, and head pose dynamics—to improve the accuracy and robustness of automated pain recognition in videos. For each video sample, all three individual models were independently applied: (1) a facial expression-based system leveraging deep spatiotemporal features, (2) an audio-based system analyzing speech characteristics, and (3) a head pose-based system modeling dynamic postural changes. Each model outputs a binary prediction (”Pain” or ”No Pain”).

The final pain assessment for each video was determined through a simple majority voting scheme: the decision with at least two out of three models in agreement was taken as the overall system prediction. This late fusion strategy helps compensate for potential errors or ambiguities in any single modality, leading to improved reliability in challenging real-world conditions.

The following subsections provide detailed descriptions of the three core components of our multimodal framework.

A , Facial Expression-Based Pain Detection

For facial expression-based pain detection, we utilized the Pain E-motion Faces Database (PEMF) [18], consisting of 272 high-resolution micro-video clips from 68 participants. This dataset includes spontaneous (algometer-induced, laser-induced) and posed pain expressions, alongside neutral expressions. Each clip was meticulously annotated by trained raters using the Facial Action Coding System (FACS), providing detailed annotations for pain intensity (0–8 scale), facial Action Units (e.g., AU4: brow lowering, AU6: cheek raiser), emotional valence, arousal ratings, and discrete emotion labels (e.g., happiness, sadness, anger). The overall pipeline for facial expression analysis comprises three main stages: preprocessing, feature extraction, and temporal modeling, as depicted in Figure 1.

Preprocessing In this initial stage, video frames underwent face detection and landmark-based alignment using Dlib's frontal face detector coupled with a 68-point landmark predictor. Faces were strictly cropped, aligned, resized to 224×224 pixels, and normalized to reduce variability due to lighting conditions. This step ensured consistency and stability of input images for subsequent feature extraction.

Feature Extraction Spatial features were extracted using the pretrained Inception-V3 convolutional neural network (CNN) [19], with all network weights frozen to leverage representations learned from the ImageNet dataset. Each aligned frame was processed through the Inception-V3 backbone, and the resulting feature maps were condensed by global average pooling and global max pooling, yielding robust 4096-dimensional vectors per frame.

Temporal Modeling The sequence of frame-level features was modeled to capture temporal dynamics crucial for detecting subtle pain-related behaviors. Temporal modeling involved applying one-dimensional convolution (Conv1D) layers to encode short-term dependencies, followed by stacked bidirectional Long Short-Term Memory (BiLSTM) layers [20] to capture longer temporal dependencies in both directions. A soft attention mechanism [21] was implemented to identify and emphasize the most informative frames indicative of pain, resulting in a context-aware feature representation suitable for both regression and classification tasks.

Training and Optimization For regression tasks (pain intensity estimation), the model was trained using mean squared error (MSE) loss, while for classification tasks (binary pain detection), binary cross-entropy loss was employed. Stratified data splits preserved intensity distribution, and data augmentation (e.g., random flips, rotations, zoom, and brightness adjustments) was applied during training to enhance robustness. Frame-level features were standardized using a StandardScaler, and sequences were padded to match the longest sequence. The model was optimized using the Adam optimizer, with early stopping, learning rate reduction upon plateau, and checkpointing based on validation metrics.

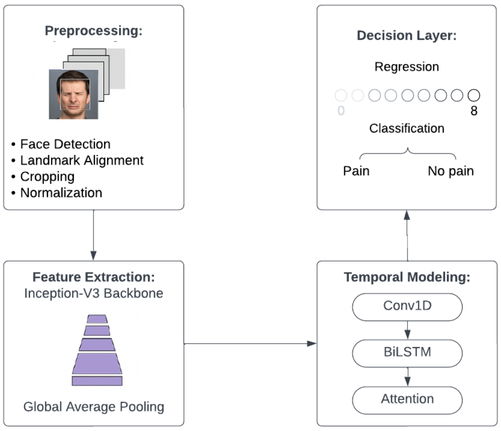


Figure 1: Flowchart illustrating the pipeline for facial expression-based pain detection

B Vocal Cues-Based Pain Detection

Dataset Overview The method employs the TAME-Pain dataset [22], which contains 7,039 speech utterances (311 minutes of audio) from 51 healthy adult participants. Each utterance is annotated with a self-reported pain score (1–10) and a binary pain label (Pain vs. No Pain), and is accompanied by an audio-quality rating on a 0–4 scale (0 = highest quality). Only recordings marked as valid and with the highest quality label (action label 0) were retained for analysis. All audio was captured at 16 kHz, 16-bit mono. An overview of this audio-based pain detection pipeline is shown in Figure 2.

Preprocessing and Feature Extraction Each retained audio recording was preprocessed to remove silence and extract spectral features. First, leading and trailing silences were trimmed using a voice-activity detector (VAD)[22]. Then, 20-dimensional Mel-frequency cepstral coefficients (MFCCs) were computed over short time frames of the trimmed speech. MFCCs are widely used acoustic features in speech analysis[23]. For each utterance, the frame-level MFCCs were summarized by computing the mean and standard deviation of each coefficient across time, yielding a 40-dimensional feature vector (20 means + 20 standard deviations).

Classification Model Prior to classification, feature vectors were z-score standardized across the dataset. To mitigate the imbalance between pain and no-pain classes, synthetic minority oversampling (SMOTE) was applied [24]. A deep feedforward neural network (multilayer perceptron) was then trained on the resulting features. The network architecture comprised seven hidden layers with 1,024→512→256→256→128→128→64 units, respectively. Each hidden layer used a ReLU activation, followed by batch normalization and dropout (dropout rates of 0.3–0.4) to promote generalization. The final layer was a single sigmoid unit for binary pain-vs-no-pain prediction. Binary cross-entropy loss was minimized using the Adam optimizer. Early stopping and a learning-rate scheduling scheme were employed as additional regularization during training.

Training Setup The dataset was split into training and test sets using an 80:20 stratified partition, preserving the pain/no-pain ratio. The model was trained on the 80% training subset (with hyperparameters tuned via cross-validation on training data), and evaluated on the held-out 20% test set.

C Head Pose-Based Pain Detection

The system is evaluated on the BioVid Heat Pain Database Part A[25], which contains recordings of 90 subjects under four levels of experimentally induced heat pain. For each video frame, head orientation angles (yaw, pitch, and roll) are estimated using Google's MediaPipe FaceMesh model[26]. From these angle trajectories, the angular velocity and acceleration are computed. These feature sequences (angles, velocities, accelerations) are normalized separately for each subject to reduce inter-subject variability.

The temporal features are fed into a bidirectional LSTM (BiLSTM) network for pain classification[20]. The BiLSTM processes the sequence in both forward and backward time directions, thereby capturing contextual dynamics of head motion. To address class imbalance, training uses the focal loss function[27], which down-weights easy (well-classified) examples in the loss, and the training set is balanced with SMOTE over-sampling. Model performance is assessed in a leave-one-subject-out cross-validation (LOSO-CV) scheme, i.e., each subject serves once as the held-out test set. Finally, the head-pose-based predictions are fused with those from other modalities (facial expression and audio) via majority voting. This late fusion leverages complementary signals to improve overall pain detection robustness. Figure 3 illustrates

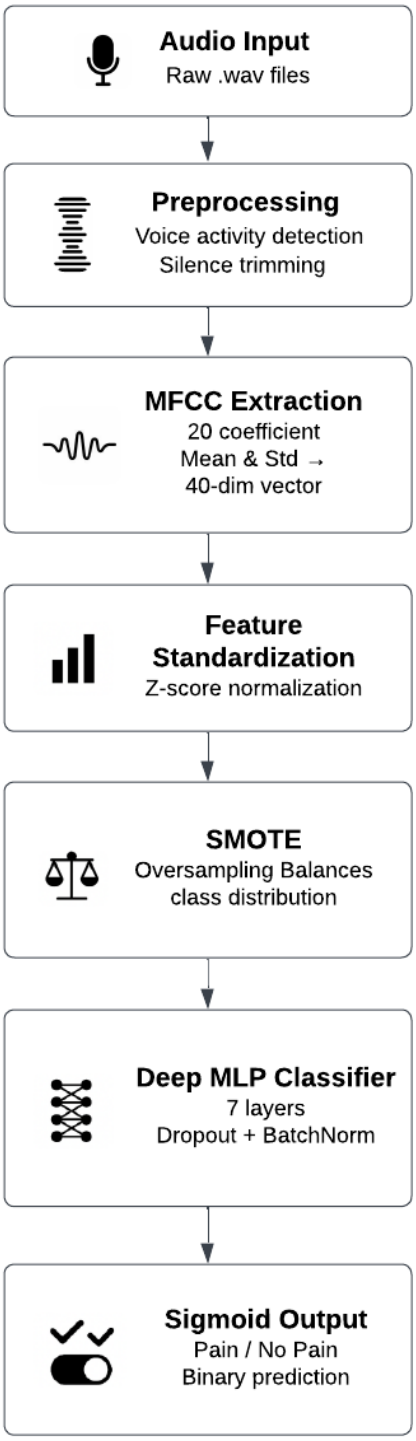


Figure 2: Flowchart illustrating the pipeline for audio-based pain detection

the processing pipeline for this method.

,

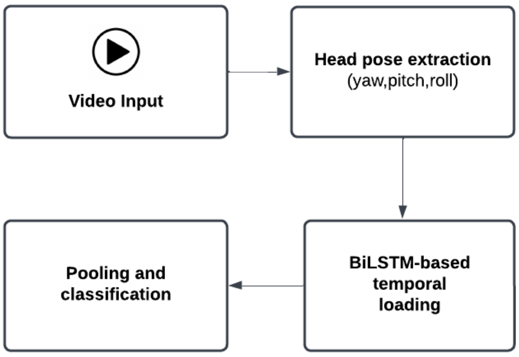

Video Input
Head pose extraction (yaw,pitch,roll)
BiLSTM-based temporal loading
Pooling and classification

Figure 3: Flowchart illustrating the pipeline for head pose-based pain detection

IV Results and discussion

In this section, we first present the performance of the three individual unimodal systems—facial expression-based, audio-based, and head pose-based pain detection—followed by the results of our primary contribution: the multimodal pain detection system that integrates these modalities. Pain expression is inherently diverse, varying significantly across individuals due to differences in personality, culture, gender, and age. As such, relying on a single modality often proves insufficient for consistent and robust pain recognition. To address this, our multimodal framework combines complementary information from facial expressions, vocal characteristics, and head pose dynamics, leveraging a late-fusion majority voting scheme.

A Facial-Based Pain Detection

We evaluated the performance of various facial expression-based models for both pain intensity regression and binary pain classification, using the PEMF database. Models were compared based on predictive accuracy, architectural design, and feature representations.

1) Regression Performance Our best-performing regression model combined **InceptionV3** (with frozen weights) as a frame-level feature extractor with a temporal modeling stack comprising **Conv1D**, **Bidirectional LSTMs**, and a **soft attention mechanism**. This network predicted pain intensity (on a 0–8 scale) with:

- **MSE**: 0.502
- **MAE**: 0.587
- **RMSE**: 0.708
- **R^2**: 0.780

These results indicate a strong fit between predicted and ground-truth pain levels. Training was performed using the **Adam optimizer** (learning rate: 1e-4), **batch size** of 8, and **early stopping** based on validation loss. Dropout (0.3) and learning-rate scheduling were used to improve generalization.

2) Comparative Architectures and Fusion Variants We evaluated several alternative models to benchmark performance (see Table 1). Notably:

- **InceptionResNetV2 (IRNV2)** backbones, whether frozen or partially fine-tuned (30–90 final layers), underperformed InceptionV3. The best IRNV2 configuration (30 unfrozen layers) yielded an MSE of 0.76.
- **AU-only** models used facial Action Unit features extracted via *OpenFace* or *py-feat*. These shallow regressors performed poorly (best MSE = 1.00), likely due to loss of spatiotemporal resolution.
- **PSPI-only** models used the Prkachin and Solomon Pain Intensity (PSPI) score—computed from selected AUs—as a scalar input. Despite its clinical relevance, this hand-crafted feature set yielded subpar results (MSE = 1.01).
- **Fusion models** that combined CNN features with AU or PSPI inputs provided moderate improvements. For example, a late-fusion model combining InceptionV3 with PSPI achieved an MSE of 0.60. However, none surpassed the InceptionV3-only model.

These findings reaffirm that **deep CNN-based spatiotemporal modeling** is more effective than handcrafted facial feature regressors in capturing subtle pain-related cues in video sequences.

3) Binary Classification Results The same feature extraction and sequence modeling pipeline was adapted for binary classification (pain vs. no pain). The final architecture included:

- Two **Bidirectional LSTM** layers
- A **soft attention layer**
- Fully connected layers with **ReLU activation** and **L2 regularization**

Training was class-balanced using inverse frequency weights, and monitored using validation accuracy. On the test set, this model achieved:

- **Accuracy**: 0.855

- **Precision**: 0.923
- **Recall**: 0.878
- **F1 Score**: 0.900

The strong F1 score and balanced precision/recall demonstrate the model's robustness across both spontaneous and posed pain expressions.

Table 1: Summary of Facial Expression-Based Pain Detection Models

Model Variant	MSE
InceptionV3 (frozen)	**0.50**
InceptionV3 + PSPI (late fusion)	0.60
InceptionResNetV2 (30 layers unfrozen)	0.76
IRNV2 + PSPI (late fusion)	0.70
AU-only (py-feat max)	1.00
PSPI-only	1.01

These results, summarized in Table 1 , show that a **frozen InceptionV3 backbone** paired with **temporal sequence modeling** and **attention mechanisms** yields the best performance for both pain intensity regression and binary pain detection. While hybrid models incorporating AU or PSPI features offer marginal improvements in some configurations, they are consistently outperformed by deep CNN-based models. These findings highlight the effectiveness of **deep spatiotemporal representations** in capturing facial expressions of pain, especially in elderly populations.

B Audio-Based Pain Detection

We evaluated multiple neural architectures for binary pain detection using vocal cues from the TAME-Pain dataset. The best-performing system was a deep feedforward neural network trained on MFCC-based features. Specifically, for each utterance, 20 Mel-frequency cepstral coefficients (MFCCs) were extracted and summarized via their mean and standard deviation, resulting in a 40-dimensional input vector. These were z-score normalized, and class imbalance was addressed using SMOTE.

The final architecture consisted of seven hidden layers (sizes: 1024 → 512 → 256 → 256 → 128 → 128 → 64), each followed by ReLU activation, batch normalization, and dropout (0.3–0.4). The model was trained using the Adam optimizer (initial learning rate = 0.001), with early stopping (patience = 10) and learning rate scheduling (ReduceLROnPlateau, patience = 5). A batch size of 16 and 200 maximum epochs were used. Cross-validation (5-fold) yielded strong and consistent results: **Accuracy = 0.905**, **Precision = 0.896**, **Recall = 0.916**, and **F1 = 0.906**, confirming excellent generalization.

Several alternative models were also tested:

- **Simpler MLPs** with 2–4 hidden layers (e.g., 128–64 or 256–128–64) achieved decent performance (Accuracy ≈ 0.849–0.860), but lacked the capacity of deeper models.
- A **sequence-aware LSTM classifier** using reshaped MFCC sequences captured temporal patterns, with moderate gains in Recall (0.847), but at a cost to Precision.
- A **pure Conv1D network** applied to MFCC sequences showed competitive results but fell short of the deep MLP in overall metrics.
- Additional experiments included augmenting features with **pitch** or applying **denoising filters**, though these showed minimal or negative impact on performance.
- Lastly, **pre-trained Wav2Vec2 embeddings** (1536-dimensional, summarized via mean and std) were tested with an MLP, but yielded subpar results, likely due to domain mismatch and lack of fine-tuning.

The final comparative metrics are summarized in Table 2.

Table 2: Summary of Audio-Based Pain Detection Models

Model Variant	Accuracy / F1
Deep MLP (7-layer, cross-validated)	0.905 / **0.906**
Deep MLP (no CV)	0.877 / 0.847
Medium MLP (4-layer)	0.860 / 0.824
MFCC + LSTM Layer	0.832 / 0.803
MFCC + CNN-BiLSTM	0.835 / 0.793
Mel-Spectrogram + CNN	0.598 / 0.004
Wav2Vec2 + MLP	0.594 / 0.557

These results clearly demonstrate the effectiveness of deep MLP architectures for pain detection from speech, particularly when combined with robust preprocessing, class balancing, and hyperparameter tuning.

C Head Pose-Based Pain Detection

We evaluated multiple automated pain detection methods leveraging head pose dynamics, using the BioVid Heat Pain dataset and a rigorous Leave-One-Subject-Out cross-validation (LOSO-CV) protocol to ensure robust and generalized performance.

Optimal Model (BiLSTM) As shown in Table 3, our most effective model utilized normalized head orientation features—specifically yaw, pitch, and roll angles—along with their corresponding velocities and accelerations. These temporal sequences were modeled using a two-layer Bidirectional Long Short-Term Memory (BiLSTM) network with 64 hidden units per layer, dropout of 0.25, and layer normalization. To address dataset imbalance, SMOTE oversampling and focal loss were applied. The model was trained using the Adam optimizer, and learning-rate scheduling (step decay every 6 epochs). This model achieved a mean macro-F1 score of 0.775 and an accuracy of 0.667.

Comparative Evaluations To benchmark the BiLSTM model, we assessed several alternative temporal architectures:

- **BiGRU-based models:** Slightly lower performance (F1 = 0.74).
- **Transformer encoders (2 layers, 4 heads):** Significantly lower performance (F1 = 0.61, Accuracy = 0.57).
- **Ensemble methods (LSTM + GRU + Transformer):** Minor improvement over BiLSTM (F1 = 0.75, Accuracy = 0.65).

Traditional and Shallow Learning Baselines To assess the standalone discriminative power of head pose features, we also evaluated several traditional methods:

- **1D CNN/TCN:** Limited performance (Accuracy = 0.51, F1 = 0.52).
- **Statistical summaries (mean/std) + Random Forest:** Modest improvement (Accuracy = 0.58, F1 = 0.59).
- **Optical Flow / Motion History Image + RF:** Performance near chance level (Accuracy = 0.51, F1 = 0.50).
- **Facial landmark-based statistics + RF:** Limited predictive power (Accuracy = 0.53, F1 = 0.51).

The results in Table 3 show that deep recurrent architectures—particularly BiLSTM—are well-suited for leveraging the temporal structure of head pose sequences. These models substantially outperform transformer-based alternatives and traditional learning baselines. However, the relatively modest performance of simpler models underscores the subtle and individualized nature of head movement as a pain indicator. While head pose provides valuable insights, its predictive power alone remains limited. Therefore, integrating it into a multimodal framework alongside facial expressions and vocal cues is essential for reliably capturing diverse and nuanced pain expressions.

Table 3: Summary of Head Pose-Based Pain Detection Models

Model Variant	**Accuracy / F1**
BiLSTM (Optimal, pose sequences)	0.667 / 0.775
Ensemble (LSTM + GRU + Transformer)	0.650 / 0.750
BiGRU (pose sequences)	0.625 / 0.740
Transformer Encoder (pose sequences)	0.570 / 0.610
Statistical Features + Random Forest	0.580 / 0.590
1D CNN / TCN (pose sequences)	0.510 / 0.519
Facial Landmarks + Random Forest	0.532 / 0.510
Optical Flow + MHI + Random Forest	0.509 / 0.497

D Multimodal Pain Detection System Performance

The multimodal system processes each input video independently through three dedicated pipelines:

- **Facial expression analysis** using a CNN-BiLSTM attention-based model.
- **Vocal cue analysis** leveraging MFCC-driven deep neural networks.
- **Head pose analysis** modeled through a BiLSTM trained on temporal sequences of head orientation.

Each modality outputs a binary prediction (*Pain* or *No Pain*), and a final decision is made through a majority voting scheme, enhancing robustness to noise or uncertainty in individual modalities.

We evaluated the multimodal system using a custom dataset we created, consisting of 30 multicultural video samples with synchronized facial, audio, and head-pose data recorded from the same individuals. This dataset was constructed specifically for this study, as no publicly available resource currently provides all three modalities annotated for pain in a unified setting. The included subjects vary in age, gender, and ethnicity, reflecting real-world variability. Each input was independently processed through the three pipelines, and the final prediction was determined via majority voting across modalities. The multimodal approach achieved the following performance:

- **Accuracy:** 0.967 (96.7%)

,

- **Precision:** 1.000 (100%)
- **Recall:** 0.929 (92.9%)

Out of 30 samples, 29 were correctly classified, with only one false negative. These results surpass the performance of the individual unimodal systems on the same dataset. Notably, inspection of the prediction results showed that the multimodal model tended to output the correct label when at least two modalities predicted pain—even in cases where the third modality disagreed. For instance, some false positives from the facial stream were correctly counteracted by the head pose or audio predictions.

The strength of the multimodal system lies in its ability to capture complementary signals: while some individuals express pain more vocally, others may show it through facial tension or subtle head movements. By integrating multiple sources of evidence, the system increases both reliability and generalizability of pain recognition across diverse populations and contexts.

V Conclusion and Future Work

This research developed and validated a comprehensive multimodal AI system for automated pain detection, effectively leveraging facial expressions, vocal characteristics, and head movement dynamics. Through rigorous experimentation, the multimodal fusion approach demonstrated superior performance compared to unimodal models, adeptly handling variability and ambiguity inherent in pain expression.

Future research could explore more integrated fusion techniques that better capture how different pain signals—like facial cues and vocal patterns—interact in real time, such as early fusion techniques, attention mechanisms, or transformer-based cross-modal learning. To facilitate these efforts, utilization of synchronized, multimodal datasets such as SenseEmotion [28] and I-XTE Pain Dataset [29], once they become publicly available, is recommended. These datasets will provide richer contextual information, enabling more nuanced pain assessments. Furthermore, targeted validation studies with elderly and cognitively impaired populations should be conducted to better align the system's capabilities with real-world clinical needs, ultimately enhancing patient care and pain management strategies.

References

[1] Ageing. https://www.who.int/health-topics/ageing. Accessed on May 27, 2025.

[2] The value of aging in place. USC Leonard Davis School of Gerontology, https://gero.usc.edu/about/our-field/the-value-of-aging-in-place/. Accessed on May 27, 2025.

[3] How does chronic pain affect older adults' mental health? june 3, 2024, https://www.ncoa.org/article/exploring-the-link-between-chronic-pain-and-mental-health-in-older-adults/.

[4] K. Brett and M. Severn. Facial analysis technology for pain detection: A potentially useful tool for people living with dementia. *Canadian Journal of Health Technologies*, 3(7), 2023.

[5] P. Thiam, H. Hihn, D. A. Braun, H. A. Kestler, and F. Schwenker. Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. *Frontiers in Physiology*, 12:720464, 2021.

[6] R. Fang, E. Hosseini, R. Zhang, C. Fang, S. Rafatirad, and H. Homayoun. Survey on pain detection using machine learning models: Narrative review. *JMIR AI*, 4:e53026, February 2025.

[7] S. Rezaei, A. Moturu, S. Zhao, K. M. Prkachin, T. Hadjistavropoulos, and B. Taati. Unobtrusive pain monitoring in older adults with dementia using pairwise and contrastive training. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1450–1462, May 2021.

[8] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The UNBC–McMaster shoulder pain expression archive database. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG 2011)*, pages 57–64. IEEE, 2011.

[9] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. Traue. Towards pain monitoring: Facial expression, head pose, a new database, an automatic system and remaining challenges. In *Proceedings of the British Machine Vision Conference*, pages 119.1–119.11. British Machine Vision Association, 2013.

[10] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. C. Traue, S. Crawcour, P. Werner, A. Al -Hamadi, and A. O. Andrade. The BioVid

heat pain database: Data for the advancement and systematic validation of an automated pain recognition system. In *Proceedings of the 2013 IEEE International Conference on Cybernetics (CYBCONF)*, pages 128–131. IEEE, 2013.

[11] S. Borna, C. R. Haider, K. C. Maita, R. A. Torres, F. R. Avila, J. P. Garcia, G. D. De Sario Velasquez, C. J. McLeod, C. J. Bruce, R. E. Carter, and A. J. Forte. A review of voice-based pain detection in adults using artificial intelligence. *Bioengineering*, 10(4):500, 2023.

[12] M. Icht, H. Wiznitser Ressis-tal, and M. Lotan. Can the vocal expression of intellectually disabled individuals be used as a pain indicator? initial findings supporting a possible novice assessment method. *Frontiers in Psychology*, 12:655202, 2021.

[13] S. Lautenbacher, M. Salinas– Ranneberg, O. Niebuhr, and M. Kunz. Phonetic characteristics of vocalizations during pain. *Pain Reports*, 2(3):e597, 2017.

[14] P. Werner, A. Al-Hamadi, K. Limbrecht-Ecklundt, S. Walter, and H. C. Traue. Head movements and postures as pain behavior. *PLOS ONE*, 13(2):e0192767, 2018.

[15] J. Walsh, C. Eccleston, and E. Keogh. Pain communication through body posture: The development and validation of a stimulus set. *Pain*, 155(11):2282–2290, 2014.

[16] J. O. Egede, S. Song, T. A. Olugbade, C. Wang, A. C. D. C. Williams, H. Meng, M. S. H. Aung, N. D. Lane, M. Valstar, and N. Bianchi-Berthouze. EMOPAIN challenge 2020: Multimodal pain evaluation from facial and bodily expressions. In *Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 849–856. IEEE, 2020.

[17] R. Gutierrez, J. Garcia-Ortiz, and W. Villegas-Ch. Multimodal AI techniques for pain detection: Integrating facial gesture and paralanguage analysis. *Frontiers in Computer Science*, 6:1424935, 2024.

[18] R. Fernandes-Magalhães, A. Carpio, D. Ferrera, D. Van Ryckeghem, I. Peláez, P. Barjola, M. E. De Lahoz, M. C. Martín-Buro, J. A. Hinojosa, S. Van Damme, L. Carretié, and F. Mercado. Pain E-motion faces database (PEMF): Pain-related micro-clips for emotion research. *Behavior Research Methods*, 55(7):3831–3844, 2022.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. *arXiv*, 2015. arXiv:1512.00567 v3.

[20] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings of the 2005 IEEE International Joint Conference on Neural Networks (IJCNN)*, volume 4, pages 2047–2052. IEEE, 2005.

[21] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014. arXiv:1409.0473 v7.

[22] T.-Q. Dao, E. Schneiders, J. Williams, J. R. Bautista, T. Seabrooke, G. Vigneswaran, R. Kolpekwar, R. Vashistha, and A. Farahi. TAME Pain: Trustworthy assessment of pain from speech and audio for the empowerment of patients (version 1.0.0) [dataset]. PhysioNet, 2025.

[23] B. Tracey, D. Volfson, J. Glass, R. Haulcy, M. Kostrzebski, J. Adams, T. Kangarloo, A. Brodtmann, E. R. Dorsey, and A. Vogel. Towards interpretable speech biomarkers: Exploring MFCCs. *Scientific Reports*, 13(1):22787, 2023.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[25] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. C. Traue. Automatic pain recognition from video and biomedical signals. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014)*. IEEE, 2014.

[26] Y. Hammadi, F. Grondin, F. Ferland, and K. Lebel. Evaluation of various state-of-the-art head pose estimation algorithms for clinical scenarios. *Sensors*, 22(18):6850, 2022.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.

[28] M. Velana, S. Gruss, G. Layher, P. Thiam, Y. Zhang, D. Schork, V. Kessler, S. Meudt, H. Neumann, J. Kim, F. Schwenker, E. André, H. C. Traue, and S. Walter. The SenseEmotion database: A multimodal database for the development and systematic validation of an automatic pain- and emotion-recognition system. In *Affective*

Computing and Intelligent Interaction, Lecture Notes in Computer Science, pages 127–139. Springer, 2017.

[29] P. Werner, A. Al-Hamadi, S. Gruss, and S. Walter. Twofold-multimodal pain recognition with the X-ITE pain database. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW 2019)*, pages 290–296. IEEE, 2019.

A Systematic and automated Meta-Analysis Pipeline applied to microbiota and Allergy

Jeremy Corriger [1], Vladislav Antipin [1], Martin Larsen [1]

[1] Sorbonne Université, Inserm UMR-S1135, Centre d'Immunologie et des Maladies Infectieuses (CIMI-Paris), 75013, Paris, France.

Introduction

Systematic reviews (SRs) and meta-analyses (MAs) aim to synthesize existing evidence in response to well-defined research questions. When conducted rigorously, they represent the highest level of scientific evidence and are widely used to support clinical guidelines and health policies. Standardized frameworks such as Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (1) and Population, Intervention/Exposure, Comparison, Outcome ± Timing/Setting (PICO/PECO) (2) provide structured methodologies to ensure transparency and reproducibility. Still, the validity and generalizability strongly depend on the consistency and quality of the included studies.

In practice, the variability across studies – particularly in study design, experimental protocols, data collection procedures, and statistical methods – can compromise the comparability of results. This heterogeneity limits most MAs to an aggregation of reported effect sizes. Yet, a deeper level of integration – including the reanalysis of underlying data through harmonized statistical or bioinformatic pipelines – would enable more robust cross-study comparisons and help mitigate the risk of compounded bias.

The initial steps of SRs, including study selection and data extraction, are particularly critical for MAs. Recently, artificial intelligence (AI) has emerged as a powerful tool offering promising opportunities to streamline these tasks (3). Machine learning approaches such as natural language processing (NLP) and supervised classification models have been proposed to automate study screening, keyword extraction, and relevance assessment. Large language models (LLMs) further extend these possibilities, although their integration into SR and MA workflows remains an ongoing challenge (4).

Here, we present our approach to streamlining study selection, which is particularly critical in fields such as microbiota research, due to the heterogeneity in experimental and statistical analysis. We developed a structured, top-down pipeline to identify relevant studies and optimize article screening for MA in the context of early-life gut microbiota and allergy. Our aim is to overcome the high technical variability (e.g., sample processing, sequencing protocols, bioinformatic pipeline) by relying on robust and standardized procedures for literature selection and metadata extraction, in line with existing reporting frameworks such as STandardized Operation Procedures for Reproducible Microbiome Science (STORMS) (5).

Methods

Literature Retrieval and Preprocessing

To identify relevant literature on the gut microbiota and allergy in early life, we first explored 18 PubMed queries of varying stringency (returned between 39 and 2863 articles). We retained a mid-to-low stringency query as the most suitable compromise: *((allergy) AND (infant* OR newborn* OR neonate* OR child*)) AND ((microbio* OR *flora) AND (human) AND (gut OR *intestin* OR digestive OR enteric OR colic OR gastric)) NOT (Review[pt])*. This query returned 933 articles after duplicate removal.

For each article, the title and abstract were concatenated and processed through a text preprocessing pipeline including tokenization, lowercasing, stop-word removal, and punctuation stripping. Word frequency tables were then computed across the corpus.

To filter out generic or context-irrelevant terms, the same pipeline was applied to a set of 1494 out-of-scope articles retrieved from JAMA (published in 2024), using a query that excluded articles with any reference to allergy or microbiota: *("2024/01/01"[Date - Publication] : "2024/12/31"[Date - Publication]) AND (english[Language]) AND ("JAMA"[Journal]) NOT (allergy) NOT (microbiota) NOT (allergy[MeSH Terms]) NOT (microbiota[MeSH Terms])*. After testing several threshold combinations, we excluded words that appeared in fewer than 5% of relevant articles or in more than 0.5% of out-of-scope articles. This filtering strategy helped isolate discriminative and domain-specific keywords while minimizing the noise and computational resource usage.

Dataset Sampling and Annotation

All 933 articles were manually annotated for inclusion in the MA based on predefined relevance criteria. From this labeled corpus, we generated 4 random subsamples of increasing sizes (60, 100, 200, and 300 articles) to serve as training sets. For each, the remaining articles were used as test sets. This design allowed us to evaluate model performance across varying amounts of annotated data.

Feature Selection

To reduce dimensionality and retain only informative features, 2 complementary feature selection strategies were applied independently to each training set: Random Forest (RF) based importance ranking, and sparse Partial Least Squares Discriminant Analysis (sPLS-DA) using stability selection. Only features with a stability score ≥ 0.75 across bootstrapped samples were retained for modeling.

Model Fitting and Evaluation

Each resulting feature set was then used to train both a Generalized Linear Model (GLM) and a Partial Least Squares Discriminant Analysis (PLS-DA) classifier, resulting in 16 model configurations in total (4 sample sizes x 2 feature selection methods x 2 classifiers). Models were trained with default hyperparameters and without internal cross-validation, in order to evaluate their performance directly on the corresponding held-out test set. The outcome was a binary classification indicating article relevance for inclusion in the MA.

Model performance was assessed using multiple metrics: Area Under the Receiver Operating Characteristic Curve (ROC AUC), Recall, Precision, F1-score, and balanced Accuracy (i.e., the average of Sensitivity and Specificity), to account for the inherent class imbalance in the dataset. Classification thresholds were selected to maximize the F1-score on the training set, aiming to balance Sensitivity and Precision. Results were compared across modeling strategies and training sizes to evaluate the relative contribution of each component in the pipeline.

,

Results

Feature Selection Outcomes

The vocabulary filtering and feature selection steps yielded increasingly rich and specific sets of discriminative words as the trainset size increased. Random Forest (RF) and sparse PLS-DA (sPLS-DA) identified both overlapping and method-specific features. Commonly selected features across models included *gut*, *microbiota*, *allergy*, *infancy* and *intestinal*, reflecting the core topics of the research question. With larger training sets, more granular or technical terms emerged – such as *16S*, *rRNA*, *microbial*, and *stool* – particularly in sPLS-DA-based models (Figure 1).

Classification Performance

Classification performance improved consistently with increasing training set size, although gains were not strictly linear. Models trained on small samples (e.g., n ≤ 100) showed limited generalization, with Recall values and F1-scores below 0.30 for testset. These configurations also tended to overfit, as indicated by substantially higher performance on trainsets compared to testsets. In contrast, models trained on ≥ 200 articles showed marked improvements, with ROC-AUC values up to 0.80 and balanced accuracy up to 0.72 (Figure 2). Among these, the PLS-DA model trained on the sample of 300 articles using RF-selected features achieved the highest global performance, with the most balanced trade-off between Precision and Recall, the best test F1-score, and the smallest gap between train and test performance. The ROC curve for this model is shown in Figure 3.

Across all configurations, PLS-DA models seemed to outperform GLMs in terms of ROC-AUC and Recall, especially when combined with sPLS-DA feature selection. In contrast, GLM models, especially those trained on small samples with RF-selected features, exhibited signs of overfitting, with larger discrepancies between training and test performance (Figure 2). Nevertheless, RF-based feature models retained a degree of interpretability, relying on compact and biologically intuitive feature sets centered around a few high-signal terms (Figure 1).

Figure 2 displays the full set of performance metrics (Recall, Precision, F1-score, Balanced Accuracy, and ROC-AUC) for each model coupled with each feature selection approach. Our results support the feasibility of building more effective supervised classifiers to support article triage in SRs, starting from a broad query. Moreover, moderate-sized train sets (n ≥ 200) seem to yield informative generalizable models when coupled with thoughtful feature selection.

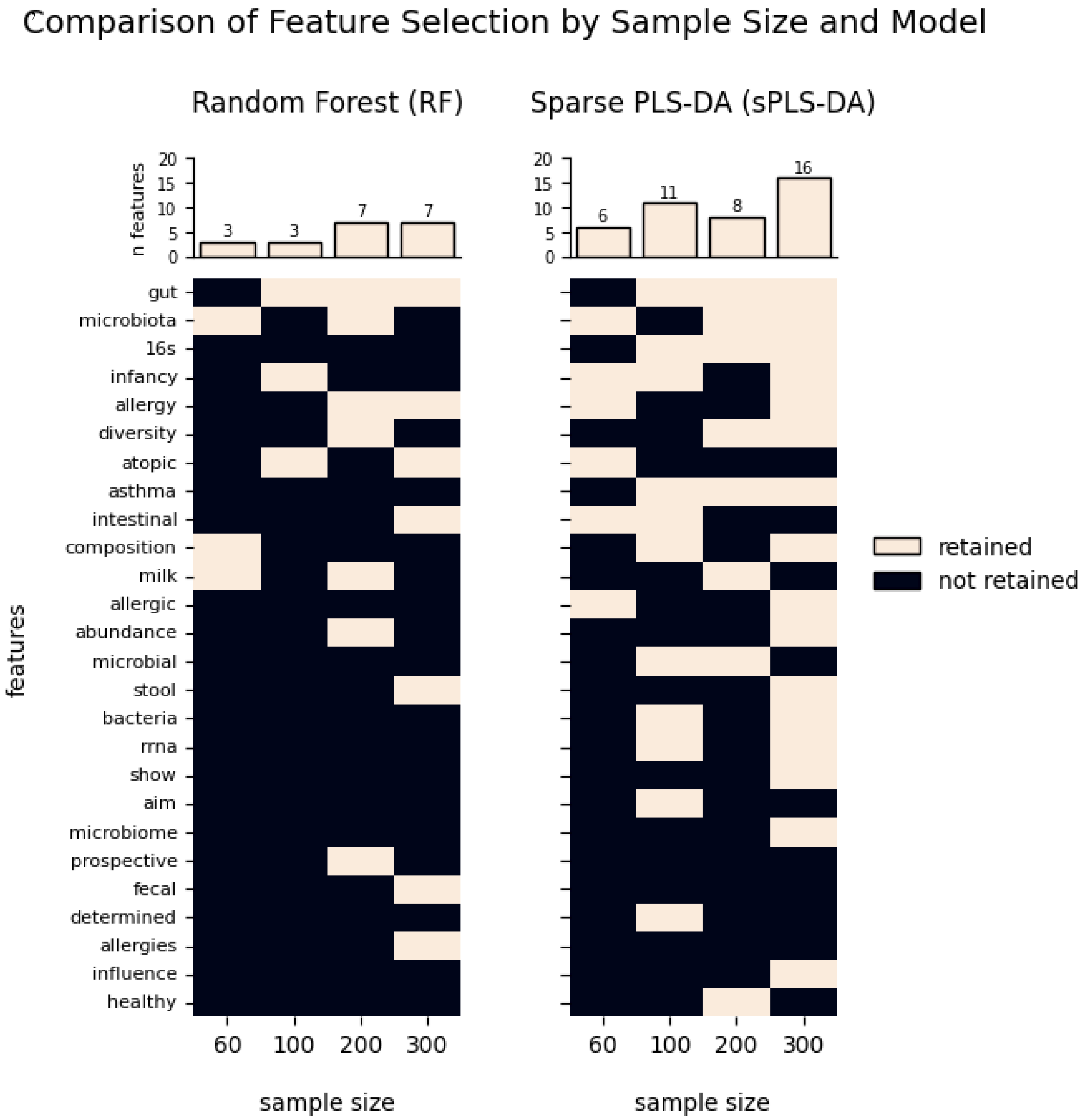


Figure 1. Word features retained across models after feature selection

Heatmap showing the presence of selected keywords (rows) across the 4 sample sizes used as trainset (columns). Models are grouped by feature selection method (RF or sPLS-DA), and training set size (n = 60, 100, 200, or 300). Relevant keywords number tends to grow with sample size, with a higher number of features retained by sPLS-DA than RF.

,

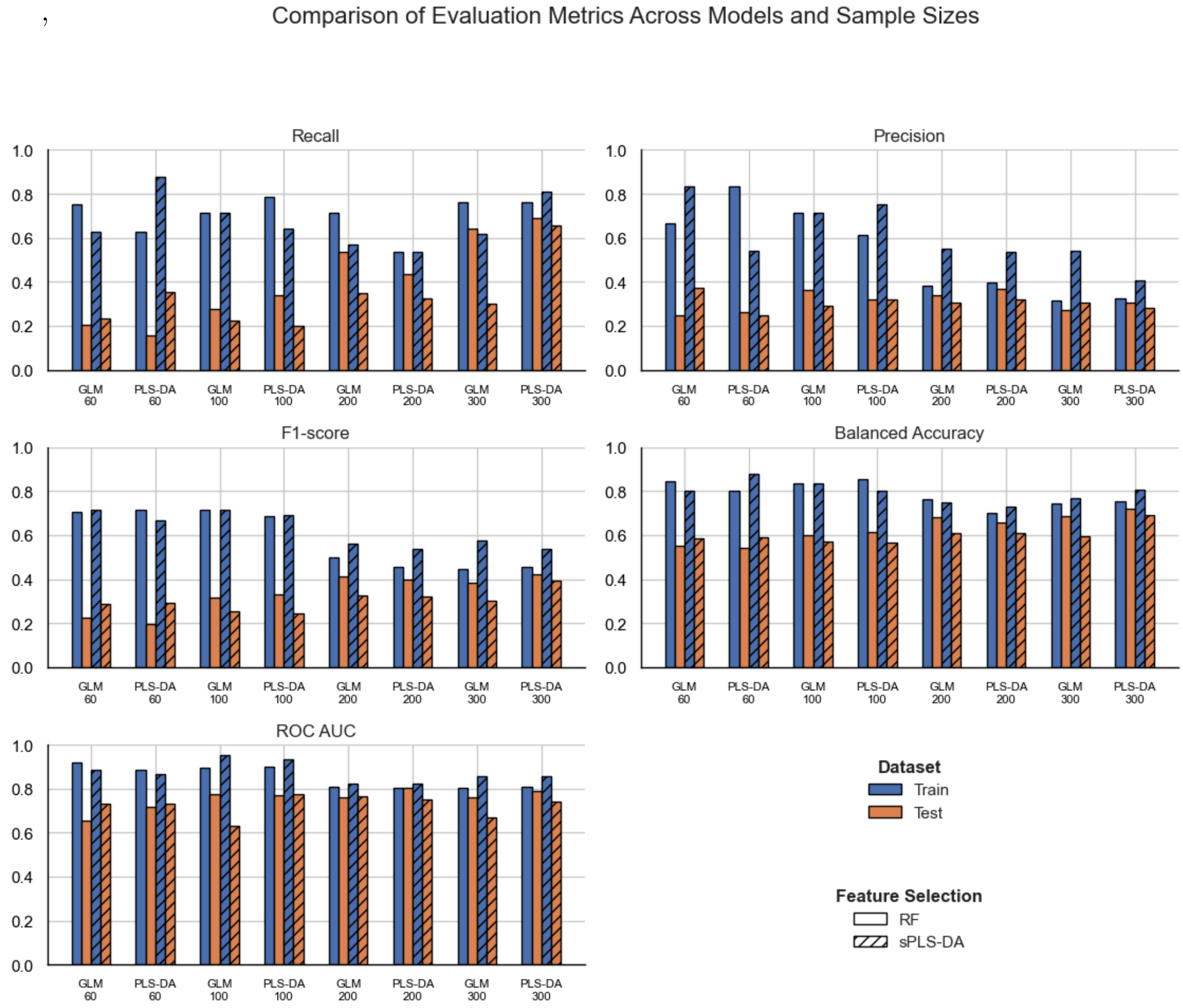


Figure 2. Performance metrics for all models on training and test sets

Bar plots showing Recall, Precision, F1-score, Balanced Accuracy, and ROC-AUC for each of the 16 models, evaluated separately on the train and test sets. The figure highlights the impact of training set size and feature selection strategy on generalization performance.

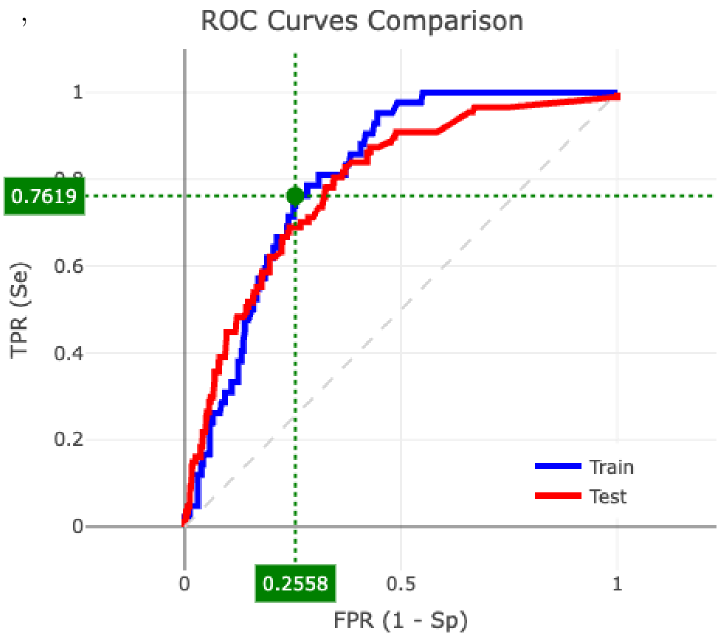


Figure 3. ROC curve for the best-performing model
ROC curve for the PLS-DA classifier (selected model threshold = 0.1789) trained on a sample of 300 articles after feature selection using Random Forest (retained words in the final model: gut, allergy, intestinal, atopic, allergies, fecal, stool). This model achieved the best overall performance across all metrics (F1-score = 0.4226, Recall = 0.6897, Precision = 0.3046, Balanced Accuracy = 0.7192, ROC AUC = 0.7885), with the smallest train-test performance gap.

Discussion

This study supports the potential of structured, feature-based machine learning approaches to support the automation of study selection in SRs and MAs. By applying a top-down pipeline combining lexical filtering and supervised classification, we were able to build predictive models that reached decent performance using moderate training sets. Our results suggest that article triage is achievable even without deep linguistic modeling, provided that relevant features are selected. In particular, combining PLS-DA with RF-based feature selection appears to offer an acceptable trade-off between interpretability and generalization.

Compared to recent advances using LLMs, our approach offers a lightweight, transparent, and reproducible alternative. While LLMs have shown impressive zero-shot performance in SR screening tasks (4,6), these methods still face challenges in particular interpretability and computational cost. In contrast, our method allows explicit control over vocabulary, model structure and performance thresholds, which may facilitate its integration into validated, auditable pipelines in biomedical research.

Nevertheless, our current approach presents several limitations. First, it relies on simple lexical representations based on word frequency, which ignore word order, grammatical structure, and semantic relationships. This limits model capacity, especially when abstracts include negations, ambiguous terms, or complex syntax. Second, performance was evaluated using a single random train/test split. Without cross-validation or stratified sampling, we cannot fully assess the stability of model performance across different subsets. Third, each model was trained independently, and no combination of classifiers (e.g., through ensembling) was explored. Using ensemble strategies could help mitigate overfitting and improve robustness, particularly when training data are limited.

To address these issues, future developments will integrate more sophisticated natural language processing (NLP) techniques, including lemmatization, pattern detection via regular expressions (regex) or stemming, and vector-based text representations such as static (e.g., Word2Vec, GloVe, FastText) and contextualized embeddings (e.g., BERT, GPT). These transformer-based approaches offer the potential to capture richer syntactic and semantic information beyond token frequency. Beyond article selection, the next stages of our pipeline will focus on automated metadata extraction from full-text articles. Named Entity Recognition (NER) techniques will be used to extract structured information on populations, sequencing methods, other confounders and outcomes.

We plan to evaluate both generic and domain-specific models – pre-trained on large biomedical corpora – such as BioBERT (7), SciBERT (8), and PubMedBERT (9), as well as general-purpose LLMs (e.g., GPT, LLaMA-4) fine-tuned for biomedical information extraction (4). These models will be evaluated in both zero-shot and fine-tuned settings for tasks such as study classification and metadata extraction. However, a fine-tuning step on annotated corpora from microbiota and allergy research is essential to enhance recognition accuracy, given the domain-specific terminology and the need for high-precision extraction of often underrepresented entities.
We also aim to complement our top-down selection strategy with a bottom-up approach based on citation network analysis, starting from manually validated seed articles. This dual strategy is expected to increase performance while maintaining relevance, and may help identify studies not well indexed by keyword searches alone.

In perspective, these efforts aim to support the construction of a fully automated meta-analysis pipeline. This pipeline will integrate sequential modules for article screening, structured data and metadata extraction, and harmonized downstream bioinformatic and statistical analysis. By automating each step while maintaining interpretability and methodological transparency, we aim to reduce human workload and subjectivity, enhance the reproducibility and scalability of SRs and MAs, and ultimately enable deeper, cross-study integration of datasets – particularly in complex domains such as microbiota research.

References

1. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

2. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc. 2006;2006:359-363.

3. Preiksaitis C, Ashenburg N, Bunney G, et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. JMIR Med Inform. 2024;12:e53787. Published 2024 May 10. doi:10.2196/53787

4. Delgado-Chaves FM, Jennings MJ, Atalaia A, et al. Transforming literature screening: The emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A. 2025;122(2):e2411962122. doi:10.1073/pnas.2411962122

5. Mirzayi C, Renson A, Genomic Standards Consortium et al. Reporting guidelines for human microbiome research: the STORMS checklist. Nat Med. 2021;27(11):1885-1892. doi:10.1038/s41591-021-01552-x

6. Wang S, Scells H, Zhuang S, et al. Zero-shot Generative Large Language Models for Systematic Review Screening Automation. ECIR 2024. arXiv preprint arXiv:2401.06320, 2024. doi:10.48550/arXiv.2401.06320

7. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682

8. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. EMNLP 2019. arXiv:1903.10676. doi:10.48550/arXiv.1903.10676

9. Han Q, Tian S, Zhang J. A PubMedBERT-based classifier with data augmentation strategy for detecting medication mentions in tweets. arXiv preprint arXiv:2112.02998, 2021. doi:10.48550/arXiv.2112.02998

,

Feature-Based Classification of High Pain Responders in Lyme Disease Patients Using Machine Learning

Teresa Ciavattini[1], Marc Shawky[1], Florian De Vuyst[2], Séverine Padiolleau-Lefèvre[3], Gordana Avramovic[4], and John Shearer Lambert[4]

[1]Université de Technologie de Compiègne, Costech Laboratory, 60203 Compiègne, France
[2]Université de Technologie de Compiègne, CNRS, Laboratory of Biomechanics and Bioengineering, 60203 Compiègne, France
[3]Université de Technologie de Compiègne, Unité de Génie Enzymatique et Cellulaire (GEC), CNRS UMR 7025, 60203 Compiègne, France
[4]Department of Infectious Diseases, Catherine Mc Auley Education Research Centre, Mater Misericordiae University Hospital, D07 A8NN Dublin, Ireland

Abstract

Lyme disease is a chronic infectious condition that presents with heterogeneous symptom profiles, complicating both diagnosis and treatment monitoring. To address the challenge of evaluating therapeutic response in the absence of standardized clinical endpoints, we developed a machine learning pipeline to classify responders and non-responders using patient-reported symptom survey data.

The study analyzed data from adult patients diagnosed with Lyme disease and treated at Mater Misericordiae University Hospital in Dublin. Symptom severity was assessed at baseline and post-treatment, and changes across multiple domains (fatigue, pain, mood, and overall symptom burden) were used to stratify patients into high responders, non-responders, and intermediate groups using a quantile-based approach.

To predict response class from baseline clinical features, we applied a filter-based feature selection method using ANOVA F-score (`f_classif`) to identify the top 20 predictive variables. These were used to train and evaluate four supervised classifiers (Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbors). We compared SMOTE oversampling and class weighting strategies for class imbalance, and assessed performance using stratified 5-fold cross-validation and multiple metrics (Accuracy, F1-score, ROC AUC, and MCC).

Our results show that baseline symptom profiles are predictive of treatment response, with Random Forest combined with SMOTE achieving the best overall performance. This study proposes a reproducible, interpretable ML framework for response classification in Lyme disease and lays the groundwork for future integrative analyses involving biomarkers and longitudinal modeling.

1 Introduction

Lyme disease is a tick-borne infection caused by *Borrelia burgdorferi*, with symptoms ranging from mild flu-like illness to severe neurological or musculoskeletal complications [1]. While antibiotic treatment is generally effective, a substantial subset of patients continues to experience long-term symptoms, a condition often referred to as post-treatment Lyme disease syndrome (PTLDS) [2].

One of the key challenges in clinical practice is predicting which patients will experience meaningful symptom improvement following treatment. Recent machine learning studies have attempted to address this, most notably the work of Vendrow et al. [3], who used symptom survey data to identify predictors of therapeutic response through a combination of feature selection techniques and deep learning classifiers.

Inspired by their approach, this study implements a simplified and interpretable machine learning pipeline tailored to a smaller clinical dataset. Unlike Vendrow et al., we do not employ neural networks due to data size

Authors are listed in the order of their contributions.

constraints, but focus on traditional supervised classifiers such as Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbors.

Our primary objective is to explore whether baseline symptom profiles can be used to classify patients into high responders and non-responders, based on post-treatment symptom change. To this end, we adopt a global feature selection strategy, performed prior to model training, to identify the most informative clinical features. This design choice, aligned with the exploratory and feature-focused nature of Vendrow et al.'s methodology, prioritizes interpretability and the identification of stable clinical markers over predictive generalization.

2 Methods

Before diving into the specific steps, our methodological pipeline can be broadly divided into three components: (1) data preprocessing and symptom-based classification, (2) feature selection, and (3) machine learning modeling and evaluation. Each step was designed to ensure robustness, interpretability, and clinical relevance of the results.

2.1 Preprocessing Strategy for Clinical Survey Data

2.1.1 Dataset Description

The dataset used in this study consists of anonymized clinical survey responses collected from patients diagnosed with Lyme disease at Mater Misericordiae University Hospital in Dublin. Surveys were administered at two timepoints: baseline (T0, pre-treatment) and follow-up (T2, post-treatment), enabling longitudinal symptom tracking. The dataset includes 176 clinical features covering physical, cognitive, and emotional symptoms, treatment details, and comorbidities.

2.1.2 Categorical Variable Encoding

Clinical survey data often contain categorical variables, such as symptom presence or gender. We applied binary encoding for dichotomous features (e.g., gender, prior diagnosis) and one-hot encoding for nominal variables with more than two categories (e.g., antibiotic tolerance). This approach preserves the categorical information without introducing artificial ordinal relationships [4]. Proper encoding is essential for model interpretability in medical machine learning contexts.

2.1.3 Handling Missing Data

To address missing values, we excluded features with more than 70% missingness, a commonly used threshold that balances model robustness and data retention [5]. Median imputation was used for continuous variables due to its robustness to outliers [6]. For binary features, missing entries were imputed with a constant value of -1, allowing models to potentially capture the informativeness of missingness itself. This strategy aligns with previous studies such as MyLymeData [3], where non-responses were explicitly retained for modeling.

2.1.4 Feature Scaling and Standardization

All continuous variables were standardized to zero mean and unit variance, a key step for optimizing performance in distance-based or regularized models [7, 4]. Such normalization practices are widely used in clinical machine learning pipelines and are known to enhance subgroup separability [7].

2.1.5 Leakage Prevention

To avoid data leakage and simulate realistic clinical deployment, we excluded from the feature set the four variables directly used to define the response classification: `severe_fatigue_rate`, `muscle_pain_rate`, `symp_today_rate`, and `mood_rate`. All other clinical survey features, including those from both T0 and T2, were retained for modeling.

,

2.1.6 Application to Lyme Disease Data

Prior work has emphasized the importance of robust preprocessing for Lyme-related survey data. Vendrow et al. [3] demonstrated the value of constant-value imputation, categorical encoding, and dimensionality reduction, while Kehoe et al. [7] highlighted the impact of normalization in biomarker classification. Building on these findings, our pipeline integrates best practices for reproducible and clinically interpretable machine learning.

2.2 Pain Response Classification

2.2.1 Calculation of Symptom Change (T2 – T0)

Treatment response was assessed by measuring self-reported changes in four key symptom domains: severe fatigue, muscle pain, symptom severity today, and mood. These symptoms were selected for their consistent clinical relevance in Lyme disease and other post-infectious syndromes [**?** 8, 9].

Each variable was measured on a 1–10 Likert scale at both baseline (T0) and follow-up (T2) [10]. The direction of improvement depended on the question: lower scores indicated improvement for fatigue and pain, while higher scores indicated improvement for mood and overall symptoms.

Table 1: Symptom variables used for response classification

Symptom	**Scale (1–10)**	**Survey Item(s)**
Severe fatigue	1 = no fatigue 10 = severe fatigue	"In the last six months, are you experiencing unexplained severe fatigue not relieved by rest?" If yes: "How would you rate the level of severe fatigue that you have experienced?"
Muscle pain	1 = no pain 10 = severe pain	"Have you experienced muscle pain in the last six months?" If yes: "How would you rate the pain in your joints or muscles?"
Symptoms today	1 = very poor 10 = very well	"How would you rate how you are feeling today regarding your symptoms?"
Mood	1 = very low mood 10 = very good mood	"Do you feel 'down' or in low mood because of your symptoms?" If yes: "How would you rate your overall mood?"

We computed the difference between T2 and T0 for each symptom. These difference scores were then used to derive symptom-specific improvement distributions.

2.2.2 Quantile-Based Classification of Responders

To account for variation in baseline severity and subjective bias, we employed a quantile-based strategy for stratification. For each symptom, patients were grouped into tertiles based on their change scores. The top third (Q1) represented the greatest improvement, and the bottom third (Q3) the least.

We then defined multi-dimensional responder categories:

- **High responders:** Q1 in at least 3 out of 4 symptoms
- **Non-responders:** Q3 in at least 3 out of 4 symptoms
- **Others:** All remaining patients (excluded from modeling)

,

2.2.3 Subsampling Strategy

To ensure balanced class representation, only high and non-responders were retained for modeling. All other patients were excluded to reduce class ambiguity [3]. When class imbalance was present, we applied stratified sampling to preserve equal proportions of high and non-responders in the training data. This helped reduce variability due to random sampling and ensured better model differentiation between the two groups.

2.3 Feature Selection and Machine Learning Pipeline

2.3.1 Feature Selection

To reduce dimensionality and improve model interpretability, we applied a filter-based feature selection using the `SelectKBest` method with ANOVA F-score (`f_classif`) scoring. This technique assesses the dependency between each feature and the target variable by comparing between-class and within-class variance, making it particularly effective for identifying features that contribute significantly to class separation in high-dimensional survey data.

To avoid data leakage, all features directly used to define treatment response were excluded. From the remaining pool, the top 20 features were selected. This threshold was chosen based on previous work suggesting it provides a balance between model expressiveness and overfitting risk in high-dimensional clinical datasets [3, 11]. Feature selection was performed prior to any class balancing to prevent synthetic samples from influencing the selection process.

Importantly, this selection step was conducted once on the full dataset prior to model evaluation. While this approach may introduce a potential risk of information leakage in strictly predictive settings, it was intentionally adopted to align with the methodology described in Vendrow et al. [3], where feature selection was performed globally to identify a subset of relevant survey items before classification. Our primary goal was to explore stable clinical predictors of treatment response, consistent with the exploratory and feature-focused nature of the original study, rather than to optimize predictive generalization.

2.3.2 Class Balancing

Given the class imbalance between high responders and non-responders, we tested two balancing strategies:

1. **SMOTE (Synthetic Minority Oversampling Technique):** Generates synthetic samples of the minority class to improve generalization and sensitivity in clinical prediction [12].

2. **Cost-sensitive learning:** Uses the `class_weight='balanced'` parameter to adjust the penalty for misclassifications based on class frequency, preserving data integrity without oversampling [13].

K-Nearest Neighbors was tested only with SMOTE, as class-weighting is not natively supported for this algorithm.

2.3.3 Model Training and Evaluation

We trained and evaluated four supervised learning models [3] commonly used in clinical tabular data:

- **Random Forest (RF):** An ensemble method based on decision trees that reduces overfitting and captures non-linear relationships. It is robust to noise and performs implicit feature selection.

- **Support Vector Machine (SVM):** A linear kernel SVM was used to construct a hyperplane separating the classes with maximum margin, particularly effective in high-dimensional spaces.

- **Logistic Regression (LR):** A probabilistic linear model suitable for binary classification, offering interpretable coefficients.

- **K-Nearest Neighbors (KNN):** A distance-based non-parametric method that classifies a sample by majority vote of its k nearest neighbors in feature space.

All models were validated using stratified 5-fold cross-validation to preserve class proportions in each fold. We evaluated model performance using four metrics:

,

- **Accuracy:** The proportion of correctly classified instances over the total number of samples.
- **F1-score:** The harmonic mean of precision and recall, which balances false positives and false negatives.
- **ROC AUC:** The Area Under the Receiver Operating Characteristic Curve, indicating the model's ability to distinguish between classes across thresholds.
- **Matthews Correlation Coefficient (MCC):** A robust measure for imbalanced datasets, taking into account true and false positives and negatives. It returns a value between −1 (total disagreement) and +1 (perfect prediction), and is considered more informative than accuracy in skewed settings [13].

No hyperparameter tuning was performed in this preliminary analysis; all models were used with scikit-learn's default or commonly recommended parameters. To ensure reproducibility, all random seeds were fixed across libraries and processes.

2.3.4 Feature Importance

To interpret model decisions, we computed and compared two feature importance measures on the trained Random Forest model: (1) impurity-based importance, which reflects the average decrease in Gini impurity, a measure of node impurity used to split decision trees, where lower values indicate purer class distributions; and (2) permutation importance, which evaluates the drop in model performance when the values of a feature are randomly shuffled [14]. This comparison enables the identification of stable predictors that consistently contribute to model discrimination.

3 Results

3.1 Patient Classification

Patients were stratified using a quantile-based classification, as described in Section 2.2.2. This approach allowed the selection of individuals at the extremes of symptom variation, thereby enabling clearer differentiation between responder categories.

Table 2 summarizes the distribution of patients across quantiles (Q1, Q2, Q3) for each symptom variable, along with the corresponding response class assignments derived from the multi-dimensional rule.

Table 2: Patient classification summary based on quantile thresholds

Parameter	**Q1**	**Q2**	**Class**	**Count**
`severe_fatigue_rate`	-4.00	-2.00	unknown	96
			high-responder	77
			non-responder	65
			low-responder	63
`muscle_pain_rate`	-3.00	-1.00	high-responder	96
			unknown	87
			low-responder	78
			non-responder	40
`symp_today_rate`	0.00	3.00	unknown	90
			low-responder	89
			high-responder	81
			non-responder	41
`mood_rate`	1.00	3.00	unknown	95
			high-responder	82
			low-responder	63
			non-responder	61

,

3.2 Model Performance

Model performance was evaluated using four metrics: Accuracy, F1-score, ROC AUC, and Matthews Correlation Coefficient (MCC). Two balancing strategies were compared: SMOTE and `class_weight='balanced'`. The results of 5-fold stratified cross-validation are reported in Table 3 and illustrated in Figure 1.

The Random Forest model combined with SMOTE yielded the highest performance across all evaluation metrics (Accuracy = 0.84, F1 = 0.84, ROC AUC = 0.84, MCC = 0.69). This indicates that synthetic oversampling contributed to improved class separation and model generalization. In contrast, models trained using class weighting performed consistently lower, particularly in terms of MCC, suggesting that cost-sensitive learning alone was insufficient to address the class imbalance.

Table 3: Model performance comparison across balancing strategies

Model	**Strategy**	**Accuracy**	**F1-score**	**ROC AUC**	**MCC**
RF + SMOTE	SMOTE	0.84	0.84	0.84	0.69
LogReg + SMOTE	SMOTE	0.79	0.78	0.79	0.59
SVM + SMOTE	SMOTE	0.78	0.77	0.78	0.58
KNN + SMOTE	SMOTE	0.76	0.71	0.76	0.56
SVM + weights	class_weight	0.74	0.81	0.72	0.42
LogReg + weights	class_weight	0.72	0.79	0.69	0.37
RF + weights	class_weight	0.75	0.83	0.67	0.38

As shown in Figure 1, SMOTE outperformed the class-weighted strategy across all models.

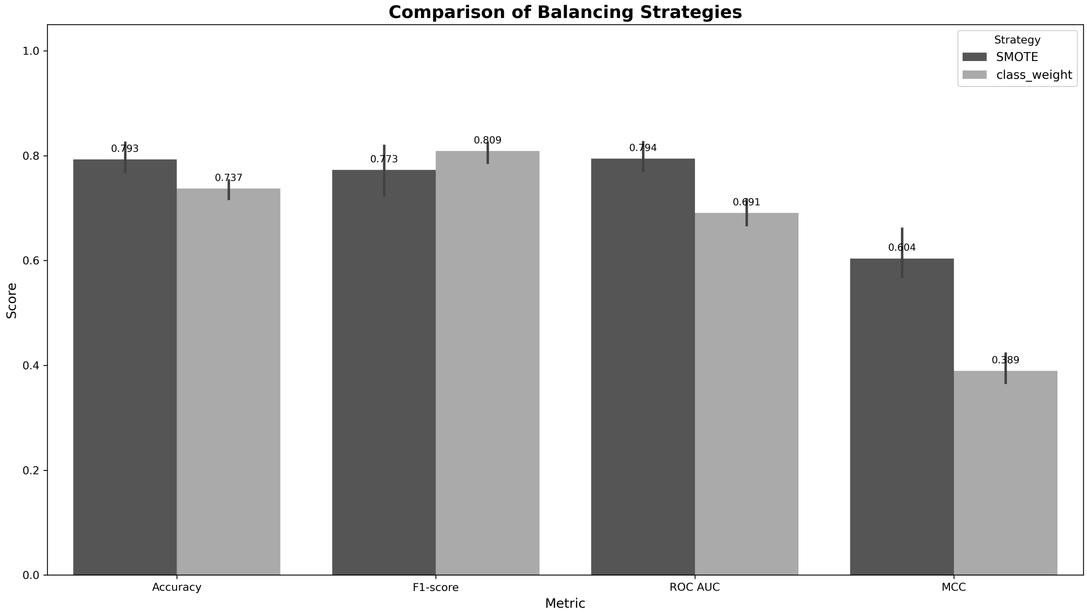


Figure 1: Comparison of class balancing strategies across performance metrics.

3.3 Feature Importance

To evaluate the contribution of individual features to model predictions, we applied both internal Random Forest importance ranking and permutation-based importance. Figure 2 shows the features most consistently ranked at the top by both methods.

The following variables were among the most frequently selected:

,

- `T2_care_crc_rate` – How would you rate the treatment and care that you have received for the past six months at the tickborne infections consultation service?
- `T2_antib_dur` – Were you prescribed antibiotic therapy at your first visit here? If yes, for how long did you take antibiotic therapy in total (in weeks)?
- `T0_swglands` – Have you experienced swollen glands in the last six months?
- `T2_prior` – In the months prior to your diagnosis, how would you rate how you were feeling about your health?

These features were identified purely based on their statistical relevance to classification, without clinical interpretation in this analysis.

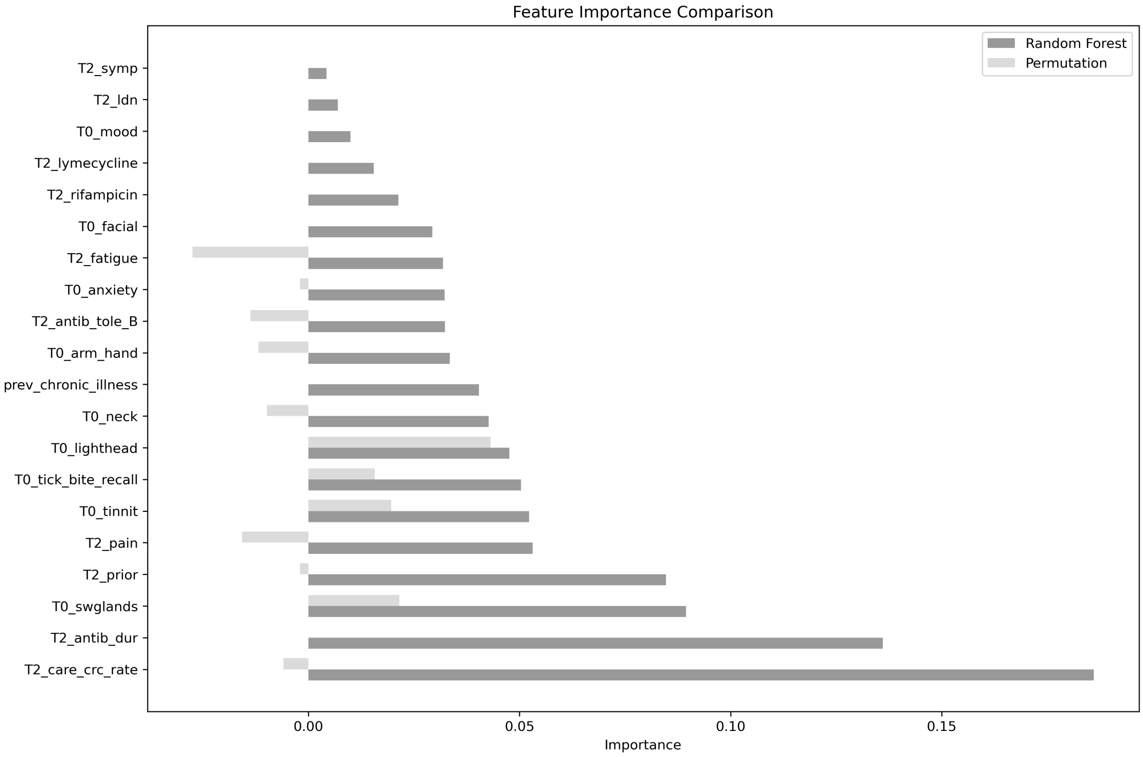


Figure 2: Feature importance comparison using Random Forest and permutation-based methods. Features are ranked by their impact on classification performance.

4 Discussion and Conclusion

The choice to perform feature selection globally, on the full dataset prior to model evaluation, represents an important methodological consideration. While this approach may inflate performance estimates in strictly predictive settings, it was intentionally adopted to remain consistent with the methodology of the original study by Vendrow et al. [3]. In their work, feature selection was conducted globally to identify a subset of clinically meaningful variables before classification. Our primary objective, similarly, was to explore stable clinical predictors of treatment response, not to optimize generalization performance. Nevertheless, in future work we plan to implement feature selection within cross-validation folds to obtain more robust performance estimates and to support the development of clinically deployable predictive models.

Several additional enhancements are planned for future iterations of the pipeline. We will implement more sophisticated imputation techniques such as random forest-based methods (e.g., MissForest), which have shown improved performance compared to traditional approaches in clinical datasets [15]. We also aim to explore advanced classification strategies, including ensemble and boosting-based models, to improve predictive robustness.

Another important development will be the integration of biomarker and blood test data into the model. Recent studies have demonstrated that serologic biomarkers and multiplex peptide profiling can enhance classification accuracy in Lyme disease [16]. By incorporating multi-modal data, we expect to create more biologically grounded and generalizable predictive models.

Lastly, we plan to compare our results with those of Vendrow et al. [3] and carry out proper clinical interpretation of the most informative features. This will provide valuable insight without drawing premature conclusions.

In summary, this work lays the groundwork for an interpretable ML framework for Lyme disease symptom data and sets the stage for future pipeline refinements and integrative, clinically informed modeling.

5 Acknowledgements

This project is co-funded by the European Union's Horizon Europe research and innovation program Cofund SOUND.AI under the Marie Sklodowska-Curie Grant Agreement No 101081674.

References

[1] Allen C Steere. Lyme disease: a growing threat to urban populations. *Proceedings of the National Academy of Sciences*, 101(47):17859–17860, 2004.

[2] Paul G Auwaerter, Johan S Bakken, Raymond J Dattwyler, J Stephen Dumler, John J Halperin, Edward McSweegan, Raphael B Nadelman, Eugene D Shapiro, Sandeep K Sood, Allen C Steere, et al. Lyme borreliosis: clinical case definitions for diagnosis and management in europe. *Clinical Microbiology and Infection*, 17(1):69–79, 2011.

[3] Joshua Vendrow, James Haddock, Deanna Needell, and Lorraine Johnson. Feature selection from lyme disease patient survey using machine learning. *Algorithms*, 13(12):334, 2020.

[4] Lukas Heumos, Philipp Ehmele, Tim Treis, Julius Upmeier zu Belzen, Eljas Roellin, and et al. An open-source framework for end-to-end analysis of electronic health record data. *Nature Medicine*, 30:3369–3380, 2024.

[5] Marziyeh Afkanpour, Elham Hosseinzadeh, and Hamed Tabesh. Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review. *BMC Medical Research Methodology*, 24(1):188, 2024.

[6] A. R. Ismail, N. Z. Abidin, and M. K. Maen. Systematic review on missing data imputation techniques with machine learning algorithms for healthcare. *Journal of Robotics and Control*, 3(2):143–152, 2022.

[7] Eric R. Kehoe, Bryna L. Fitzgerald, Barbara Graham, M. Nurul Islam, Kartikay Sharma, Gary P. Wormser, John T. Belisle, Adriana Marques, Paul G. Auwaerter, Kristin Kirkland, John N. Aucott, Mark J. Soloski, Howard Gewurz, et al. Biomarker selection and a prospective metabolite-based machine learning diagnostic for lyme disease. *Scientific Reports*, 12:1478, 2022.

[8] Shannon M. Smith, Robert H. Dworkin, Dennis C. Turk, and et al. Interpretation of chronic pain clinical trial outcomes: IMMPACT recommended considerations. *Pain*, 161(11):2446–2461, 2020.

[9] Kunal Garg, Abbie Thoma, Gordana Avramovic, Leona Gilbert, Marc Shawky, Minha Rajput Ray, and John Shearer Lambert. Biomarker-based analysis of pain in patients with tick-borne infections before and after antibiotic treatment. *Antibiotics*, 13(8):693, 2024.

,

[10] David Xi, Abbie Thoma, Minha Rajput-Ray, Anne Madigan, Gordana Avramovic, Kunal Garg, Leona Gilbert, and John S. Lambert. A longitudinal study of a large clinical cohort of patients with lyme disease and tick-borne co-infections treated with combination antibiotics. *Microorganisms*, 11(9):2152, 2023. Published 24 August 2023.

[11] Farideh Mohtasham, Mohamad Amin Pourhoseingholi, Seyed Saeed Hashemi Nazari, Kaveh Kavousi, and Mohammad Reza Zali. Comparative analysis of feature selection techniques for covid-19 dataset. *Scientific Reports*, 14:18627, 2024.

[12] Yuxuan Yang, Hadi A. Khorshidi, and Uwe Aickelin. A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. *Frontiers in Digital Health*, 6:1430245, 2024.

[13] Mabrouka Salmi, Dalia Atif, Diego Oliva, Ajith Abraham, and Sebastian Ventura. Handling imbalanced medical datasets: review of a decade of research. *Artificial Intelligence Review*, 57:273, 2024.

[14] Joonhyuk Cho, Qingyang Xu, Chi Heem Wong, and Andrew W. Lo. Predicting clinical trial duration via statistical and machine learning models. *Contemporary Clinical Trials Communications*, 45:101473, 2025.

[15] Lucy Grigoroff, Reika Masuda, John Lindon, and et al. Evaluation of imputation strategies for multicentre studies: Application to a large clinical pathology dataset. *Research Square*, 2024. Random Forest-based methods (MissForest) outperform MICE.

[16] Tingting Zhang, Laurie Baert, Neal W. Woodbury, and Laimonas Kelbauskas. Serologic biomarker discovery for differentiating lyme disease from diseases with similar clinical symptoms. *Frontiers in Immunology*, 2025.

,

Interpretable AI for Classifying Human- and LLM-Generated Medical Misinformation with Multi-Modal Features

Mahshid Baharifar, Anima Kujur, Zahra Monfared

Interdisciplinary Center for Scientific Computing (IWR), Department of Mathematics and Computer Science, Heidelberg University, Heidelberg, 69120, Germany

ABSTRACT: The growing spread of medical misinformation, particularly content generated by Large Language Models (LLMs), presents serious public health risks. In this contribution, we present an interpretable AI framework for classifying human- and LLM-generated medical misinformation using the Med-MMHL dataset , which provides multi-modal features including text, metadata, and source labels. Our approach combines comprehensive Exploratory Data Analysis (EDA), advanced resampling strategies, and interpretability techniques to develop robust and explainable classifiers. We benchmark a range of machine learning and deep learning models, including Multinomial Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Feedforward Neural Networks, Bidirectional Encoder Representations from Transformers (BERT), and BioBERT . To mitigate class imbalance, we applied Synthetic Minority Oversampling Technique, undersampling, and a hybrid approach, which significantly improved model performance. BioBERT, fine-tuned with hybrid sampling, achieved the highest performance. We further employ Local Interpretable Model-agnostic Explanations and Shapley Additive Explanations (SHAP) to enhance model interpretability, identifying lexical complexity, sentiment, and source metadata as key predictors. SHAP-based meta-analysis reveals distinctive syntactic patterns in LLM-generated misinformation , enabling a preliminary taxonomy of misinformation types and exposing weaknesses in short-text classification by traditional models. Transformer models like BioBERT shows superior contextual understanding of medically nuanced statements, while simpler models struggled with ambiguous or densely packed information. Our contributions are threefold: a rigorous benchmark of classical and neural models on a publicly available multi- modal dataset, establishing strong performance baselines for future research; an interpretable pipeline that integrates EDA, imbalance mitigation, and explainable tools for medical misinformation detection; and novel insights into the linguistic and contextual patterns of LLM-generated misinformation, informing downstream tasks. Future work will explore improved multi-modal fusion and adversarial robustness to advance the detection and understanding of medical misinformation.

Keywords: LLM, Machine Learning, Deep Learning, Explainable AI, Multi Modal Features

1. INTRODUCTION

The spread of false medical information has become a serious problem in today's world. With the rise of the internet and social media, people can easily share information—both true and false—very quickly. In recent years, powerful language models like ChatGPT and GPT-4 have made this issue even more complex. These models can produce text that looks correct and trustworthy, even if the information is actually wrong. This makes it harder to detemine what's true and what's not, especially because the language used by these models is often smooth, clear, and free of obvious mistakes. In contrast, human-written misinformation sometimes includes emotional or unusual writing patterns that can help with detection. However, AI-generated text often sounds more neutral and professional which can make people more likely to trust it- even when they should not.

This kind of misinformation can be especially harmful in the medical field. When people read false advice about diseases, vaccines, or treatments, they may make poor health decisions. This can cause confusion, reduce trust in doctors and health organizations, and even lead to dangerous actions. Because of the size and seriousness of this problem, we need AI tools that can detect misinformation clearly, correctly, and in a way that is easy to understand.

In this study, we build a system that can detect both human-written and AI-generated medical misinformation. We use a dataset called Med-MMHL, which includes medical texts along with extra information like their sources and truth labels. First, we observed a class imbalance in a data set- certain types of misinformation were significantly underperformed compared to other. These methods help the models learn better by giving more attention to the rare types of misinformation. Next, we tested several machine learning models. These included simple models like Logistic

Regression and Support Vector Machines, as well as more advanced models like BERT and BioBERT, which are specially trained to understand language. We found that while the simple models did a decent job—especially when the data was balanced—BioBERT gave the best overall performance because it understands complex language patterns better. To make sure our system is not just accurate but also understandable, we used two tools: SHAP and LIME. These tools help explain why the model made a certain decision by showing which words or patterns were most important. For example, the model often focused on word choices, tone, and how confident the statements were. One category of misinformation (referred to as category 1) remained difficult to classify accurately, as it continued numerous ambiguous or context-dependent phrases. This shows that further improvements—like better data cleaning or adjusting how features are used—could help in the future.

Overall, our findings highlight the importance of developing systems that not only achieve high accuracy in detecting misinformation but also provide transparent and interpretable outputs to support user trust.

2. Related Work

The detection of medical misinformation has become a crucial challenge at the intersection of Natural Language Processing (NLP), biomedical informatics, and trustworthy AI. Prior work in this domain can be broadly categorized into three areas: datasets and benchmarks, modeling strategies, and approaches for explainability and robustness in medical AI systems.

Datasets and Benchmarks for Medical Misinformation Detection: In recent years, several domain-specific benchmarks have been introduced to better capture the complexities and subtleties of medical misinformation. Sun et al. [1] introduced the Med-MMHL dataset, a multimodal resource that incorporates text, metadata, and source credibility to support misinformation detection across both human- and LLM-generated content. The dataset spans formats such as tweets and fake news, enabling fine-grained multimodal analysis. Building on this, Sun et al. [6] investigated LLM vulnerabilities, finding that despite strong performance, LLMs are prone to subtle hallucinations. Pal et al. [2] contributed Med-HALT, a benchmark designed to evaluate hallucinated biomedical claims in LLM outputs. Similarly, MedHal by Mehenni and Zouaq [3] focused on hallucination detection within clinical contexts, emphasizing frequency and severity of false content. Chen et al. [4] extended this work to vision-language models, arguing that standard factual consistency metrics are inadequate for the medical domain and advocating for domain-sensitive alternatives. Tian et al. [5] underscored the significance of contextual and source metadata in shaping user interpretation of medical information. To enhance global applicability, the ICHI 2024 Doctoral Consortium [7] introduced multilingual and multimodal benchmarks to address cross-cultural misinformation dynamics.

Modeling Approaches for Misinformation Detection: Medical misinformation detection strategies span classical and deep learning methods. Transformer-based models such as BioBERT and ClinicalBERT have shown high efficacy in capturing domain-specific semantics due to their pretraining on biomedical corpora [6,11]. Nevertheless, classical models like Logistic Regression and Support Vector Machines (SVMs) remain relevant, especially when paired with domain-aware features [8,9]. These approaches offer interpretability and efficiency, particularly in low-resource settings. Class imbalance is a persistent issue, as factual content far outweighs misinformation. Techniques like SMOTE and random undersampling have been used effectively by Lei [10] and Gupta et al. [13] to improve recall and macro-F1. Zhou and Liu [14] further identified linguistic cues—such as semantic drift and hedging—characteristic of LLM-generated misinformation, providing useful features for both classical and neural models.

Explainability and Robustness in Medical AI: Trustworthy AI in healthcare demands interpretable and resilient systems. Model-agnostic methods like SHAP and LIME offer local interpretability by highlighting feature contributions. Kim and Nguyen [12] applied SHAP to misinformation classification, uncovering clinically meaningful attribution patterns. Robustness under perturbation is equally critical: Chen and Wang [16] showed that LLM-based detectors are vulnerable to minimal input changes, and Hassan and Ali [17] demonstrated deep learning models' fragility under noisy COVID-19 scenarios. Multimodal architectures have been proposed to improve generalization. Patel and Sharma [15] integrated metadata and text through transformer encoders, while Torres and Gomez [18] used attention mechanisms to extract salient features. Singh and Jain [19] advocated for scalable models tailored for the dynamic and high-volume nature of social media misinformation.

3. Materials

3.1. Data Description

This study employs the Med-MMHL dataset, a large-scale, domain-specific benchmark designed to support the detection of medical misinformation on social media, particularly Twitter. The dataset contains over 60,000 health-related tweets collected using Twitter's API, filtered for content relevant to medicine, public health, and scientific communication. In addition to textual data, the dataset includes metadata such as source labels and links to associated images to support multimodal analysis. However, this work focuses solely on the textual component, allowing the model to concentrate on linguistic patterns without introducing visual complexity. The dataset is split into training (44,799 samples), validation (6,405 samples), and test (12,800 samples) subsets, following a 70%-10%-20% ratio. To support sentiment-aware analysis, each tweet was also processed using the VADER sentiment analyzer, assigning polarity labels (positive or negative) and enabling an overview of sentiment distribution.

Label reliability was ensured through a two-phase annotation process. First, tweets were matched against trusted sources like the WHO and CDC. Then, public health experts validated each label. Tweets are labeled as real (1) or fake (0). These labeled subsets are now fully prepared for further preparation and preprocessing and model development.

3.2. Data Preparation

The Med-MMHL dataset was created by integrating two independent sources: one targeting COVID-19 misinformation and another covering general medical content. To ensure structural consistency, column names across both datasets were standardized to id, message, and label. The merged dataset was then shuffled using a fixed random seed (random_state = 42) to ensure reproducibility across experimental runs. Entries with missing or null values were systematically removed, resulting in a clean, harmonized dataset ready for preprocessing and model development.

To examine the dataset's linguistic characteristics, we conducted Exploratory Data Analysis (EDA). A word frequency analysis revealed key terms such as "study," "health," and "research" as the most common, reflecting the dataset's medical and scientific orientation. For further analysis, we compared the terms used in real and fake tweets across the training, validation, and test sets using various visualizations, including frequency charts and word clouds. Fake tweets often referenced institutions such as "Mayo Clinic" or "National Institutes," likely to enhance their perceived credibility. In contrast, real tweets displayed a broader range of discourse, including critical expressions like "opposite opinion."

A custom preprocessing function was applied to normalize the message text. The function converted text to lowercase, removed URLs, mentions, hashtags, punctuation, and stopwords—including custom social media tokens such as "u," "dont," and "ure"—and applied lemmatization. This yielded a new clean_message column with significantly reduced noise. This cleaning step led to a 39.6% reduction in average token count across all splits, decreasing the training set average from 66 to 40 tokens. Additionally, frequent token distribution has been shifted in before and after cleaning: high-frequency stopwords like "the" and "to" were replaced by semantically meaningful terms such as "study" and "health." These improvements enhanced the dataset's quality and interpretability for downstream modeling tasks.

3.3. Pre_processing

To prepare the cleaned Med-MMHL dataset for modeling, a comprehensive preprocessing pipeline was applied to address distributional irregularities, class imbalance, and feature representation. The initial analysis of message lengths revealed several extreme outliers, with a small number of messages exceeding 20,000 characters. These unusually long messages posed challenges for tokenization and model convergence. Therefore, outliers were removed prior to vectorization. This step yielded a more stable distribution of message lengths and improved computational efficiency without substantial data loss.

Subsequent exploratory analysis employed Kernel Density Estimation (KDE) and box plots to examine message length distribution. The analysis revealed a pronounced right-skew, with the majority of messages falling under 500 characters. To further support model interpretability, we stratified the dataset based on text length into short (≤1,000 characters) and long (1,000–10,000 characters) categories. This categorization improved downstream evaluation by allowing performance metrics to be assessed across different text length groups. The class distribution within the dataset was highly imbalanced, with fake messages constituting approximately 72% of the data, compared to 28% real. To address this, we applied a combination of undersampling, oversampling, and the Synthetic Minority Over-sampling Technique (SMOTE). These balancing techniques mitigated bias and enhanced model generalization, particularly for the minority class. For feature extraction, we employed CountVectorizer and TF-IDF to transform the cleaned textual content into structured numerical vectors. These methods were applied to the clean_message column, ensuring that only semantically meaningful tokens contributed to model input. This preprocessing strategy significantly improved the signal-to-noise ratio and supported robust training of both traditional classifiers and deep learning architectures.

4. Methodology

The methodology employed in this study as shown in the Fig.1 encompasses several critical phases to ensure effective data handling and analysis. The process initiates with Data Mining, where over 60,000 tweets and sentiments are collected, annotated for misinformation using official and verified sources, and validated by medical experts. This is followed by Data Preparation, which includes utilizing two separate datasets, removing null/missing entries, shuffling with a random state of 42, and merging the datasets for consistency. The subsequent Data Cleaning phase involves lowercasing text, removing URLs, mentions, hashtags, numbers, stop words, and non-alphanumeric characters, followed by tokenization, lemmatization, and joining clean words into a string. Next, Data Pre-processing comprises outlier filtering, noise/extreme value removal, vectorization using CountVectorizer or TF-IDF, categorization of short and long messages, and class balancing through undersampling, oversampling with SMOTE, or a hybrid approach. The final Data Training phase entails splitting the dataset into training, validation, and test sets, performing hyper-parameter tuning using Grid Search or Random Search with cross-validation, and evaluating the models on validation/test sets to assess performance metrics.

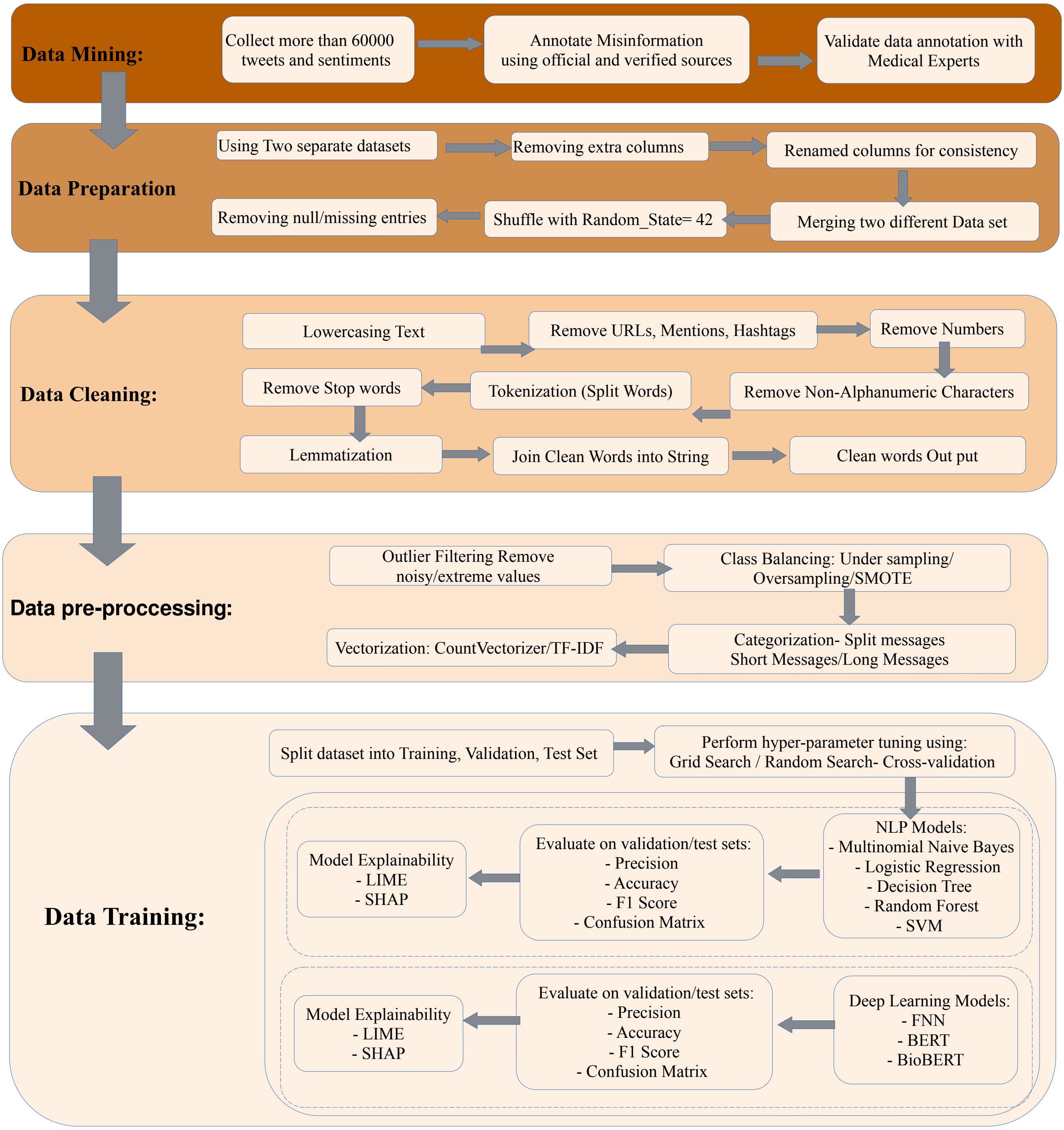

Data Mining:
Collect more than 60000 tweets and sentiments
Annotate Misinformation using official and verified sources
Validate data annotation with Medical Experts
Data Preparation
Using Two separate datasets
Removing extra columns
Renamed columns for consistency
Merging two different Data set
Shuffle with Random_State= 42
Removing null/missing entries
Data Cleaning:
Lowercasing Text
Remove URLs, Mentions, Hashtags
Remove Numbers
Remove Non-Alphanumeric Characters
Tokenization (Split Words)
Remove Stop words
Lemmatization
Join Clean Words into String
Clean words Out put
Data pre-proccessing:
Outlier Filtering Remove noisy/extreme values
Class Balancing: Under sampling/ Oversampling/SMOTE
Categorization- Split messages Short Messages/Long Messages
Vectorization: CountVectorizer/TF-IDF
Data Training:
Split dataset into Training, Validation, Test Set
Perform hyper-parameter tuning using: Grid Search / Random Search- Cross-validation
NLP Models: - Multinomial Naive Bayes - Logistic Regression - Decision Tree - Random Forest - SVM
Evaluate on validation/test sets: - Precision - Accuracy - F1 Score - Confusion Matrix
Model Explainability - LIME - SHAP
Deep Learning Models: - FNN - BERT - BioBERT
Evaluate on validation/test sets: - Precision - Accuracy - F1 Score - Confusion Matrix
Model Explainability - LIME - SHAP

Figure1. Methodology Workflow Diagram

4.1 Experimental Results

We extended our analysis by evaluating five ML classifiers across both the test and validation sets. Performance was assessed using standard evaluation metrics, including accuracy, precision, recall, F1-score, and confusion matrix. Given its balanced consideration of both precision and recall, the F1-score was selected as the primary metric for subsequent comparisons and discussion.

To address the dataset's class imbalance, we applied three commonly used balancing techniques. The results indicated that SMOTE consistently yielded the most stable and improved F1-scores across all classifiers. As other balancing methods did not demonstrate significant improvements, only SMOTE-based results are presented for clarity and relevance. Furthermore, the models were evaluated under three experimental settings based on message-length categories. However, due to a limited number of samples in the long-message category, the third setting was excluded from further analysis. All reported results, therefore, correspond to the short and long message categories under SMOTE-balanced training.

A. Classical Natural Language Processing using Machine Learning Models

Multinomial Naive Bayes: Model performance varied across categories. In Category 1, MNB achieved moderate F1-scores between 0.69 and 0.70 under SMOTE variants, while the hybrid method slightly underperformed. Category 2 yielded the strongest results, with all resampling methods producing high F1-scores. Notably, SMOTE led to the highest performance, reaching an F1-score of 0.96, supported by excellent precision and recall, highlighting the model's strong capability in distinguishing misinformation in this class.

Logistic Regression: As the result illustrated, in Category 1, the model produced satisfactory results, with F1-scores ranging from 0.82 to 0.79 under SMOTE-based strategies. The hybrid method underperformed slightly. In Category 2, model delivered its most consistent and accurate predictions. All balancing techniques led to high F1-scores, with SMOTE achieving a peak score of 0.98. This was reinforced by high precision and recall, indicating reliable detection of relevant patterns in this class.

Decision Tree: As outlined , Category 1 yielded moderate classification outcomes, with F1-scores between 0.67 and 0.70 across all resampling techniques. While these scores indicate some discriminatory power, they reflect limited generalization across the dataset. In Category 2, the model demonstrated significantly better performance, achieving F1-scores between 0.92 and 0.95 regardless of the resampling strategy. In summary, the Decision Tree model showed reliable performance in well-separated classes but struggled to maintain stability in less clearly defined or imbalanced scenarios.

Random Forest: According to result, Category 1 showed strong performance under SMOTE and under-sampling, with F1-scores ranging from 0.67 to 0.70. These results reflect the model's robustness in capturing subtle distinctions in class-specific word usage. In Category 2, RF excelled across all sampling methods, consistently achieving high F1-scores between 0.92 and 0.95. This stability demonstrates its capacity to detect nuanced misinformation cues in structured categories.

Support Vector Machine: Table 1 illustrates that Category 1 produced moderate classification performance, with F1-scores ranging from 0.71 to 0.78 under SMOTE and under-sampling. In Category 2, SVM exhibited its highest scores. All balancing techniques led to strong outcomes, with F1-scores ranging between 0.97. This indicates excellent alignment between the model's decision boundaries and the underlying class separability in this category.

		MNB	LR	DT	RF	SVM
		F1- Score				
Balanced Dataset (SMOTE)	SHORT	0.69 – 0.70	0.82 – 0.79	0.67 – 0.70	0.81 – 0.85	0.78 – 0.71
	LONG	0.95 – 0.96	0.79 – 0.98	0.95 – 0.92	0.96 – 0.96	0.97 - 0.97

Table 1. Observed metrics from ML Models

B. Deep Learning Models

Feedforward Neural Networks: In this section, we systematically optimize performance by fine-tuning various parameters and hyperparameters, including units, layers, learning rate, batch size, dropout, and optimizers, while initially setting epochs at 30 before repeating the process with 50 epochs to identify the optimal configuration. Table 2 presents the observed performance for the Long Category. It can be seen that F1-Scores reaches a maximum of 0.98 with losses as low as 0.89, significantly outperforming other methods on complex real-world text. The Short Category lags behind (F1 0.57–0.75) due to constrained contextual depth, though it shows improvement over baseline models. Refinements such as reduced batch sizes and lower dropout rates enhance outcomes, yet the identical Test/Validation F1-Scores indicate a need for rigorous validation split analysis to confirm reliability.

Experiment ,	Short Category						Long Category						Learning Rate	Batch Size	Units	Dropout	
	Test			Validation			Test			Validation							
	Precision	Recall	F1- Score	Precision	Recall	F1-Score	Precision	Recall	F1- Factor	Precision	Recall	F1- Score					
Baseline	0.71	0.57	0.63	0.70	0.70	0.70	0.90	0.85	0.88	0.91	0.89	0.90	0.001	64	[64, 64]	0.10	Epoch=30
Higher LR	0.68	0.69	0.68	0.71	0.68	0.69	0.89	0.90	0.89	0.92	0.91	0.91	0.010	64	[64, 64]	0.10	
Smaller Batch	0.75	0.79	0.76	0.70	0.71	0.70	0.92	0.91	0.91	0.94	0.95	0.94	0.001	32	[64, 64]	0.10	
Deeper Model	0.79	0.79	0.79	0.77	0.70	0.73	0.93	0.94	0.94	0.96	0.95	0.95	0.001	64	[128, 64, 32]	0.10	
Lower Dropout	0.49	0.63	0.55	0.76	0.59	0.66	0.88	0.82	0.86	0.90	0.90	0.90	0.001	64	[64, 64]	0.05	
Baseline	0.79	0.69	0.73	0.61	0.78	0.69	0.89	0.90	0.89	0.88	0.92	0.90	0.001	64	[64, 64]	0.10	Epoch=50
Higher LR	0.71	0.70	0.70	0.79	0.60	0.78	0.90	0.89	0.89	0.91	0.89	0.90	0.010	64	[64, 64]	0.10	
Smaller Batch	0.78	0.75	0.76	0.70	0.71	0.70	0.94	0.91	0.93	0.91	0.85	0.88	0.001	32	[64, 64]	0.10	
Deeper Model	0.75	0.70	0.72	0.70	0.75	0.72	0.96	0.90	0.93	0.95	0.90	0.92	0.001	64	[128, 64, 32]	0.10	
Lower Dropout	0.72	0.75	0.73	0.65	0.70	0.68	0.91	0.89	0.90	0.88	0.90	0.89	0.001	64	[64, 64]	0.05	

Table 2. The Performance of FNN

Bidirectional Encoder Representations from Transformers : In this analysis, we explore performance optimization by experimenting with key hyperparameters—such as learning rate, dropout, and batch size—starting with a 5-epoch training cycle, then scaling to 10 epochs to pinpoint the ideal configuration. As shown in Table 3, our BERT outcomes reveal impressive accuracy in identifying medical misinformation. The Long Category achieves peak F1-Scores of 0.94 alongside losses dropping to 0.89, outperforming competitors on intricate real-world texts. The Short Category trails with F1-Scores ranging from 0.79 to 0.85 due to contextual limitations, though it surpasses baseline results. Adjustments like elevated learning rates and fine-tuning enhance performance, yet identical Test/Validation F1-Scores call for a thorough review of validation splits to validate consistency.

	Short Category						Long Category						Dropout	Learning rate	Batch Size	
	Test			Validation			Test			Validation						
	Precision	Recall	F1- Score	Precision	Recall	F1-Score	Precision	Recall	F1- Factor	Precision	Recall	F1- Score				
Cross-Entropy loss	0.70	0.75	0.72	0.68	0.63	0.65	0.93	0.90	0.92	0.91	0.89	0.90	0.1	2e-5	8	Epoch = 3
	0.76	0.79	0.75	0.78	0.73	0.74	0.93	0.90	0.91	0.95	0.95	0.95	0.1	2e-5	16	
	0.69	0.65	0.68	0.70	0.65	0.68	0.91	0.89	0.89	0.88	0.92	0.91	0.1	3e-5	8	
	0.56	0.49	0.52	0.57	0.60	0.58	0.83	0.84	0.83	0.81	0.85	0.83	0.2	2e-5	8	
	0.81	0.74	0.77	0.83	0.79	0.80	0.93	0.88	0.90	0.95	0.90	0.92	0.1	2e-5	8	Epoch = 5
	0.79	0.71	0.75	0.75	0.68	0.71	0.98	0.97	0.97	0.95	0.94	0.94	0.1	2e-5	16	
	0.61	0.60	0.60	0.61	0.57	0.59	0.90	0.89	0.89	0.91	0.89	0.90	0.1	3e-5	8	
	0.65	0.68	0.66	0.68	0.70	0.69	0.88	0.81	0.85	0.86	0.85	0.85	0.2	2e-5	8	

Table 3. The Performance of BERT

Bidirectional Encoder Representations from Transformers for Biomedical Text: In order to refine the results further we employed hyperparameter tuning such as learning rate, dropout, and batch size. The model was initially trained with a 5-epochs, followed by an extension to 10 epochs to identify the most effective setup. As illustrated in Table 4, BioBERT highlights its potential in detecting medical misinformation. The Long Category achieves top F1-Scores of 0.97 with losses as low as 0.95, outperformed remaining classifiers on complex real-world biomedical texts. The Short Category results with F1-Scores from 0.68 to 0.75 due to limited contextual depth, though it edges past baseline performance. Optimizations like adjusted learning rates and sequence lengths enhance results, but identical Test/Validation F1-Scores necessitate a deeper validation split review to ensure robustness.

,

	Short Category						Long Category										
	Test			Validation			Test			Validation			Dropout	Learning rate	Max sequence length	Batch Size	
	Precision	Recall	F1- Score	Precision	Recall	F1-Score	Precision	Recall	F1-Factor	Precision	Recall	F1- Score					
Cross-Entropy loss Optimizer AdamW	0.67	0.69	0.68	0.65	0.63	0.64	0.96	0.97	0.96	0.95	0.96	0.95	0.1	2e-5	512	8	Epoch=5
	0.74	0.76	0.75	0.71	0.69	0.70	0.96	0.97	0.96	0.95	0.96	0.95	0.1	2e-5	128	16	
	0.68	0.70	0.69	0.66	0.64	0.65	0.97	0.98	0.97	0.96	0.97	0.96	0.1	3e-5	512	16	
	0.70	0.75	0.73	0.70	0.68	0.69	0.95	0.96	0.95	0.94	0.95	0.94	0.1	2e-5	128	8	Epoch = 3
	0.69	0.71	0.70	0.67	0.65	0.66	0.95	0.96	0.95	0.94	0.95	0.94	0.2	3e-5	128	16	
	0.70	0.72	0.71	0.68	0.66	0.67	0.94	0.95	0.94	0.93	0.94	0.93	0.1	3e-5	128	8	
	0.65	0.67	0.66	0.63	0.61	0.62	0.98	0.98	0.98	0.97	0.98	0.97	0.2	2e-5	128	8	

Table 4. The Performance of BioBert

4.2. Explainable AI Methods

SHapley Additive exPlanations : SHAP is a ML framework that interprets complex model predictions by quantifying each feature's contribution to the output. It is particularly valuable for black-box models like neural networks, which are difficult to interpret. SHAP assigns importance scores to features, making it easier to debug models, choose relevant features, and build trust in their predictions. In text classification tasks like detecting medical misinformation, SHAP helps reveal how specific words impact the model's decisions -- an important step for validating how the model works and spotting potential biases. This is especially useful in datasets with varied message lengths, where context significantly impacts interpretation.

The SHAP summary plots provide insight into the decision-making process, used for distinguishing between real and fake tweets in our dataset. The first plot (Fig 2)corresponds to Category 1, which represents fake tweets. Here, words such as "opinion," "also," "opposite," and "said" are associated with higher SHAP values, indicating a stronger influence in pushing the model's prediction toward the fake class. These words tend to be subjective or general, commonly found in opinion-based content rather than in scientifically grounded statements. Additionally, clinical and scientific terms like "disease," "health," "clinical," and "researcher" either have minimal impact or are negatively associated with fake tweets, meaning their presence in a tweet reduces the likelihood of it being labeled as fake. This suggests that fake tweets often lack the domain-specific language typically found in credible medical sources.

In contrast, the SHAP plot for Category 2, assumed to represent real tweets, highlights the significance of scientific and medical terminology in supporting the model's predictions. Words such as "vaccine," "cancer," "covid19," "clinic," "researcher," and "therapy" have high positive SHAP values, meaning they contribute strongly toward classifying a tweet as real. These terms are more frequently used in medically accurate and evidence-based tweets, which aligns with the expectation that real tweets are grounded in professional or clinical discourse. The presence of these domain-specific terms enhances the classifier's confidence in labeling content as real, indicating that the model has successfully learned meaningful patterns that distinguish credible medical information from misinformation.

Local Interpretable Model-agnostic Explanations: LIME is a technique used to explain the predictions of ML models by approximating their behavior locally with simpler, interpretable models. It works by perturbing the input data around a specific instance and observing how these changes affect the model's predictions, thereby identifying which features—such as words in text data—most influence the outcome for that instance. The LIME analysis for Test Example 1, Fig. 3, shows a strong prediction probability of 0.91 for Category 1 and 0.09 for Category 2. Key terms like "fdric brocard," "institut de neuroscience," "discovered," "hyperexcitability," "deregulation," and "sodium channel neuron" heavily influence the classification toward Category 1, indicating a strong association with credible, research-based content. For Test Example 2, the model assigns a 0.79 probability to Category 1 and 0.21 to Category 2, with "vitamin," "greater," "low blood vitamin level," and "reduced kidney function" driving the prediction, reinforcing the link to evidence-based health information.

Overall, the LIME analysis highlights that Models depend on domain-specific terminology to differentiate between categories. For Category 1, likely representing real tweets, the presence of scientific and medical terms boosts the model's confidence in labeling content as credible. In contrast, Category 2, likely associated with fake tweets, shows minimal influence from these terms, suggesting that the absence or reduced use of specialized language is a key indicator of misinformation. This pattern underscores the model's effectiveness in identifying authentic medical discourse based on linguistic cues.

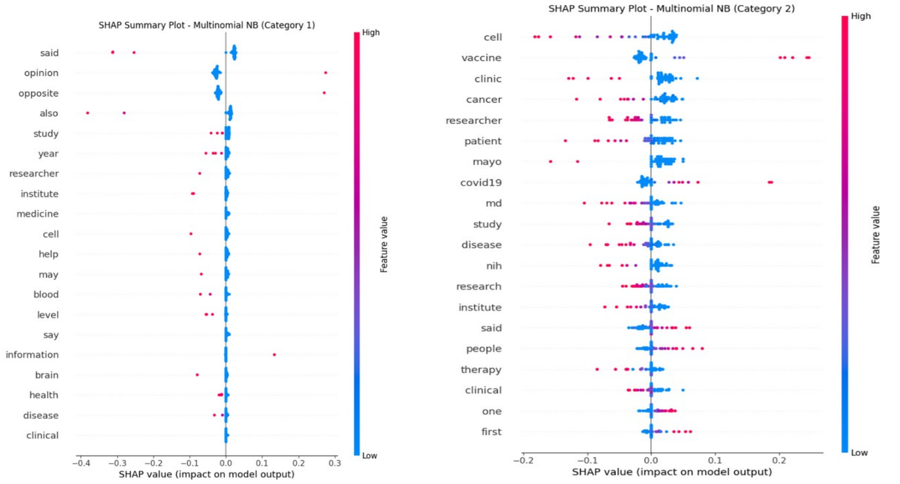


Figure 2. SHAP Plot Example

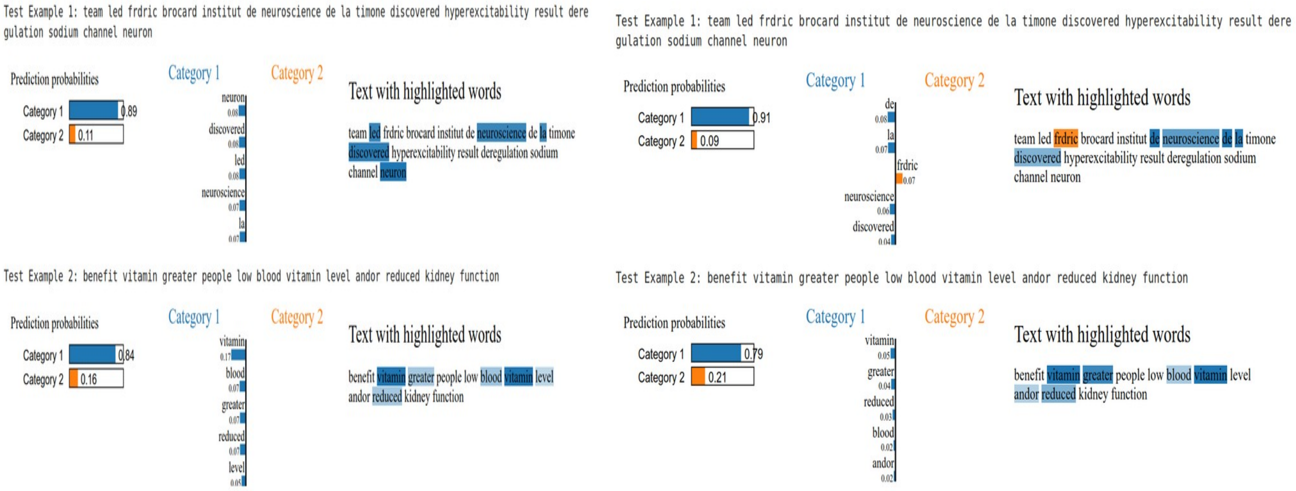


Figure 3. LIME Plot Example

5. Conclusion

Our results showed that combining data balancing, message-length categorization, and advanced models like FNN, BERT, and BioBERT led to strong improvements in detecting medical misinformation. BioBERT outperformed all other models, especially on long texts (F1 up to 0.98), while short texts remained challenging with lower F1-scores (0.68–0.75), mainly due to limited context. Classical models like Logistic Regression and SVM performed well in certain cases, especially with SMOTE, but struggled in more ambiguous categories. SHAP and LIME helped us interpret predictions, revealing that general or emotional words in short tweets often misled the models. Scientific and domain-specific terms, on the other hand, were strong indicators of credibility.

To improve short-text performance, we suggest targeted preprocessing to filter misleading terms and a refined sampling strategy, followed by re-training.

Future work

Research could focus on enhancing short-text classification by incorporating context-aware preprocessing, such as phrase disambiguation or semantic filtering. Leveraging multi-modal inputs -- including metadata and image features -- may also enrich the signal for better performance. Additionally, techniques like domain-adaptive pretraining, contrastive learning, and curriculum learning can help models distinguish subtle misinformation patterns. Evaluating models under adversarial and real-world noisy conditions would further ensure reliability in practical deployment.

Acknowledgements

zahra.M: is grateful to the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research) for funding through project OIDLITDSM, No. 01IS24061.

References

[1] Y. Sun, J. He, S. Lei, L. Cui, and C.-T. Lu, "Med-MMHL: A Multi-Modal Dataset for Detecting Human- and LLM-Generated Misinformation in the Medical Domain," Jun. 2023, [Online]. http://arxiv.org/abs/2306.66571

[2] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Med-HALT: Medical Domain Hallucination Test for Large Language Models," Sep. 2023, [Online]. Available: http://arxiv.org/abs/2307.15343

[3] G. Mehenni and A. Zouaq, "MedHal: An Evaluation Dataset for Medical Hallucination Detection," Apr. 2025, [Online]. https://arxiv.org/abs/2504.08596

[4] J. Chen, D. Yang, T. Wu, Y. Jiang, X. Hou, M. Li, S. Wang, D. Xiao, K. Li, and L. Zhang, "Detecting and Evaluating Medical Hallucinations in Large Vision Language Models," Jun. 2024, [Online]. http://arxiv.org/abs/2406.10185

[5] S. Tian, Q. Jin, L. Yeganova, P.-T. Lai, Q. Zhu, et al., "Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health," Jun. 2023, [Conference Proceedings].

[6] Y. Sun, J. He, S. Lei, L. Cui, and C.-T. Lu, "Silver Lining in the Fake News Cloud: Can Large Language Models Help Detect Misinformation?," IEEE Trans. Artif. Intell., vol. 1, no. 1, pp. 1–12, Jan. 2025, [Online]. https://www.computer.org/csdl/journal/ai/2025/01/10631663/1ZfjnHxxsf6

[7] "An Enhanced Multimodal Multilingual Dataset for Medical Misinformation Detection," in Proc. IEEE Int. Conf. Healthcare Informatics (ICHI) Doctoral Consortium, 2024, [Online]. https://www.zhaoyisun.com/assets/files/ICHI2024DoctoralConsortium.pdf

[8] L. Cui, "Google Scholar Profile: Research Contributions on Medical Misinformation and Dataset Development," [Online]. Available: https://scholar.google.com/citations?hl=en&user=pfd4pUkAAAAJ

[9] J. He, "Google Scholar Profile: Research Contributions on Misinformation Detection and Dataset Creation," [Online]. https://scholar.google.com/citations?hl=en&user=agAf96sAAAAJ

[10] S. Lei, "Google Scholar Profile: Research Contributions on Multimodal Misinformation Detection," [Online]. https://scholar.google.com/citations?hl=en&user=vm368LkAAAAJ

[11] H. Li, Q. Zhang, and S. Patel, "Combating Medical Misinformation with Transformer-Based Models: A Multi-Modal Approach," 2024.

[12] J. Kim and T. Nguyen, "Explainable AI for Health Misinformation Detection: A SHAP-Based Analysis," 2023.

[13] R. Gupta, A. Singh, and V. Kumar, "Addressing Class Imbalance in Medical Misinformation Datasets: Novel Resampling Techniques," 2025.

[14] Y. Zhou and M. Liu, "Linguistic Patterns in AI-Generated Medical Misinformation: Detection and Mitigation," 2024.

[15] N. Patel and R. Sharma, "Multi-Modal Misinformation Detection in Healthcare: Leveraging Text and Metadata with LLMs," 2023.

[16] L. Chen and X. Wang, "Robustness of Large Language Models Against Medical Misinformation Attacks," 2025.

[17] M. Hassan and K. Ali, “Deep Learning for COVID-19 Misinformation: A Comparative Study of Model Architectures,” 2024.

[18] E. Torres and P. Gomez, “Interpretable Detection of Health Misinformation Using Attention Mechanisms,” 2023.

[19] P. Singh and R. Jain, “Scalable Machine Learning Solutions for Medical Misinformation in Social Media,” 2025.

[20] T. Wu and S. Lee, “Hybrid Models for Medical Misinformation Detection: Combining LLMs and Knowledge Bases,” 2024.

2nd Sorbonne-Heidelberg Workshop on AI in Medicine

Dȩep Spatio-temporal Learning in fMRI Sequence Prediction for Alzheimer's Disease

Ahmed Alshembari[a], Anima Kujur[a], Zahra Monfared[a]

[a] *Interdisciplinary Center for Scientific Computing (IWR), Department of Mathematics and Computer Science, Heidelberg University, Heidelberg, 69120, Germany*

Abstract

In this study we introduce a deep Autoregressive (AR) framework for predicting the next temporal representation of resting-state brain function in Alzheimer's Disease (AD) patients, capturing evolving spatial-temporal patterns from functional magnetic resonance imaging (fMRI) sequences in AD patients. Addressing the limitations of traditional linear AR models and standard Recurrent Neural Networks (RNNs) based architectures, we implement a spatio-temporal deep learning model designed to preserve spatial coherence and temporal dynamics throughout the predictive pipeline. Our proposed architecture utilizes time-distributed convolutional blocks followed by temporal sequence modeling and progressive spatial reconstruction, enabling high-fidelity prediction of the next step sequences in fMRI. In contrast to Convolutional Neural Network + Long Short-Term Memory (CNN+LSTM) setups, which vectorize spatial features before temporal modeling – leading to increased parameter complexity and loss of 2D spatial coherence – our proposed ConvLSTM2D approach embeds convolutional operations directly within LSTM units. This integration preserves 2D spatial structure, reduces computational cost, and enhances prediction performance. A custom loss function combining Mean Squared Error (MSE) and Structural Similarity (SSIM) further reinforces structural accuracy. Grid search optimization reveals that deeper convolutional filters and dual sequence layers yield superior performance. Cross-validation confirms robustness across subjects, following best practices for model validation in neuroimaging, and interpretability analysis shows alignment with brain regions affected in early AD. This AR learning framework not only advances predictive modeling in neuroimaging but also holds promise for early biomarker identification and progression monitoring in clinical AD research.

Keywords: , LSTM, fMRI, Alzheimer Disease, CNN, Deep Learning

1. Introduction

Alzheimer's Disease (AD), a leading cause of dementia, is a progressive brain disorder marked by memory loss, cognitive decline, and neural damage. Key biological features, such as amyloid-beta plaques and tau tangles, disrupt brain networks like the hippocampus and default mode network (DMN) [1, 2]. Early detection of these changes is essential for timely intervention and slowing disease progression [3].

fMRI offers a non-invasive way to study brain activity through blood-oxygen-level-dependent (BOLD) signals [4]. Static functional connectivity methods reveal network disruptions in AD but assume steady activity, missing dynamic changes over time [5]. Dynamic Functional Connectivity (dFC) methods improve temporal analysis but are limited by fixed window sizes and noise sensitivity [6].

Deep learning has advanced functional Magnetic Resonance Imaging (fMRI) analysis by capturing complex patterns. Convolutional Neural Network (CNNs) extract spatial features, while Long Short-Term Memory (LSTM) networks model temporal changes [7, 8, 9, 10]. However, these models, when used separately, struggle to capture the combined spatial and temporal dynamics critical for understanding AD progression [11].

To address this, we propose a ConvLSTM2D-based Autoregressive (AR) framework [12], which integrates convolutional and LSTM operations to model spatial and temporal patterns in fMRI sequences simultaneously. We compare this model against a CNN+LSTM hybrid [13, 14], optimizing performance with a custom loss function combining Mean Squared Error (MSE), Structural Similarity (SSIM), Mean Absolute Error (MAE), and Peak Signal-to-Noise Ratio (PSNR). Grid search and cross-validation ensure robust performance.

This framework enhances fMRI-based analysis of AD, facilitating both early diagnosis and long-term monitoring. Section 2 provides a review of prior research on fMRI and deep learning methods; Section 3 outlines the datasets and methodological approaches; Section 4 reports the experimental results and analysis; and Section 5 concludes with a summary of key findings and directions for future research.

2. Related Work

Early fMRI Analysis of AD: Early fMRI studies on AD employed static functional connectivity methods, such as seed-based correlations and independent component analysis (ICA), to identify disruptions in networks like the DMN [2]. These methods revealed reduced connectivity in AD patients but failed to capture dynamic neural activity. For instance, [15] observed inconsistent connectivity patterns across subjects. Dynamic functional connectivity (dFC) methods, such as sliding window correlations, addressed some temporal variability but were limited by parameter selection and noise sensitivity [6].

Deep Learning in AD fMRI: Deep learning has significantly advanced fMRI analysis for AD. CNNs have been effective in extracting spatial features from 3D fMRI volumes. For example, [16] applied a CNN to resting-state fMRI data, achieving 96.86% accuracy in classifying AD patients versus controls. Additionally, [17] proposed a deep learning framework combining resting-state fMRI and structural MRI, achieving an AUC of 85.12 for AD classification. These approaches demonstrate robust performance in distinguishing AD stages but are primarily designed for classification, not sequence prediction, which is critical for modeling AD progression.

Hybrid CNN-LSTM Models in AD fMRI: Hybrid CNN-LSTM architectures integrate spatial and temporal features for AD analysis. For instance, [18] proposed a multi-modal CNN-LSTM for AD classification with $\approx$ 86% accuracy, but its static setup – vectorizing fMRI features before LSTM – limits future fMRI sequence prediction by losing spatial-temporal coherence. Similarly, [19] employed a 3D-CNN-LSTM model to classify AD progression across cognitively normal, mild cognitive impairment, and AD stages, reporting high classification accuracy. Furthermore, [20] proposed a 3D-CNN and bidirectional LSTM framework for 4D fMRI data, achieving 94.82% accuracy in AD classification. However, these methods focus on classification tasks, limiting their ability to perform integrated spatial-temporal forecasting needed for understanding long-term AD dynamics, which our study addresses.

ConvLSTM2D for AD fMRI Sequence Prediction: ConvLSTM2D, which embeds convolutional operations within LSTM units, has shown promise in video prediction

[12], and in limited fMRI classification [21]. However, its application to AR fMRI sequence prediction in AD remains underexplored. Our study leverages ConvLSTM2D to enable integrated spatial-temporal modeling for predicting AD fMRI images, overcoming the limitations of prior classification-focused approaches. Optimized for AD-specific brain dynamics, our framework offers a novel tool for neurodegenerative research, as detailed in Section 1.

3. Methodology

This section outlines the methodology employed in this study, as depicted in the flow diagram in Figure 1. The methodology encompasses four key stages: data collection, preprocessing, model architectures, and performance evaluation. Detailed results and further analyses are illustrated in Section 4.

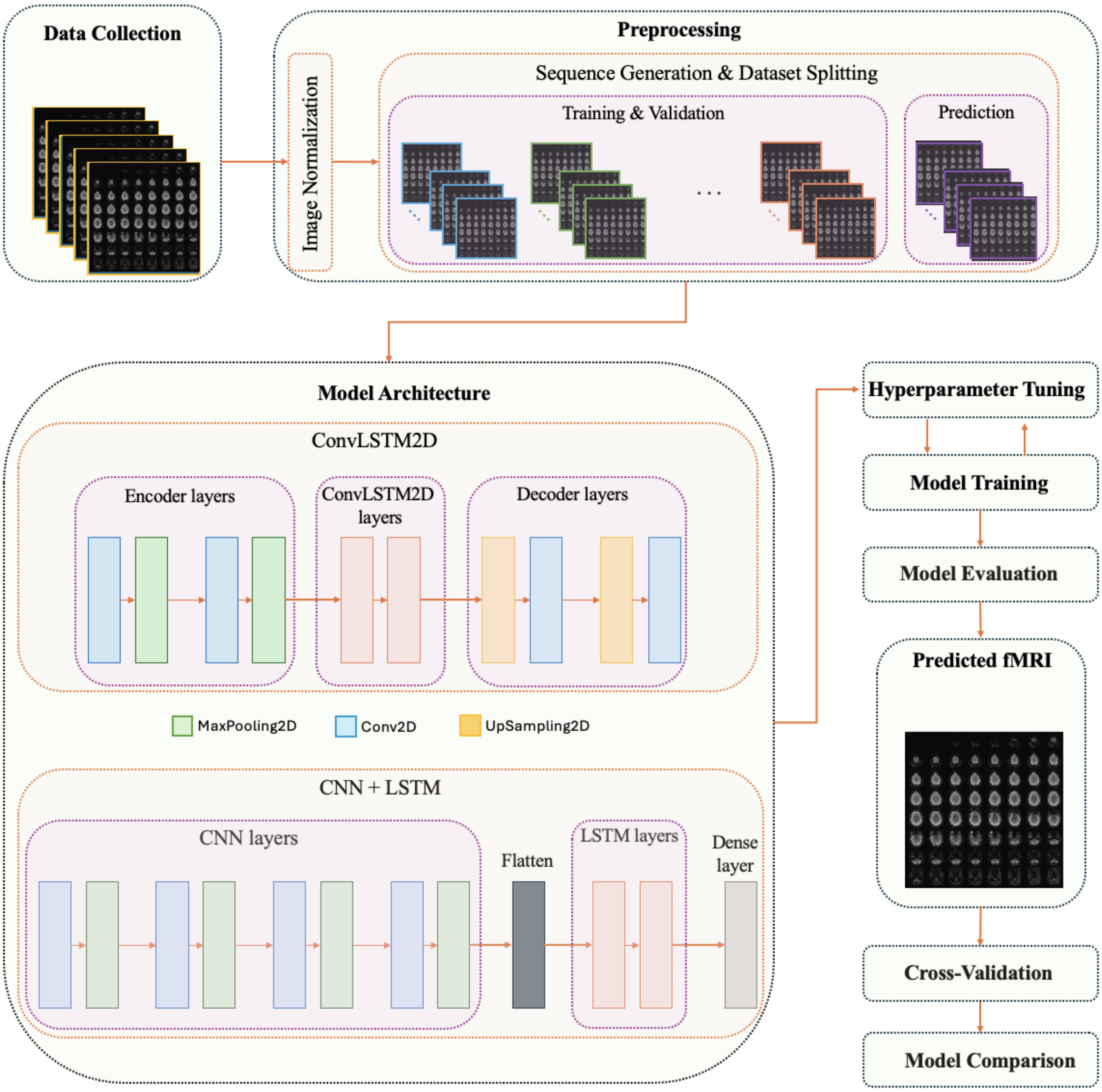


Figure 1: The Flow Diagram of the proposed methodology.

3.1. Data Description

This study uses resting-state fMRI data (eyes open) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`http://adni.loni.usc.edu`, accessed 31 October 2024). The dataset consists of 2026 3D fMRI volumes from five AD patients, labeled P1 through P5, capturing brain activity over time with varying numbers of time points per patient. Specifically, P1, P2, and P3 each have 482 volumes; P4 has 369; and P5 has 211. Each 3D volume is reformatted into a 2D (704×704-pixel) image by arranging axial slices in a grid, as shown in Fig. 2, and represented as $\mathbf{X} \in \mathbb{R}^{T\times 704\times 704}$, where T is the number of time points. These images form the foundation of this study. The dataset includes participants of varying sexes and ages, providing a sample to investigate AD progression.

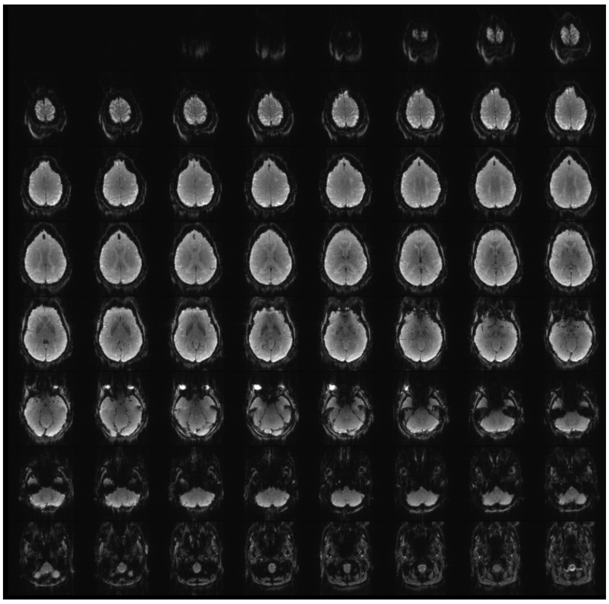


Figure 2: 2D fMRI image formed from axial slices in a grid.

3.2. Preprocessing

The preprocessing pipeline was designed to prepare the data for AR models by normalizing voxel intensities, structuring sequences to capture temporal dependencies, and splitting the data.

3.2.1. Image Normalization

To address variability in intensity scales across scans, each fMRI image was normalized by dividing its pixel intensities by the maximum intensity value within the image, scaling all values to the range [0, 1], using the formula I^{norm} = I / max(I), where I represents the original pixel intensity. This normalization ensures consistent intensity ranges across participants, improving the model's ability to learn and compare brain activity patterns effectively.

3.2.2. Sequence Generation for Temporal Modeling

We used a sliding-window approach to generate sequences for temporal modeling using normalized images. Each sequence consists of 10 consecutive images as input, denoted as $X_j = [I_{\mathrm{j}}^{norm}, \ldots, I_{\mathrm{j+9}}^{norm}] \in \mathbb{R}^{10\times 704\times 704}$, with the 11th image as the target label

$Y_j = I_{\mathrm{j+10}}^{norm} \in \mathbb{R}^{704\times704}$. Here, I_j^{norm} represents the normalized fMRI image at time step j. The index j ranged from 1 to M, where $M = N - 10$, and $N = 2026$ is the total number of time steps that result in sequences $M = 2016$. Preserving the temporal order of these sequences is critical for prediction, and the chosen sequence length balances temporal context with computational efficiency.

3.2.3. Dataset Splitting

Before data splitting, a single sequence of 10 images was set aside as a held-out prediction set for final evaluation, as illustrated in Figure 1. This approach ensured that model generalization could be assessed on truly unseen data. The remaining sequences were divided into 80% for training and 20% for validation, with balanced representation maintained across both subsets. To preserve the temporal dependencies critical for sequential modeling, shuffling was disabled during the split. Model training was performed using a batch size of 1 (i.e., one sequence per batch), due to GPU memory constraints associated with processing high-dimensional fMRI data. Training proceeded for 100 epochs to encourage convergence and stability. The training set was used to fit the model and optimize its parameters, while the validation set supported performance monitoring and helped prevent overfitting.

3.3. Model Architectures

This study employs two deep learning models to predict the fMRI image: ConvLSTM2D and CNN+LSTM. ConvLSTM2D integrates convolutional operations within the LSTM framework, effectively preserving both spatial and temporal features. In contrast, CNN+LSTM processes spatial features through convolutional layers followed by temporal feature extraction with LSTM layers, which limits integrated spatio-temporal modeling. The ConvLSTM2D architecture outperformed other models in our experiments, as detailed in Section 4, making it the focus of this study.

3.3.1. ConvLSTM2D Architecture

The ConvLSTM2D model employs an encoder-decoder framework. The encoder applies TimeDistributed Conv2D layers to extract spatial features from each input sequence frame, followed by MaxPooling2D layers to downsample and reduce computational complexity. The core includes one or more ConvLSTM2D layers, extending traditional LSTMs with convolutional operations, as defined by:

$$\begin{aligned}
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + b_i)\\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + b_f)\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)\\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + b_o)\\
H_t &= o_t \odot \tanh(c_t)
\end{aligned}$$

where $*$ denotes convolution, X_t is the input at time t, H_t is the hidden state, and c_t is the cell state. The decoder uses UpSampling2D layers to restore spatial dimensions and a Conv2D layer to predict frames with the same dimension as the input. The model iteratively predicts frames by feeding outputs as inputs.

Training uses a custom loss function combining MSE and SSIM, defined as $\mathcal{L} = \alpha \cdot \mathrm{MSE} + (1-\alpha)\cdot(1-\mathrm{SSIM})$. Here, $\alpha \in [0,1]$ balances the contributions of MSE (pixel-wise accuracy) and SSIM (structural fidelity), tuned empirically to optimize performance for AD fMRI sequences, as detailed in Section 4.1.

3.4. Performance Metrics

Model performance is evaluated using four established metrics: MSE, MAE [22], SSIM [23], and PSNR [24].

The MSE measures pixel-level accuracy by computing the average squared difference between predicted ($\hat{y}_i$) and actual (y_i) values:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The MAE quantifies the average absolute error, providing a robust measure of prediction error magnitude:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

The SSIM assesses perceptual similarity between images, with values ranging from -1 (dissimilar) to 1 (identical):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

The PSNR evaluates image quality in decibels, where higher values indicate superior fidelity:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right)$$

4. Results and Discussion

4.1. Hyperparameter Optimization via Grid Search

Hyperparameter tuning was performed to optimize the ConvLSTM2D model using a grid search over parameters including convolutional depth, number of filters, ConvLSTM2D units, and number of layers. These parameters significantly influence the model's ability to capture spatio-temporal patterns in fMRI data.

To ensure consistent spatial feature extraction, all convolutional layers used a fixed filter size of $(3, 3)$, and max-pooling layers adopted a pool size of $(2, 2)$. Intermediate layers utilized ReLU activation functions, while the final output layer employed a sigmoid activation to constrain predictions within the normalized range of $[0, 1]$.

The grid search examined convolutional depths of 1 to 4, convolutional filter counts of $\{16, 32, 64\}$, ConvLSTM2D units of $\{64, 128, 256, 512\}$, and ConvLSTM2D layers of $\{1, 2\}$. Each configuration was assessed using a custom loss function for ConvLSTM2D, with the optimal setup determined by the lowest validation set loss. For comparison, a baseline CNN+LSTM model was optimized using MSE loss under similar grid search conditions.

Table 1 presents the optimal hyperparameter configurations for both models. The ConvLSTM2D model achieved a validation loss of 0.1316 using the MSE+SSIM loss function, while the CNN+LSTM model achieved an MSE of 3.9e-3, optimized for capturing spatio-temporal dynamics in AD fMRI sequence prediction.

The grid search showed that convolutional depths beyond 2 gave diminishing returns for ConvLSTM2D, while a depth of 4 performed better for the CNN+LSTM model. Using two ConvLSTM2D layers effectively captured long-term patterns in fMRI data, making the model suitable for AD image prediction. These optimized hyperparameters form the basis for our models evaluations and comparisons.

,Table 1: Optimal Hyperparameter Configurations for ConvLSTM2D and CNN+LSTM Models

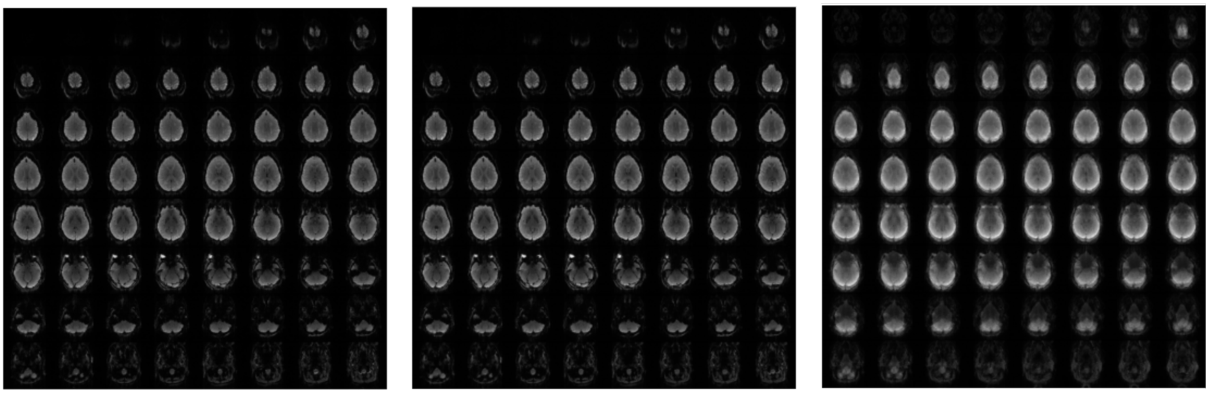
Parameter	**ConvLSTM2D**	**CNN+LSTM**
Convolutional Depth	2	4
Number of Filters	64	64
LSTM Units	512	256
Number of Layers	2	2
Learning Rate	1e-4	1e-3
Filter Size	(3,3)	(3,3)
Pool Size	(2,2)	(2,2)
Epochs	100	50
Batch Size (seq.)	1	1
Optimizer	Adam	Adam

4.2. Final Model Training and Performance Evaluation

This section evaluates the performance of the ConvLSTM2D model to predict next step fMRI frames using the optimal hyperparameters identified through the grid search (Section 4.1). The model was trained on the training data set as described in Section 3.2.3.

On the predict sequence reserved during dataset splitting (Section 3.2.3) and generated using the sliding-window approach (Section 3.2.2), the ConvLSTM2D model achieved high predictive accuracy and structural fidelity, with a MSE of 0.00028, MAE of 0.0082, SSIM of 0.9621, and PSNR of 35.8921.

In comparison, the baseline CNN+LSTM model, trained with its optimal hyperparameters, yielded an MSE of 0.0032, MAE of 0.0286, SSIM of 0.7400, and PSNR of 24.9910. These results underscore the superior ability of the ConvLSTM2D model to capture spatio-temporal patterns iteratively, critical for AR prediction in AD research. Higher SSIM and PSNR values indicate better preservation of structural details and image quality, as visually confirmed in Figure 3, which illustrates the close similarity between predicted and ground truth fMRI frames.



(a) Actual fMRI Image

(b) ConvLSTM Prediction

(c) CNN+LSTM Prediction

Figure 3: Comparison of Actual and Predicted fMRI Images

4.3. Cross-Validation Results

To evaluate the robustness and generalizability of the proposed AR ConvLSTM2D model for predicting the next step fMRI image, a 5-fold cross-validation approach was employed using the optimal hyperparameters identified through grid search (Section 4.1). Performance was assessed using MSE, MAE, SSIM, and PSNR, with metrics averaged across all folds. The results are summarized in Table 2.

The ConvLSTM2D model achieved low MSE and MAE, indicating high pixel-level accuracy, and high SSIM, reflecting excellent structural fidelity critical for AD fMRI sequences. Its high PSNR confirms low-noise predictions. In contrast, the CNN+LSTM model exhibited higher errors and lower structural similarity and image quality, underscoring ConvLSTM2D's superior spatiotemporal modeling due to its convolutional LSTM layers. These findings, detailed in Table 2, highlight the model's effectiveness for Alzheimer's research.

Table 2: Average Performance Metrics from 5-Fold Cross-Validation

Metric	**MSE**	**MAE**	**SSIM**	**PSNR**
ConvLSTM2D	0.0003	0.0086	0.9609	35.5123
CNN+LSTM	0.0179	0.0655	0.4210	17.9718

5. Conclusion

This contribution introduces a novel AR framework utilizing a ConvLSTM2D model to predict next brain state in AD patients using resting-state fMRI data. By integrating convolutional operations for spatial patterns and recurrent operations for temporal dependencies, the model effectively captures the spatio-temporal dynamics critical for understanding AD progression.

The ConvLSTM2D model was optimized through grid search, identifying an optimal configuration with 2 ConvLSTM2D layers, 64 filters, and 512 LSTM units. Evaluated via 5-fold cross-validation and on a reserved predict subsequence from one patient, it demonstrated outstanding predictive accuracy and anatomical fidelity, significantly outperforming the baseline CNN+LSTM model, as detailed in Table 5.

These results underscore the ConvLSTM2D model's superior ability to model long-term spatiotemporal dependencies, significantly outperforming the baseline CNN+LSTM model, which exhibited higher error rates and lower structural fidelity. This advancement highlights the value of an integrated architecture for AR forecasting in neurodegenerative research.

This work offers a substantive advancement toward early biomarker discovery and the predictive modeling of Alzheimer's progression via accurate forecasting of brain state dynamics. This framework provides a robust tool for research and clinical applications, offering deeper insights into AD's neural dynamics. Future work could extend this approach to larger datasets or incorporate multimodal data to further improve predictive accuracy and clinical utility.

Acknowledgements

Z.M. is grateful to the Bundesministerium für Bildung und Forschung (BMBF, Federal Ministry of Education and Research) for funding through project OIDLITDSM, No. 01IS24061.

References

[1] C. R. Jack, D. A. Bennett, K. Blennow, M. C. Carrillo, B. Dunn, S. B. Haeberlein, et al., Nia-aa research framework: Toward a biological definition of alzheimer's disease, Alzheimer's & Dementia 14 (2018) 535–562.

[2] R. L. Buckner, J. R. Andrews-Hanna, D. L. Schacter, The brain's default network: Anatomy, function, and relevance to disease, Annals of the New York Academy of Sciences 1124 (2008) 1–38.

[3] Y. Yakoub, N. J. Ashton, C. Strikwerda-Brown, L. Montoliu-Gaya, T. K. Karikari, P. R. Kac, F. Gonzalez-Ortiz, J. Gallego-Rudolf, P.-F. Meyer, F. St-Onge, et al., Longitudinal blood biomarker trajectories in preclinical alzheimer's disease, Alzheimer's & Dementia 19 (2023) 5620–5631.

[4] M. D. Greicius, G. Srivastava, A. L. Reiss, V. Menon, Default-mode network activity distinguishes alzheimer's disease from healthy aging: Evidence from functional mri, Proceedings of the National Academy of Sciences 101 (2004) 4637–4642.

[5] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, et al., Dynamic functional connectivity: Promise, issues, and interpretations, NeuroImage 80 (2013) 360–378.

[6] M. G. Preti, T. A. Bolton, D. Van De Ville, The dynamic functional connectome: State-of-the-art and perspectives, NeuroImage 160 (2017) 41–54.

[7] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (1998) 2278–2324.

[8] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[9] H. Sun, A. Wang, S. He, Temporal and spatial analysis of alzheimer's disease based on an improved convolutional neural network and a resting-state fmri brain functional network, International Journal of Environmental Research and Public Health 19 (2022) 4508.

[10] J. Li, B. Song, C. Qian, Diagnosis of alzheimer's disease by feature weighted-lstm: a preliminary study of temporal features in brain resting-state fmri, Journal of Integrative Neuroscience 21 (2022) 56.

[11] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annual Review of Biomedical Engineering 19 (2017) 221–248.

[12] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional lstm network: A machine learning approach for precipitation nowcasting, in: Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 802–810.

[13] R. Wang, Q. He, C. Han, H. Wang, L. Shi, Y. Che, A deep learning framework for identifying alzheimer's disease using fmri-based brain network, Frontiers in Neuroscience 17 (2023) 1177424.

[14], S. Wein, A. Schüller, A. M. Tomé, W. M. Malloni, M. W. Greenlee, E. W. Lang, Forecasting brain activity based on models of spatiotemporal brain dynamics: A comparison of graph neural network architectures, Network Neuroscience 6 (2022) 665–701.

[15] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, V. D. Calhoun, Tracking whole-brain connectivity dynamics in the resting state, Cerebral Cortex 24 (2014) 663–676.

[16] S. Sarraf, G. Tofighi, Deep learning-based pipeline to recognize alzheimer's disease using fmri data, in: 2016 future technologies conference (FTC), IEEE, 2016, pp. 816–820.

[17] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, et al., Development and validation of an interpretable deep learning framework for alzheimer's disease classification, Brain 143 (2020) 1920–1933.

[18] S. Spasov, L. Passamonti, A. Duggento, P. Liò, N. Toschi, A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease, NeuroImage 189 (2019) 276–287.

[19] J.-H. Noh, J.-H. Kim, H.-D. Yang, Classification of alzheimer's progression using fmri data, Sensors 23 (2023) 6330.

[20] C. Feng, A. Elazab, P. Yang, T. Wang, F. Zhou, H. Hu, X. Xiao, B. Lei, Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm, IEEE Access 7 (2019) 63605–63618.

[21] Y. Chen, J. Wang, C. Wang, M. Liu, Q. Zou, Deep learning models for disease-associated circrna prediction: a review, Briefings in bioinformatics 23 (2022) bbac364.

[22] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: Data mining, inference, and prediction, 2 ed., Springer, 2009.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.

[24] R. C. Gonzalez, R. E. Woods, Digital image processing, 3 ed., Pearson, 2008.

,

FractalPINN-Flow: A Fractal-Inspired Network for Unsupervised Optical Flow Estimation with Total Variation Regularization

Sara Behnamian[1], Rasoul Khaksarinezhad[2], and Andreas Langer[3]

[1]Globe Institute, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark, sara.behnamian@sund.ku.dk
[2]rasoul.khaksari@gmail.com
[3]Centre for Mathematical Sciences, Lund University, Box 118, 221 00 Lund, Sweden, andreas.langer@math.lth.se

Abstract

We present FractalPINN-Flow, an unsupervised deep learning framework for dense optical flow estimation that learns directly from consecutive grayscale frames without requiring ground truth. The architecture centers on the Fractal Deformation Network (FDN)—a recursive encoder-decoder inspired by fractal geometry and self-similarity. Unlike traditional CNNs with sequential downsampling, FDN uses repeated encoder-decoder nesting with skip connections to capture both fine-grained details and long-range motion patterns. The training objective is based on a classical variational formulation using total variation (TV) regularization. Specifically, we minimize an energy functional that combines L^1 and L^2 data fidelity terms to enforce brightness constancy, along with a TV term that promotes spatial smoothness and coherent flow fields. Experiments on synthetic and benchmark datasets show that FractalPINN-Flow produces accurate, smooth, and edge-preserving optical flow fields. The model is especially effective for high-resolution data and scenarios with limited annotations.

Keywords: Optical Flow, Unsupervised Learning, Neural Networks, Total Variation, Encoder-Decoder, Motion Estimation

1 Introduction

Optical flow estimation seeks to recover the apparent motion field between two consecutive grayscale images. Let $I_1, I_2 : \Omega \to [0,1]$ denote two discrete images defined on a spatial domain $\Omega \subset \mathbb{Z}^2$, and let $w : \Omega \to \mathbb{R}^2$, $w(x) = (u(x), v(x))$, be the displacement field to be estimated. The fundamental assumption underlying most optical flow methods is brightness constancy, which posits that the intensity of each point remains constant as it moves, i.e.,

$$I_1(x) = I_2(x + w(x)) \quad \text{for all } x \in \Omega. \tag{1}$$

This results in a nonlinear constraint on w. Since the problem is underdetermined (one equation, two unknowns), additional regularization is required.

To simplify the nonlinear data term, many methods approximate the brightness constancy relation (1) by a linear model:

$$I_2(x + w(x)) \approx I_2(x) + \nabla I_2(x) \cdot w(x),$$

where ∇I_2 denotes the image gradient. This yields the approximate linear constraint

$$\nabla I_2(x) \cdot w(x) + I_2(x) - I_1(x) = 0,$$

as used in the classical Horn–Schunck formulation [7]. There the flow field is recovered by minimizing

$$\|\nabla I_2 \cdot w + I_2 - I_1\|_2^2 + \lambda \|\boldsymbol{\nabla} w\|_2^2,$$

where $\lambda > 0$ is a regularization parameter, $\|\boldsymbol{\nabla} w\|_2^2 = \|\nabla u\|_2^2 + \|\nabla v\|_2^2$ encourages global smoothness of the displacement field and $\|\cdot\|_2$ denotes the standard Euclidean norm. This quadratic regularization penalizes rapid variations in the flow but tends to oversmooth motion boundaries.

A more robust alternative replaces the L^2-norm with the L^1-norm in both the data fidelity and smoothness terms, yielding a total variation (TV) regularization model, allowing the flow field to exhibit discontinuities. This leads to the energy

$$\|\nabla I_2 \cdot w + I_2 - I_1\|_1 + \lambda \|\boldsymbol{\nabla} w\|_1,$$

where the last term represents the total variation of the flow field w. In practice, this formulation encourages piecewise-smooth flow fields while preserving motion discontinuities [22]. To efficiently solve TV-regularized optical flow problems, in [5] a primal-dual finite element method was introduced and combined with an iterative warping algorithm to handle large displacements. Building on this in [1, 8] adaptive discretizations schemes were proposed, which can be interpreted as an adaptive multigrid method, further accelerating convergence and sometimes even improving the quality of the estimated flow field.

With the advent of deep learning, optical flow estimation has seen dramatic advances. FlowNet [4] introduced a convolutional neural network (CNN) architecture capable of predicting flow directly from image pairs, opening the door to data-driven solutions, see also [18]. However, supervised methods require large-scale annotated datasets, which are difficult to obtain for many real-world domains. To address this challenge, unsupervised approaches have gained momentum by optimizing photometric consistency and smoothness priors without ground-truth supervision. These methods can be further classified according to the number of input frames used during training, including multiframe models [19], which leverage temporal consistency across sequences, and two-frame models [9, 14], which are based solely on image pairs. Although most of such models incorporate smoothness losses, they typically do not employ TV regularization explicitly.

In this paper, we propose *FractalPINN-Flow*, a novel unsupervised framework that integrates an encoding-decoding strategy with total variation regularization to estimate dense optical flow from grayscale image sequences. Our use of TV minimization in a neural setting is motivated by recent work such as DeepTV [12], which demonstrated how classical regularization techniques can be effectively combined with deep architectures to produce sharp, structurally coherent outputs. At the core of our model is a *Fractal Deformation Network (FDN)*, a recursive encoder-decoder architecture inspired by the self-similar principles of FractalNet [13]. Unlike conventional hierarchical CNNs, which apply sequential downsampling, the FDN recursively nests encoder-decoder modules at multiple scales, each equipped with skip connections. This design builds a deep multiscale representation that maintains local texture while capturing long-range deformation structures, properties that are especially important in scenes with fine motion patterns or limited training data. The FDN output is passed to a lightweight CNN that predicts the optical flow field. Training is fully unsupervised, uses only two frames and is guided by a composite loss function that consists of a combined L^1/L^2 data term and TV regularization.

2 FractalPINN-Flow: Architecture and Implementation

We consider the problem of learning the optical flow from two input images I_1, I_2 by minimizing

$$E_{\mathrm{TV}}(w) := \lambda_1 \|\nabla I_2 \cdot w + I_2 - I_1\|_1 + \lambda_2 \|\nabla I_2 \cdot w + I_2 - I_1\|_2^2 + \lambda_{\mathrm{TV}} \|\boldsymbol{\nabla} w\|_1, \tag{2}$$

where $\lambda_1, \lambda_2, \lambda_{\mathrm{TV}} \geq 0$, w is represented by a neural network, see Section 2.1 for a detailed description of the network structure, and

$$\|\boldsymbol{\nabla} w\|_1 := \|\nabla_x u\|_1 + \|\nabla_y u\|_1 + \|\nabla_x v\|_1 + \|\nabla_y v\|_1 \tag{3}$$

is the anisotropic TV. Here, ∇_x and ∇_y denote the finite differences along the horizontal and vertical axes, respectively. This regularization penalizes abrupt discontinuities while preserving meaningful motion boundaries, thereby promoting spatially coherent and piecewise-smooth flow estimates. We note that while the gradients in (3) could also be implemented in a pointwise manner, as is common in physics-informed neural networks [16], we refrain from doing so following the rationale in [12].

,

In particular, pointwise evaluation may completely miss jump discontinuities, especially in piecewise constant outputs, by sampling points where the gradient happens to vanish. In contrast, a finite difference discretization reliably captures such variations and provides a more accurate measure of total variation.

We note that a combined L^1/L^2 data fidelity term together with total variation regularization, as used in (2), was first introduced in [6] and analyzed in [11] in the context of variational image restoration. We consider this formulation here for optical flow estimation.

2.1 Neural Network

We propose a deep learning architecture for dense optical flow estimation that integrates our novel FDN with a CNN-based optical flow prediction head. The method estimates pixel-wise motion between two consecutive frames by leveraging multi-scale fractal features.

2.1.1 Network Architecture

High-Level Structure. The architecture of the neural networks we use follows a sequential flow of processing stages, outlined as follows:

Input: Two consecutive grayscale frames (I_1, I_2)
$\downarrow$
FDN: Encoder-decoder network
$\downarrow$
Projection Layer: 32 → 64 channels
$\downarrow$
Optical Flow Network: CNN with 5 layers
$\downarrow$
Output: Dense optical flow field (u, v)

The FDN encodes multi-scale motion features, which are then mapped to dense flow fields by a convolutional regression head. This modular structure mirrors successful encoder-decoder designs widely used in optical flow [20, 21] and image registration [3].

Fractal Deformation Network (FDN). The FDN is based on a symmetric U-Net-style encoder–decoder architecture with configurable depth d. In this work, we fix the depth to $d = 4$ without hyperparameter tuning. Each downsampling (encoder) block consists of two 3×3 convolutional layers, each followed by batch normalization and ReLU activation, and concludes with a 2×2 max pooling operation. For $d = 4$, the channel configuration expands as

$$\texttt{2} \rightarrow \texttt{32} \rightarrow \texttt{64} \rightarrow \texttt{128} \rightarrow \texttt{256}$$

enabling progressive abstraction of features at increasingly coarser spatial resolutions.

The overall architecture is designed to promote multi-scale feature extraction through this hierarchical structure. Although we refer to it as a Fractal Deformation Network for consistency with the naming of our framework, the term ”fractal” here is used loosely to suggest repeated block-level processing across scales, rather than strict self-similarity or recursion. This design allows the network to capture both fine-scale deformations and large displacements efficiently.

The decoder mirrors the encoder and progressively increases spatial resolution using 2×2 transposed convolutions. These operations serve as learned upsampling layers: they first insert zeros between pixels (to increase spatial resolution) and then apply a learnable 2×2 kernel to interpolate meaningful values, while also reducing the number of channels by half (e.g., 256 → 128). The overall flow of channel dimensions of the decoder is given by:

$$\texttt{256} \rightarrow \texttt{128} \rightarrow \texttt{64} \rightarrow \texttt{32} \rightarrow \texttt{32.}$$

To preserve high-resolution details, skip connections are added from the encoder to decoder at each intermediate resolution level (excluding the input and final output layers, as their channel dimensions differ and cannot be connected directly). These connections use bilinear interpolation of the encoder feature maps to match the size of the decoder features, followed by element-wise addition. After this fusion, each decoder block includes two 3 × 3 convolutions, each followed by batch normalization and ReLU activation, to refine the combined features. To ensure compatibility with the downsampling operations, the input images are zero-padded, maintaining consistent spatial dimensions throughout the network. Unlike classical U-Net models that concatenate features, our architecture uses element-wise addition instead of concatenation, preserving multiscale information while reducing the number of trainable parameters. The final output of the decoder and thus the FDN is a 32-channel feature map matching the input spatial size, which is passed to the subsequent module. This higher dimension (32-channel) representation allows the network to encode more complex information about the image differences and potential motion cues than a lower-dimensional (e.g., 2-channel) feature representation could encode.

Projection Layer To adapt the FDN output to the optical flow predictor, a 1×1 convolution projects 32 channels to 64. This shallow transformation learns to emphasize motion-relevant features while maintaining full spatial resolution. It ensures architectural compatibility without adding significant computational burden.

Optical Flow Prediction Network This subnetwork is a compact CNN that predicts dense flow fields. It consists of five 3 × 3 convolutional layers with ReLU activations and no pooling. The channel pattern follows:

$$64 \rightarrow 128 \rightarrow 256 \rightarrow 128 \rightarrow 64 \rightarrow 2$$

No downsampling is performed, preserving exact spatial correspondence between features and motion vectors. The final layer has two output channels without ReLU activations, representing horizontal and vertical displacement per pixel. The use of a bottleneck architecture enables expressive mapping while retaining efficiency.

End-to-End Information Flow The full architecture forms a feedforward pipeline:

$$\begin{aligned}
\text{Input Pair: } & I = [I_1, I_2] \in \mathbb{R}^{B\times 2\times H\times W} \\
\text{FDN Output: } & F \in \mathbb{R}^{B\times 32\times H\times W} \\
\text{Projected Features: } & P \in \mathbb{R}^{B\times 64\times H\times W} \\
\text{Predicted Flow: } & w \in \mathbb{R}^{B\times 2\times H\times W}
\end{aligned}$$

where B denotes the batch size and $H \times W$ the number of pixels in the images I_1 and I_2. This separation of concerns ensures that multiscale feature learning and dense motion regression are individually optimized yet trained jointly.

3 Numerical Evaluation

3.1 Evaluation Metrics

Average Endpoint Error (AEE) The AEE metric quantifies the Euclidean distance between the predicted and ground truth flow vectors across all pixels:

$$\text{AEE} = \frac{1}{N}\sum_{i=1}^{N} \left\| w_{\text{pred}}^{(i)} - w_{\text{gt}}^{(i)} \right\|_2,$$

where $w_{\text{pred}}^{(i)}$ and $w_{\text{gt}}^{(i)}$ denote the i-th pixel of the predicted and ground truth flow vectors w_{pred} and w_{gt}, respectively, and N is the number of pixels.

Average Angular Error (AAE) To assess directional consistency, the AAE metric measures angular discrepancies between predicted and ground truth flow vectors:

,

$$\mathrm{AAE} = \frac{1}{N}\sum_{i=1}^{N} \arccos\left(\cos(\theta^{(i)})\right), \quad \cos(\theta) = \frac{u_{\mathrm{pred}}u_{\mathrm{gt}} + v_{\mathrm{pred}}v_{\mathrm{gt}} + 1}{\sqrt{(u_{\mathrm{pred}}^2 + v_{\mathrm{pred}}^2 + 1)(u_{\mathrm{gt}}^2 + v_{\mathrm{gt}}^2 + 1)}}.$$

The additive constant term in both the numerator and denominator ensures numerical stability, particularly in zero-flow regions, and prevents division by zero.

3.2 Training Configuration

To evaluate the impact of regularization techniques on optical flow estimation, we optimize our functional (2) with $\lambda_1 = 0.2$, $\lambda_2 = 0.8$, and varying TV weights λ_{TV}. All experiments are conducted on a synthetic image and the Middlebury dataset using PyTorch [15], with deterministic settings to ensure reproducibility. Logging and visualization are automated for all runs.

In all configurations, we use the Adam optimizer [10] with a learning rate of 10^{-4}. Training is conducted for a fixed number of epochs, which serves as the stopping criterion. Each input consists of a pair of normalized grayscale frames, concatenated along the channel dimension. Since we use only one image pair—two frames I_1 and I_2—for training, the batch size is set to 1.

We employ the FDN with a fixed depth of $d = 4$. Although the architecture supports configurable depth, we do not perform hyperparameter tuning on this parameter; all models use the same setting to ensure comparability. This depth balances model capacity and computational efficiency.

The neural network selection is based on the training loss, and the best model checkpoint (i.e., with the lowest loss) is saved for each configuration. At each epoch, we compute the data and regularization terms, along with endpoint and angular error metrics: average endpoint error (AEE), standard deviation of endpoint error (SDEE), average angular error (AAE), and standard deviation of angular error (SDAE). These metrics are used to evaluate convergence behavior and to guide the final result visualization and comparison.

Infrastructure and Logging All experiments have been executed on NVIDIA GPUs using PyTorch with CUDA acceleration. The training framework supports comprehensive experiment tracking through structured logging. For each configuration, the system creates a dedicated directory containing JSON-formatted configuration files, full training logs, and performance summaries.

Loss curves and evaluation metrics (AEE, AAE, and their standard deviations) are recorded at every epoch. To maintain memory stability during long training runs, the system applies aggressive memory management strategies. After each epoch, the CUDA cache is explicitly cleared, and Python's garbage collector is invoked to prevent memory accumulation. This ensures robustness when running multiple configurations sequentially on large datasets.

3.3 Shepp-Logan Phantom with Synthetic Motion

To validate our method under idealized and interpretable conditions, we construct a controlled synthetic experiment based on the Shepp-Logan phantom [17]—a canonical analytic image commonly used in medical imaging and tomographic reconstruction due to its smooth grayscale transitions and well-defined elliptical structures. The phantom is set to 256 × 256 pixels and augmented with two distinct circular regions to simulate localized anatomical structures. The first circle, with moderate intensity (0.5), is positioned in the upper quadrant, while the second circle, with higher intensity (0.75), is placed in the lower quadrant. These intensity values create sufficient contrast against the phantom background while maintaining realistic tissue-like appearance.

Each circular region is assigned opposing vertical motion to simulate anatomical displacement patterns: the upper circle is translated upward by 3 pixels while the lower circle is shifted downward by the same amount. We warp the original image according to this synthetic flow field to generate the target frame. The obtained image frames and the respective ground-truth optical flow fields are depicted in Figure 1. All flow vectors are normalized by the maximum displacement magnitude to ensure optimal color saturation, allowing regions with maximum motion to appear as fully saturated colors. This synthetic setup provides an idealized benchmark for evaluating unsupervised optical flow models, combining well-defined anatomical structure with ground-truth motion fields.

,

We train the model for a fixed amount of 10,000 epochs, which allows the model to achieve highly accurate reconstruction of the imposed deformation, using two separate runs with two different TV weights, $\lambda_{\mathrm{TV}} = 0$ and $\lambda_{\mathrm{TV}} = 10^{-5}$. Figure 1 illustrates the predicted flow for both choices of λ_{TV} and demonstrates that the learned network captures the opposite vertical displacement of the two circles and shows increased smoothness with regularization.

In Figure 2, the training dynamics for $\lambda_{\mathrm{TV}} = 10^{-5}$ are summarized. Specifically, we observe that the loss curve (left) shows a rapid decrease within the first 1,000 epochs and continues to decline more gradually thereafter, indicating successful reconstruction of the synthetic motion patterns. AEE (middle) demonstrates an accurate magnitude estimation of flow vectors, while AAE (right) shows precise directional learning with consistent improvement throughout training. These metrics confirm that the network successfully learns both accurate magnitude and directional representations of the opposing circular motions. Note that the final best loss is 1.23×10^{-7}, indicating that the constant brightness assumption (1) is closely satisfied by the estimated optical flow.

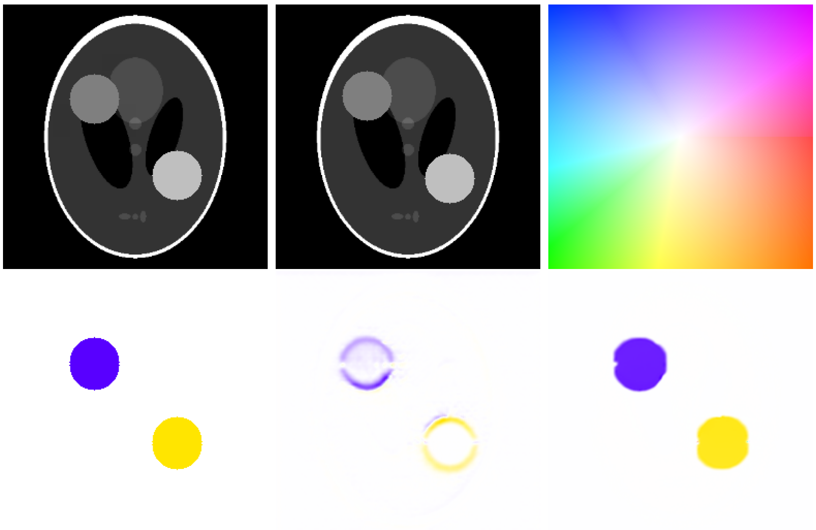


Figure 1: Synthetic Shepp-Logan phantom experiment. Top row (left to right): original phantom, synthetic frame 1 with embedded circles, warped frame 2, and color wheel. Bottom row: ground truth flow, predicted flows for $\lambda_{\mathrm{TV}} = 0$ and 10^{-5}, respectively, all trained for 10,000 epochs.

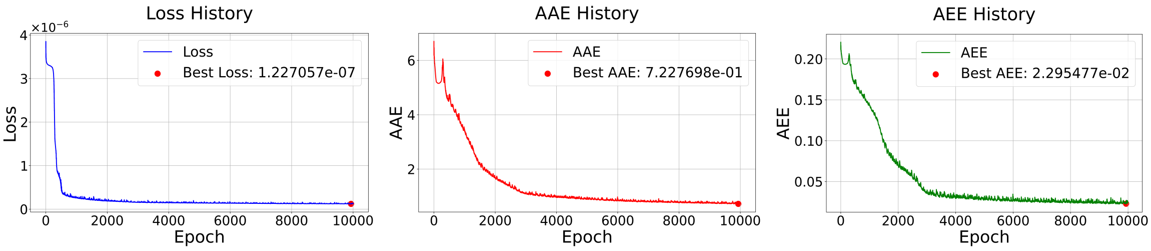


Figure 2: Training curves for the Shepp-Logan phantom experiment using total variation regularization with $\lambda_{\mathrm{TV}} = 10^{-5}$. The model is trained for 10,000 epochs. Left: Loss history showing stable convergence with a final best loss of 1.23×10^{-7}. Middle: AEE curve indicating accurate magnitude estimation of flow vectors, with a best AEE of 2.30×10^{-2} and SDEE of 1.88×10^{-1}. Right: AAE decreasing to a final value of 7.23×10^{-1}, with SDAE of 5.87.

Although the stopping criterion is a fixed number of epochs, the downward trends in all error curves of Figure 2 suggest that additional training could yield further improvements. The use of total variation regularization contributes to the smoothness and sharpness of the estimated flow field, particularly in preserving boundaries of the moving structures. Together, these results validate the model's ability to resolve localized motion in an interpretable, noise-free environment.

,

3.4 Results on Middlebury Benchmark

We evaluate FractalPINN-Flow on the Middlebury optical flow benchmark [2] to assess both quantitative accuracy and visual quality across a range of TV weights $\lambda_{\mathrm{TV}} \in \{0, 10^{-3}, 10^{-2}, 10^{-1}\}$. All models are trained for 20,000 epochs with fixed fractal depth $d = 4$. Table 1 reports the best training loss, epoch of best performance, and evaluation metrics—AEE, SDEE, AAE, and SDAE—for each benchmark scene. Across most benchmarks, intermediate regularization values ($\lambda_{\mathrm{TV}} = 10^{-2}$ or 10^{-3}) tend to yield the lowest AEE and AAE, striking a favorable balance between motion detail preservation and flow smoothness. For instance, in the *Dimetrodon* and *RubberWhale* scenes, $\lambda_{\mathrm{TV}} = 10^{-2}$ achieves the lowest AEE (0.33 and 0.17, respectively). Similarly, *Venus* shows optimal performance at $\lambda_{\mathrm{TV}} = 10^{-2}$ (AEE = 0.31, AAE = 0.08), while *Hydrangea* reaches its best results at the same setting (AEE = 0.43, AAE = 0.12). However, exceptions are present: in the *Grove3* scene, the best AEE (1.16) and AAE (0.17) are obtained at $\lambda_{\mathrm{TV}} = 10^{-1}$, outperforming lower regularization levels. In contrast, strongly regularized configurations degrade performance in noisy or high-disparity settings such as *Urban2*, where $\lambda_{\mathrm{TV}} = 10^{-1}$ yields an AEE of 7.64 versus 2.61 at $\lambda_{\mathrm{TV}} = 10^{-2}$. These results emphasize the critical role of appropriately tuned regularization in guiding unsupervised optical flow, particularly in complex scenes with variable texture and motion.

Benchmark	λ_{TV}	Best Loss	Best Epoch	AEE	SDEE	AAE	SDAE
Dimetrodon	0	0.000243	19867	0.77	0.63	0.31	0.31
	10^{-1}	0.001957	19625	0.42	0.46	0.16	0.23
	10^{-2}	0.001187	19928	0.33	0.47	0.13	0.23
	10^{-3}	0.000726	19794	0.41	0.47	0.17	0.24
Grove2	0	0.000822	19648	0.52	0.55	0.12	0.16
	10^{-1}	0.006699	19284	1.18	1.54	0.52	0.73
	10^{-2}	0.002833	19972	0.2	0.42	0.05	0.12
	10^{-3}	0.001276	19610	0.34	0.43	0.08	0.12
Grove3	0	0.001730	19779	1.94	2.31	0.45	0.58
	10^{-1}	0.011981	19866	1.16	1.67	0.17	0.31
	10^{-2}	0.004666	19292	1.18	1.87	0.2	0.33
	10^{-3}	0.002215	19914	1.34	1.97	0.26	0.39
Hydrangea	0	0.000414	19749	0.74	1.25	0.18	0.34
	10^{-1}	0.005052	19867	0.48	1.04	0.13	0.32
	10^{-2}	0.001893	19964	0.43	1.12	0.12	0.32
	10^{-3}	0.000810	19773	0.56	1.21	0.15	0.33
RubberWhale	0	0.000190	19822	0.43	0.53	0.22	0.27
	10^{-1}	0.002276	19969	0.34	0.6	0.19	0.35
	10^{-2}	0.001068	19914	0.17	0.38	0.1	0.24
	10^{-3}	0.000488	19722	0.25	0.42	0.13	0.24
Urban2	0	0.001047	19639	3.5	5.64	0.28	0.42
	10^{-1}	0.010732	19757	7.64	7.77	0.75	0.43
	10^{-2}	0.003046	19877	2.61	5.06	0.11	0.23
	10^{-3}	0.001351	19579	2.74	5.26	0.14	0.29
Urban3	0	0.000675	19779	3.4	4.47	0.38	0.7
	10^{-1}	0.007248	19541	4.68	4.28	0.47	0.72
	10^{-2}	0.003074	19637	3.26	3.95	0.33	0.63
	10^{-3}	0.001127	19331	2.6	4.17	0.3	0.7
Venus	0	0.000583	19828	0.73	0.89	0.17	0.33
	10^{-1}	0.009652	19799	1.7	1.96	0.62	0.7
	10^{-2}	0.001779	19646	0.31	0.66	0.08	0.29
	10^{-3}	0.000997	19828	0.46	0.92	0.11	0.32

Table 1: Benchmark results for various λ_{TV} configurations, based on training for 20,000 epochs.

Figure 3 visualizes the predicted flow fields for each configuration, revealing the qualitative effects of λ_{TV} on spatial smoothness and edge preservation. High regularization improves visual coherence but

,

risks oversmoothing fine structures, while low or zero regularization retains motion discontinuities but introduces noise and instability. These findings demonstrate the capacity of our fractal-based model to generalize across a wide spectrum of motion patterns and visual complexities, while also emphasizing the practical importance of hyperparameter tuning in unsupervised flow models.

The results in Table 1 highlight the sensitivity of FractalPINN-Flow to the choice of total variation regularization weight λ_{TV}. Across most benchmarks, introducing moderate regularization ($\lambda_{\text{TV}} = 10^{-2}$) consistently yields the lowest AEE and AAE, suggesting that total variation plays a key role in suppressing noise while preserving meaningful motion boundaries.

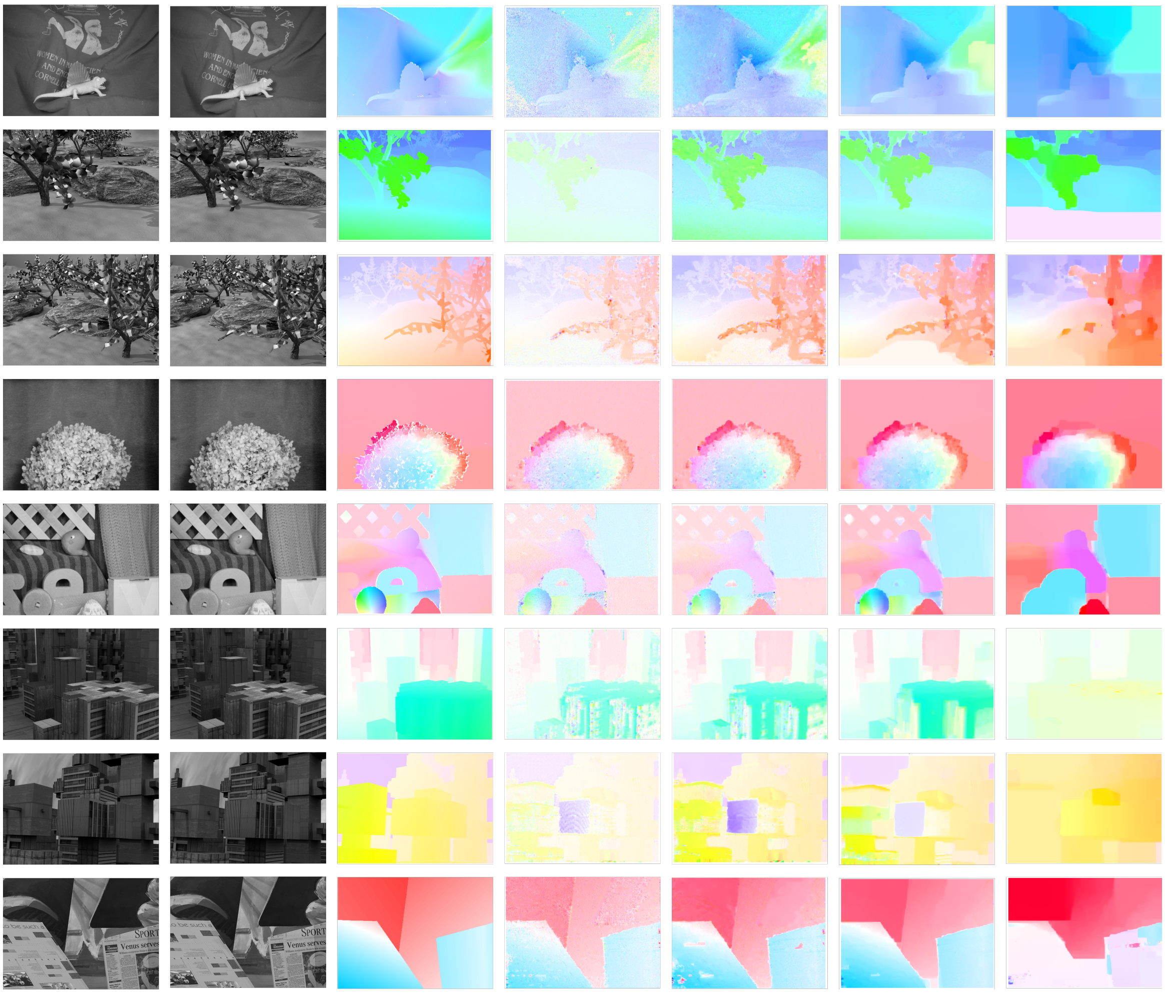


Figure 3: Middlebury Optical Flow Benchmark visualizations corresponding to the results in Table 1. Columns from left to right: I_1, I_2, ground truth optical flow, and predicted flows for $\lambda_{\text{TV}} = 0$, 10^{-3}, 10^{-2}, 10^{-1}. Predicted flow fields are taken from the best-loss epoch for each configuration. Benchmarks from top to bottom: *Dimetrodon*, *Grove2*, *Grove3*, *Hydrangea*, *RubberWhale*, *Urban2*, *Urban3*, *Venus*.

References

[1] Martin Alkämper, Stephan Hilb, and Andreas Langer. A primal-dual adaptive finite element method for total variation minimization. *arXiv preprint arXiv:2404.03125*, 2025.

[2] Simon Baker, Daniel Scharstein, J.P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

,

[3] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019.

[4] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015.

[5] Stephan Hilb, Andreas Langer, and Martin Alkämper. A primal-dual finite element method for scalar and vectorial total variation minimization. *Journal of Scientific Computing*, 96(1):24, 2023.

[6] Michael Hintermüller and Andreas Langer. Subspace correction methods for a class of nonsmooth and nonadditive convex variational problems with mixed L^1/L^2 data-fidelity in image processing. *SIAM Journal on Imaging Sciences*, 6(4):2134–2173, 2013.

[7] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial Intelligence*, 17(1-3):185–203, 1981.

[8] Thomas Jacumin and Andreas Langer. An adaptive finite difference method for total variation minimization. *Numerical Algorithms*, pages 1–36, 2025.

[9] Rico Jonschkowski, Austin Stone, Jonathan T. Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 557–572, 2020.

[10] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

[11] Andreas Langer. Automated parameter selection in the L^1-L^2-TV model for removing Gaussian plus impulse noise. *Inverse Problems*, 33(7):074002, 2017.

[12] Andreas Langer and Sara Behnamian. DeepTV: A neural network approach for total variation minimization. *arXiv preprint arXiv:2409.05569*, 2024.

[13] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. FractalNet: Ultra-deep neural networks without residuals. In *International Conference on Learning Representations (ICLR)*, 2016.

[14] Simon Meister, Junhwa Hur, and Stefan Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[15] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035, 2019.

[16] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[17] Larry A. Shepp and Benjamin F. Logan. The Fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

[18] Xiaoyu Shi, Zhaoyang Huang, Wenjie Bian, Daquan Li, Minghang Zhang, Ka Chun Cheung, Shijian Lu, Hongwei Qin, Jifeng Dai, and Hongsheng Li. VideoFlow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10381–10391, 2023.

,

[19] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. SMURF: Self-teaching multi-frame unsupervised RAFT with full-image warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2886–2895, 2021.

[20] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, 2018.

[21] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[22] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for real-time TV-L1 optical flow. In *Pattern Recognition (DAGM)*, pages 214–223. Springer, 2007.

,

2nd Sorbonne-Heidelberg Workshop on AI in Medicine

Transforming Normal ECG to Myocardial Infarction Ones using Invertible Conditional GANs

Sara Battiston[1], Roberto Sassi[1] and Massimo W. Rivolta[1]

***Abstract*— Recent advances in deep learning have enabled effective style transfer methods for biosignal synthesis, particularly for generating pathological variations of physiological signals. This work investigates the application of Invertible Conditional Generative Adversarial Networks (IcGANs) to modify 12-lead ECG heartbeats from normal sinus rhythms into patterns typical for myocardial infarction (inferior and antero-septal). In contrast to CycleGAN, which requires multiple models for each direction of transformation, IcGANs only require the training of a single conditional GAN along with an encoder, offering a more direct and lightweight framework. We trained both IcGAN and CycleGAN models using heartbeats from the PTB-XL dataset. The quality of the generated ECGs was assessed using both qualitative and quantitative techniques, including visual inspection, GAN-train and GAN-test scores, and comparisons of ST-segment amplitudes. Our results indicate that IcGAN can realistically and meaningfully alter ECG signals to exhibit myocardial infarction traits while retaining their core physiological structure. Comparisons showed IcGAN to be more efficient and effective than CycleGAN under similar architectural conditions. This approach shows promise for generating rare pathological cases, adapting models across domains, and supporting generalization in clinical applications for more personalized diagnostics.**

I. Introduction

Style transfer via a deep learning (DL) approach has gained traction in biomedical signal processing, particularly for synthesizing physiological signals when paired data is unavailable. In fact, in typical supervised setups, DL-based transformations need corresponding input-output pairs, which is rarely obtainable in clinical contexts where recording signals before and after acute conditions like myocardial infarction (MI) is a rare occasion. A similar challenge exists in computer vision, where unpaired image-to-image translation has been addressed by CycleGANs [1], which learn bidirectional mappings between domains through adversarial training and without requiring paired samples. In earlier work [2], we applied CycleGANs to convert normal ECGs into those exhibiting infarction-related features with respectable success. However, CycleGANs demand the training of four DL models per transformation, making them resource-intensive. To address this, we exploit an alternative adversarial framework called Invertible Conditional GANs (IcGANs), proposed by Perarnau et al. [3], to tackle the unpaired style transfer task. IcGAN combines a conditional GAN (cGAN) [4] with an encoder that maps input data back to the latent space. This pairing allows targeted alterations of the input signals while maintaining their key characteristics, which is especially beneficial for generating synthetic clinical data.

In this study, we evaluate the use of IcGANs for ECG style transfer, specifically transforming heartbeats from normal sinus rhythm into rhythms associated with myocardial infarction (inferior and antero-septal MI). We assess the quality of the generated signals using visual and quantitative measures. Lastly, we benchmark the performance of IcGAN generation against CycleGANs, while keeping architectural choices constant (same generator architecture), in order to isolate the effect of the training strategy and address the common practice of repurposing existing architectures.

II. Methods

A. ECG Data

The dataset employed in this study is a subset of the PTB-XL ECG dataset [5], [6], comprising standard 10 s 12-lead ECGs from a cohort of 18,885 patients. For these data, we select three ECG subpopulations according to their diagnosis: healthy patients (NORM, 80%), antero-septal myocardial infarction (ASMI, 12%), and inferior myocardial infarction (IMI, 8%). Then, we apply to these ECGs classes a series of preprocessing steps, including the filtering by means of a zero-phase Butterworth pass-band filter (order 3, cutoff freq. of 0.67 and 15 Hz) and the segmentation into windows of 0.40 s containing a single heartbeat, meaning a QRS complex and T-wave (R- preaks identified through the WFDB library [7], [6]). The final dataset consists of 35,353 12-lead 0.40 s ECG signals distributed as: 12,000 NORM (34%), 12,650 ASMI (36%), and 10,703 IMI (30%). Finally, this dataset is split into 70% training and 30% test sets, ensuring that no patient's heartbeat appears in both sets. The models are trained on the training set, and their performance is assessed using the test set.

B. Style Transfer through IcGAN

To achieve ECG class transformation and thus perform style transfer, we utilize IcGANs [3], which extend standard conditional GANs (cGANs) [4] by adding an encoder that approximates the inverse mapping from data to latent space. While a typical cGAN learns to generate a sample x' from a noise vector $z \sim \mathcal{N}(0, I)$ and a class label y, i.e., $x' = G(z, y)$, IcGANs introduce an encoder E that estimates the original latent input from a generated signal: $E(x') = (\hat{z}, \hat{y})$. This enables editing real signals by first encoding a given ECG x to its latent form, $x \rightarrow E(x) = (z, y)$ and then generating a new version x' conditioned on a different label y': $x' = G(z, y')$. An idea of the pipeline is reported in fig. 1

[1] Authors are all affiliated with the Department of Computer Science, Università degli Studi di Milano, Via Celoria 18, 20133, Milan, Italy. Corresponding author: sara.battiston@unimi.it

,

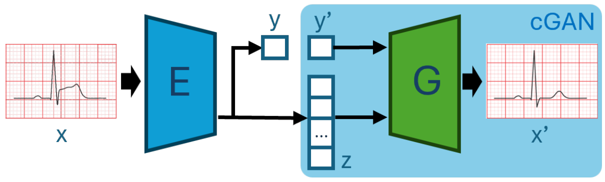


Fig. 1: IcGAN scheme.

TABLE I: Architecture of the IcGAN generator. The first and second Conv2dT layer has output_padding set to 1, while for the second Conv2dT the dilation is set to $(1, 2)$.

Layer	Filters	Pad	Stride	Kernel	IN	Activ.
Fully connected: in_shape 100, out_shape 1728						
Reshape in $[64, 9, 3]$						
Conv2dT	32	1	2	5x3	Yes	ReLU
Conv2dT	16	2	2	5x3	Yes	ReLU
Conv2d	1	same	1	5x3	No	No

TABLE II: Architecture of IcGAN Discriminator.

Layer	Filters	Pad	Stride	Kernel	IN	Activ.
Embedding Layer for labels: in_shape 1, out_shape 480						
Reshape labels in $[1, 40, 12]$						
Concatenate with signal, out_shape $[2, 40, 12]$						
Conv2d	16	1	2	3x3	No	ReLU
Conv2d	32	1	2	3x3	Yes	ReLU
Conv2d	64	1	1	3x3	Yes	ReLU
Conv2d	1	same	1	3x3	No	ReLU
Flatten						
Fully connected: in_shape 30, out_shape 1						
Sigmoid Activation						

TABLE III: Architecture of IcGAN Encoder.

Layer	Filters	Pad	Stride	Kernel	IN	Activ.
Conv2d	16	same	1	5x3	Yes	ReLU
Conv2d	32	1	2	5x3	Yes	ReLU
Conv2d	64	1	2	5x3	Yes	ReLU
Flatten						
Fully connected: in_shape 1728, out_shape 100						

The IcGAN training procedure is the same as specified in [3]: first the cGAN is trained; then with the paired random values and labels (z, y) and their respective cGAN generated signals x', we train the Encoder E. The goal for E is to minimize the reconstruction loss:

$$\mathcal{L}_E = \mathbb{E}_{z \sim p_z, y \sim p_y} \left[\|(z, y) - E(G(z, y))\|_2^2 \right] \tag{1}$$

C. Network Design, Training and Synthetic Data Creation

The IcGAN network is composed by a cGAN with generator G and Discriminator D, and an Encoder E. The networks designed are summarized in tables I, II and III, respectively.

In particular, G receives a noise vector $z \in \mathbb{R}^{99}$ concatenated with an integer class label y, resulting in a 100-dimensional input. This is passed through a fully connected layer and reshaped to a tensor of shape $[64, 9, 3]$. It then flows through 2 2D-convolutional blocks and then is passed to a final 2D convolution. The output is a tensor of shape $[1, 40, 12]$, representing a single heartbeat segment with 40 time steps and 12 ECG leads. The architecture of D takes as input an ECG segment of shape $[1, 40, 12]$ and stacks it with an embedded version of the class label (reshaped to the same dimensions), producing a combined tensor of shape $[2, 40, 12]$. This is processed through a series of Conv2d layers, ReLU activations and instance normalization. The output is flattened and passed to a fully connected layer with a sigmoid function for binary classification. The encoder E, is trained from synthetic samples sending ECGs back to the latent space. It accepts inputs of shape $[1, 40, 12]$ and outputs a vector in $\mathbb{R}^{100}$, where the first 99 components estimate the original noise z, and the last element predicts the label y. The architecture includes three Conv2d blocks with instance normalization and ReLU, followed by a fully connected layer. All three networks (G,D and E) are optimized using Adam with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate is fixed at 10^{-3}, and training is performed using mini-batches of size 128 for 50 epochs. After training the cGAN, we generate 1,000 synthetic ECGs for each of the three classes (NORM, ASMI, IMI), resulting in 3,000 (z, y) and $x' = G(z, y)$ pairs for training the encoder. Once trained, the encoder and generator are combined to build the full IcGAN. To simulate class transitions, we feed real ECGs from the test set through the encoder to obtain (z, y) and used the generator to create modified signals with alternate labels $y' \neq y$. For each test ECG, two synthetic versions are generated corresponding to the other two classes. This results in a new dataset of 21,212 generated ECGs: 7,700 ASMI, 6,384 IMI, and 7,128 NORM samples.

D. Synthetic Data Quality Evaluation

To assess the realism and clinical plausibility of the generated ECGs, we employed three evaluation strategies: visual analysis and classification-based scores.

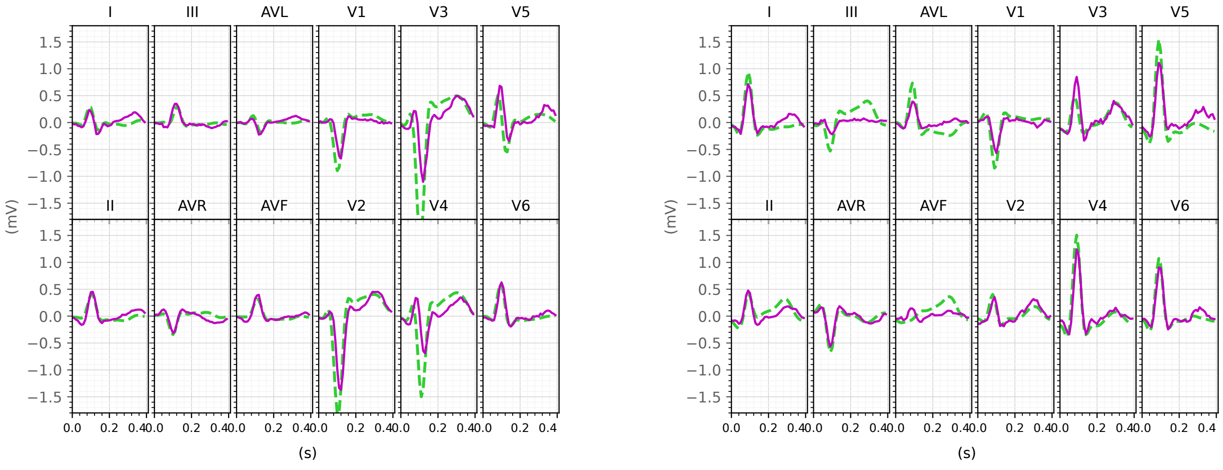
a) Visual Comparison: We plot 90% confidence intervals across all leads using 2,500 randomly chosen real and synthetic ECGs. Additionally, we use the UMAP [8] algorithm (with 15 neighbors and Euclidean distance) to project high-dimensional ECGs into a 2D space. The real signals from the test set are used to fit the UMAP model, and the corresponding synthetic signals are overlaid on the resulting density map for visual comparison.

b) GAN-related evaluation Scores: These metrics from [9] evaluate how well the synthetic data align with real data distributions. GAN-train is the accuracy of a supplementary classifier trained on synthetic data and tested on real data, while GAN-test is the reverse. We use a CNN adapted from a prior work on myocardial infarction classification [10], modified by adjusting the input layer and reducing the ResNet blocks. The classifier is trained to distinguish between the three classes (NORM, ASMI, IMI) using 85% of each dataset and validated on the remaining 15%. Each model is trained for 4 epochs.

E. Benchmarking against CycleGAN

To benchmark the performance of IcGAN, we implement a CycleGAN using nearly identical architectural components. The CycleGAN generator is constructed by linking the

,



(a)

(b)

Fig. 2: Examples of original ECGs (green, dashed) and their IcGAN-generated counterparts (magenta): (a) Real NORM and ASMI generated from NORM, (b) Real IMI and NORM generated from IMI.

IcGAN encoder and generator, with minor changes: a ReLU activation is added after the encoder, and a dropout layer ($p = 0.1$) is inserted before the final activation in the generator. A table summarizing the cycleGAN generator is shown in IV. The discriminators are adapted from the IcGAN version by removing the label embedding, since CycleGAN does not use conditional inputs. Despite these changes, the total number of parameters in the CycleGAN generator matches that of the IcGAN (424,869), while the discriminator has slightly fewer parameters (23,840 vs 25,488). We train the CycleGAN for 50 epochs using the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), a learning rate of 10^{-3}, and a batch size of 128, identical to the IcGAN training setup.

TABLE IV: Architecture of cycleGAN generator. The first and second Conv2dT layer have output_padding set to 1, while for the second Conv2dT the dilation is set to $(1, 2)$.

Layer	**Filters**	**Pad**	**Stride**	**Kernel**	**IN**	**Activ.**
Conv2d	16	same	1	5x3	Yes	ReLU
Conv2d	32	1	2	5x3	Yes	ReLU
Conv2d	64	1	2	5x3	Yes	ReLU
Flatten						
Fully connected: in_shape 1728, out_shape 100						ReLU
Fully connected: in_shape 100, out_shape 1728						
Dropout:$p = 0.1$						ReLU
Reshape in $[64, 9, 3]$						
Conv2dT	32	1	2	5x3	Yes	ReLU
Conv2dT	16	2	2	5x3	Yes	ReLU
Conv2d	1	same	1	5x3	No	No

III. Results and Discussions

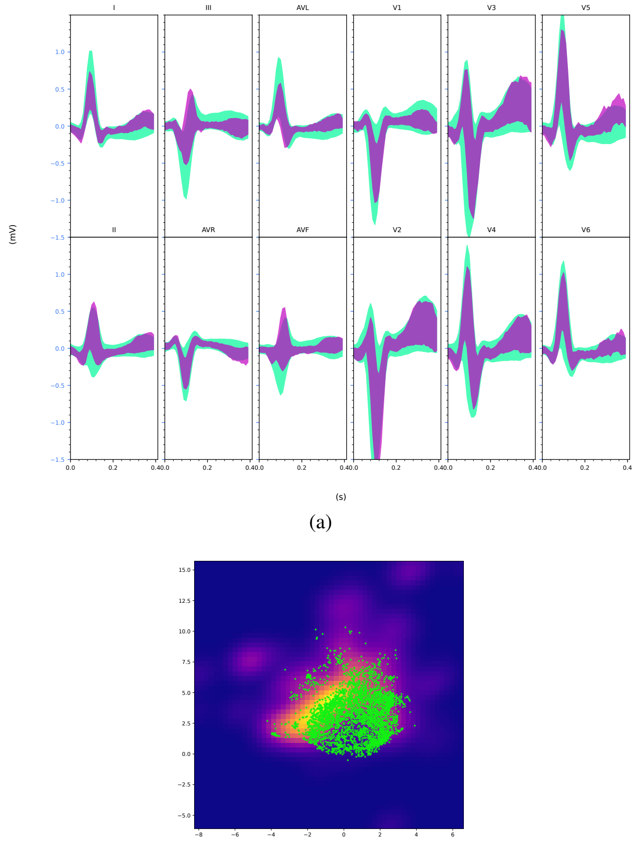
We evaluate the IcGAN-generated ECGs using both qualitative and quantitative methods, comparing them to real signals and those synthesized by CycleGAN. Only two sample comparison between cases are reported here, as the others show similar results.

a) Visual Inspection: Figure 2 illustrates two transformation examples: a synthetic NORM ECG generated from an ASMI sample (fig. 2a) and a NORM ECG synthesized from an IMI sample, (fig. 2b). It can be seen that the IcGAN successfully altered the leads most relevant to the target condition. In fact, in the examples reported, the amplitude of the ST-segment of both the original real ASMI and IMI signals (green, dashed) is flattened in their NORM synthetic counterparts (magenta), precisely in leads V_1, V_2, V_3 for the ASMI transformation and in II, III, aVF for the IMI transformation, all while preserving the other unrelated leads.

b) Confidence Bands and UMAP Analyses: Figures 3 and 4 report the Confidence Bands (fig. 3a, 4a) and UMAP plots (fig. 3b, 4b) relative to the comparison between synthetic IMI ECG generated from a NORM signal and synthetic NORM ECG generated from an ASMI signal respectively. The 90% confidence intervals for IcGAN-generated signals largely overlap with those of real ECGs, especially in the NORM class, fig. 4a, while for the IMI case, fig. 3a, the bands likely suggest that the IcGAN generated a sub-distribution of the real IMI ECG cohort, being the magenta bands fully contained into the green ones. UMAP projections confirm these findings by showing substantial overlap between real data (heatmap) and synthetic data (scattered green points) in the NORM case (fig. 4b), indicating that the generated signals reside near the real data manifold. In contrast, in the IMI case (fig. 3b), the overlap is only partial and limited to the densest region of the real IMI distribution.

c) GAN-based Scores: The IcGAN achieves a GAN-test accuracy of 84% (baseline: 90%) and a GAN-train score of 73% (baseline: 100%). These results suggest high fidelity of the generated data to the real distribution but slightly reduced variability in the synthetic signals.

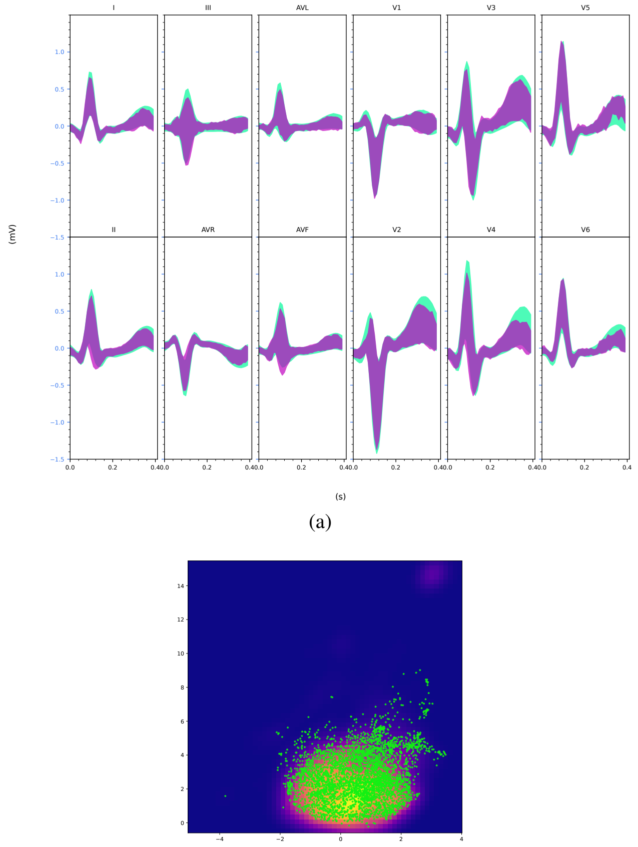
,



(a)

(b)

Fig. 3: IcGAN vs. real ECG evaluation. (a) 90% confidence bands: IMI from NORM (magenta) vs. real IMI (green). (b) UMAP heatmap of real IMI ECG overlaid with synthetic scatterplot of IMI generated from NORM.



(a)

(b)

Fig. 4: IcGAN vs. real ECG evaluation. (a) 90% confidence bands: NORM from ASMI (magenta) vs. real NORM (green). (b) UMAP heatmap of real NORM ECG overlaid with synthetic scatterplot of NORM generated from ASMI.

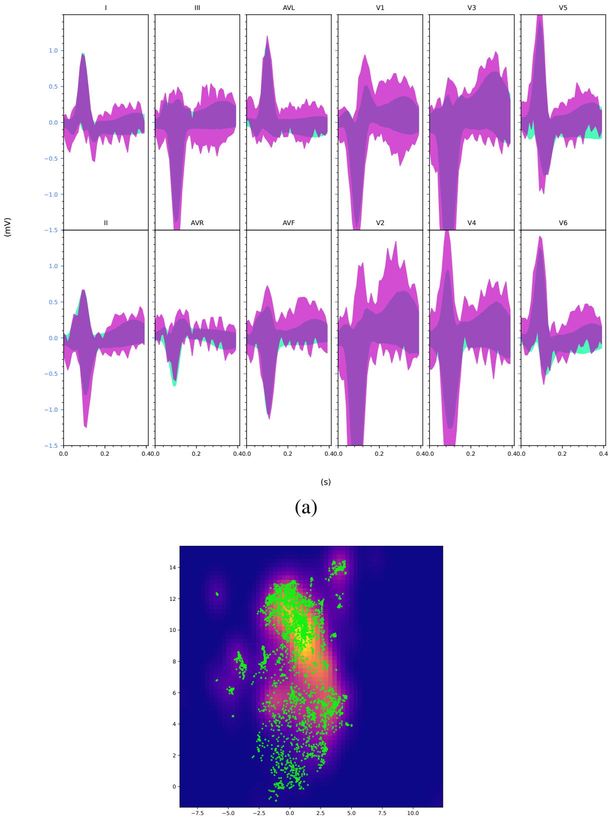
d) CycleGAN Baseline: The CycleGAN model, trained with a similar architecture and number of epochs, performs worse across all metrics. Its GAN-test score drops to 74%, and GAN-train drops to 60%, indicating poorer generalization and less alignment with the real ECG distribution. The sample visualizations in fig. 5 show very noisy confidence bands (fig. 5a) for the synthetic ASMI signals generated from IMI (magenta) compared to the real bounds for the ASMI ECGs. The UMAP plot (fig. 5b) shows that the synthetic ASMI data are mostly scattered outside the densest real ASMI signal regions, thus indicating a poor distribution fit.

e) Overall Findings: The findings demonstrate that the IcGAN model provides reliable and class-consistent ECG transformations. It mainly modifies diagnostic features while preserving overall ECG morphology. The synthetic data closely resemble real samples, especially in the NORM class. Although the generated IMI signals represent a sub-distribution of the original population, their quality remains acceptable. Visual and quantitative evaluations confirm strong alignment between synthetic and real data. This is reflected in the performance metrics: the IcGAN achieves a GAN-test score of 84% and a GAN-train score of 73%, outperforming the CycleGAN baseline at 74% and 60%, respectively. These results suggest that the IcGAN-generated data form a faithful subspace of the real ECG distribution and indicate that reusing architectures like CycleGANs without adaptation may not yield optimal results. During training, we notice that the IcGAN Encoder is acting like an effective output regularizer, though the use of an excessive amount of training data introduced noise. Limiting the encoder training to 50 epochs proves beneficial. Future work could explore early stopping or adaptive sample sizes to reduce overfitting. A further advantage of IcGANs is efficiency: unlike CycleGANs, which require separate models per direction, IcGANs use a single encoder-generator pair, simplifying training and deployment across multiple classes.

IV. Conclusion

In conclusion, we showed that IcGAN offers an efficient solution for unpaired ECG translation without requiring a separate model for each class shift. It reliably alters key ECG features while preserving the input's identity. The generated signals reflect clinically relevant patterns across ASMI, IMI, and NORM, as shown by confidence intervals, and UMAP

,



(a)

(b)

Fig. 5: (a) 90% confidence intervals: real ASMI (green) vs. CycleGAN-generated ASMI from IMI (magenta). (b) UMAP visualization: real ASMI density (heatmap) and corresponding synthetic ASMI points from IMI (scatter).

projections in our examples. Its use of label encoding enables precise control over output features, making it a flexible tool for synthesizing targeted abnormalities. This approach holds promise for tasks like device or population adaptation, rare data generation, and improving model robustness in clinical settings.

Acknowledgements

Part of the research was supported by the MUSA – Multilayered Urban Sustainability Action – project, funded by the European Union – NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strengthening of research structures and creation of R&D “innovation ecosystems”, set up of “territorial leaders in R&D”. The work was also supported by the project FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[2] S. Battiston, R. Sassi, and M. W. Rivolta, “Evaluating the quality of cyclegan generated ECG data for myocardial infarction classification,” in *2024 Computing in Cardiology Conference*, vol. 51, 2024.

[3] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, “Invertible conditional GANs for image editing,” *ArXiv preprint arXiv:1611.06355*, 2016.

[4] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *ArXiv preprint arXiv:1411.1784*, 2014.

[5] P. Wagner, N. Strodthoff, R. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter, “PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). *PhysioNet*,” 2022, https://doi.org/10.13026/kfzx-aw45.

[6] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals,” *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[7] C. Xie, L. McCullum, A. Johnson, T. Pollard, B. Gow, and B. Moody, “Waveform Database Software Package (WFDB) for Python (version 4.1.0),” 2023, PhysioNet. DOI: 10.13026/9njx-6322.

[8] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” *ArXiv preprint arXiv:1802.03426*, 2018.

[9] K. Shmelkov, C. Schmid, and K. Alahari, “How good is my GAN?” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 213–229.

[10] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr, and Others, “Automatic diagnosis of the 12-lead ECG using a deep neural network,” *Nature Communications*, vol. 11, no. 1, pp. 1–9, 2020.