Detailed report of the PhD thesis titled "Sentiment Analysis for Pain Detection Using Multimodal Data" by Anay Ghosh submitted to the Department of Information Technology, Faculty Council of Engineering & Technology, Jadavpur University, Kolkata, India

Reviewer -1

Comment-1. Page 47: Change 'Unnecessary punctuation' to 'punctuations'.

Response-1: We are thankful to the reviewer for the observation. The suggested update has been incorporated on page 47 under the **'Text Cleaning'** section of **Chapter-2.**

Comment-2. Page 47: What do you mean by 'repeated letters'? Such as letter 'r' in correction'? **Elaborate with example. Why is it necessary?**

Response-2: We sincerely thank the reviewer for the insightful observation. Repeated letters refer to cases where a letter appears more times than necessary, usually because of informal typing, emphasis, or errors. For example, people may write:

- "soooo good" instead of "so good".
- "cooool" instead of "cool".
- "happppy" instead of "happy".

Cleaning repeated letters is necessary because NLP models treat each spelling variation as a different word, so leaving exaggerated forms like "soooo" or "cooool" increases noise, makes the vocabulary unnecessarily large, and prevents the model from recognizing that these words actually mean the same thing; by reducing such repetitions to their standard form, the text becomes more consistent, easier to process, and helps the model learn clearer and more meaningful patterns.

(The answer has been incorporated on page 48 under Section 'Text Cleaning' of Chapter-2)

14.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

13.01.2026
Dr. Saiyed Umer
Assistant Professor
of CSE, Aliah Unive
Ukata-700156

Comment-3. Page 47: If text cleaning is already successfully complete, then how is tokenization using hyphen, apostrophes or capital letter done?

Response-3: We acknowledge the reviewer for bringing this important point to our attention. It has been observed that text preprocessing is not fully robust even after text cleaning is complete, tokenization still uses elements like hyphens, apostrophes, and capital letters because cleaning only removes unnecessary noise; it does not remove meaningful structures inside valid words or sentences. Tokenizers rely on these features to correctly split the text. For example, a hyphen in "well-being" or an apostrophe in "can't" is preserved during cleaning because they change the meaning of the word, so the tokenizer uses them to decide whether the word should stay whole or be split into smaller parts. Similarly, capital letters are kept in sentence boundaries, helping the tokenizer recognize where one sentence ends and another begins. In this way, cleaning prepares the text by removing irrelevant symbols, while tokenization uses the remaining meaningful punctuation and structure to break the text into accurate, analyzable units.

(The answer has been incorporated on page 48 under the 'Tokenization' section of Chapter-2)

Comment-4. Page 50, in the example: how does 'the' 'on' etc. be present after stop word removal as suggested text preprocessing? OR stop word removal is not applied to form the corpus (or BoW) and also the vector?

Response-4: We appreciate the observation of the reviewer. It is a typographical mistake on our end. The raised issue has been resolved by removing stopwords from the corpus of unique words, and the vector representations have been updated accordingly.

(The updation has been incorporated on pages 51 and 52 under the section 'Feature Extraction' of Chapter-2)

Comment-5. Page 51, equn. 2.4: Write c' in place of d to maintain conformity with equn. (2.1).

14.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot 8, Sector 3
Salt Lake, Kolkata-700106, India

14.01.2026
Dr. Sajjad Umer
Assistant Professor
Dept. of CSE, Alah University
Kolkata-700156

Response-5: We are grateful to the reviewer for the constructive feedback. The suggested changes are incorporated into equation 2.4 on page 52 of **Chapter-2**.

Comment-6. Page 62 & 63: In figures 2.10 and 2.11, since all the classes do not have same number of samples, percentage (wherever possible) would represent the performance better than the absolute values. Same is true for other confusion matrices.

Response-6: We are thankful to the reviewer for the thoughtful observation. The suggested changes are incorporated under the section **'Experiments and Results'** of each contributing chapter.

Comment-7. Page 128, Fig. 5.2: based on what criterion are the end-to-end model and the deep learning models separated? Is not a pre-trained model also a deep learning model?

Response-7: We acknowledge the reviewer for highlighting this issue. We agree that a pre-trained model is indeed a deep learning model. In this work, the separation between end-to-end models and deep learning models is based on their training strategies and usage, rather than on their underlying learning paradigms. The proposed end-to-end model and ensemble models are trained from scratch using our dataset, with feature extraction and classification jointly optimized in a single learning process. In contrast, pre-trained deep learning models rely on weights trained on large external datasets and are subsequently fine-tuned or used as feature extractors within our framework. Therefore, the distinction is made to emphasize the difference between end-to-end training from scratch and transfer learning–based approaches, not to imply that pre-trained models are outside the scope of deep learning.

(The answer has been incorporated on page 131 under the 'Proposed PSA_{video} Systems' section of Chapter-5)

Comment-8. Page 129, Figs. 5.3 & 5.4: How is change between frames calculated? Elaborate.

Response-8: We deeply appreciate the observation of the reviewer. The change between frames is calculated with the help of the Sum of Absolute Difference (SAD) by comparing each frame with the next

14-01-26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

Dr. Saiyed Umer
Assistant Professor
Dept. of CSE, Alipur University
Kolkata-700156

one to measure how much the frames are changing over time. First, both frames are usually converted to grayscale and optionally resized (e.g., to 100×100 pixels) to simplify computation. For each pair of consecutive frames (F_i) and (F_{i+1}), the pixel-wise difference is computed using the formula $D(x, y) = |F_i(x, y) - F_{i+1}(x, y)|$, where (x, y) represents each pixel location. All these absolute differences are then summed to produce a single change score for that transition, given by $ChangeScore_i = \sum_{x,y} D(x, y)$. A low score means the two frames are similar and contain little movement, while a high score indicates significant motion or abrupt change. Plotting these scores over time provides a clear view of where and how strongly the frame variations occur within a video.

(The answer has been incorporated on page 134 under the 'Video Preprocessing' section of Chapter-5.)

Comment-9. Page 159, Table 6.2: Why not combination of Text and Video modality included?

Response-9: We are grateful to the reviewer for the observation. In this thesis, three multimodal PSA systems have been proposed, such as $MPSA_{TA}$System (Multimodal system combination of Text and Audio), $MPSA_{AV}$System (Multimodal system combination of Audio and Video) and $MPSA_{TAV}$System (Multimodal system combination of Text, Audio and Video) (refer to Table 6.2 of the revised thesis). From these multimodal systems, it has been observed that the performance of $MPSA_{AV}$System is better than MPSATA System, and the performance of $MPSA_{TAV}$ System is much better than $MPSA_{AV}$System and $MPSA_{TA}$ System. There is cumulative performance improvement that has been seen from $MPSA_{TA}$ System to $MPSA_{AV}$System to $MPSA_{TAV}$ System. While the use of Multimodal system combination of Text and Video has a drop-off in the cumulative performance growth. At the same time, the possibility of getting data from the same patient is more challenging than the other three multimodal systems considered. In this work, PSA_{audio}Systems has been used as the bridge between PSA_{text}Systems and PSA_{video}Systems. It has also been observed that the fusion combination with only Text and Video reduces the performance; thus,

19.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

13.01.26

Assistant Professor
Dept. of CSE, Aliah University
Kolkata-700156

this combination has not been included.

(The answer has been incorporated on page 165 under the Section 'Modalities Fusion' in Chapter-6)

Comment-10. Since multi-modal inputs are not coming from the same source, how is this post-classification fusion meaningful for real problem? Elaborate.

Response-10: We appreciate the reviewer for the valuable comment. We considered data from heterogeneous sources. When the data comes from a single source, there exists correlation within different modalities of data. In the case of a real problem, when the data comes from a single subject, the post-classification fusion ensure more accurate evaluation.

Although the multi-modal inputs originate from different sources, the post-classification fusion remains meaningful because the modalities are complementary and temporally aligned representations of the same underlying event or subject state. In real-world scenarios, heterogeneous sensors (e.g., text, audio, and visual-based sources) are commonly deployed in parallel, and each modality captures distinct aspects of the same phenomenon. Post-classification fusion allows each modality-specific model to independently learn discriminative patterns within its own feature space, while the fusion stage aggregates their class-level confidence scores to produce a more robust and reliable final decision. This strategy is particularly suitable when the modalities have different sampling rates, noise characteristics, or missing data, as it avoids early-stage feature incompatibility and reduces error propagation. Therefore, even though the inputs are acquired from different sources, their fusion at the decision level reflects realistic multi-sensor deployments and improves generalization in practical applications.

(The answer has been incorporated on page 153 under the introductory segment of Chapter-6.)

19.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

13.01.2026

Dr. Saiyed Umer
Assistant Professor
Dept. of CSE, Aliah University
Kolkata-700156

Reviewer -2

Comment-1: Literature search and understanding of prior art :

The review effectively situates the study within current research trends and identifies critical gaps such as data imbalance, multimodal synchronization, and subjectivity in pain perception that justify the research problem. Relevant and recent sources (up to [89]) are cited appropriately, though a few citations could have been replaced with more recent works (2023–2024) to reflect the latest developments in multimodal fusion and transformer

based architectures.

Response-1: We are grateful to the reviewer for the positive evaluation of the literature review. The suggestion of the reviewer regarding the inclusion of more recent works (2023–2024) is well noted and listed below, and relevant updates are incorporated to further strengthen the discussion on multimodal fusion and recent architectures.

Recent works (2023–2024):

[154] C. P. Cheng, T. Owusu, P. Shekane, and A. M. Patel, "Sentiment analysis of pain physician reviews on healthgrades: a physician review website," Regional Anesthesia and Pain Medicine, vol. 49, no. 9, pp. 656–660, 2024.

[155] D. A. P. Nunes et al., "Computational analysis of the language of pain: A systematic review," arXiv preprint arXiv:2404.16226, 2024.

[156] I. Aggarwal, S. Joseph, N. Jaganathan, et al., "Sentiment analysis in health-care: A comparison of vader, bert, and flair nlp models on patient reviews of pain management hysicians," Cureus, 2024. Early online publication.

[156] A. Ghosh, S. Umer, B. C. Dhara, D. K. Jain, R. K. Rout, and A. Hussain,"A novel pain sentiment detection system utilizing a paincapsule model and textual facial patterns," Neurocomputing, p. 130907,

15.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

15.01.2026

Dr. Saiyed Umer
Assistant Professor
Dept. of CSE, Aliah University
Kolkata-700156

2025.

[158] R. Fang, E. Hosseini, R. Zhang, C. Fang, S. Rafatirad, H. Homayoun, et al., "Survey on pain detection using machine learning models: Narrative review," JMIR AI, vol. 4, no. 1, p. e53026, 2025.

[159] R. Fernandez-Rojas, C. Joseph, N. Hirachan, B. Seymour, and R. Goecke, "The ai4pain grand challenge 2025: Advancing pain assessment with multi-modal physiological signals," in Companion Proceedings of the 27th International Conference on Multimodal Interaction, pp. 147–152, 2025.

[226] 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, et al., "A review of voice-based pain detection in adults using artificial intelligence," Bioengineering, vol. 10, no. 4, p. 500, 2023.

[227] T.-Q. Dao, E. Schneiders, J. Williams, J. R. Bautista, T. Seabrooke, G. Vigneswaran, R. Kolpekwar, R. Vashistha, and A. Farahi, "Tame pain: Trust-worthy assessment of pain from speech and audio for the empowerment of patients," 2025.

[228] A. Lu, M. Kajol, W. Lu, and D. Sullivan, "Poster: Painnova: Privacy-aware voice-based pain-level detection," in Proceedings of the 2025 ACM SIGSAC Conference on Computer and communications Security, pp. 4761–4763, 2025.

[229] A. Ghosh, S. Umer, B. C. Dhara, and G. M. N. Ali, "A multimodal pain sentiment analysis system using ensembled deep learning approaches for iot-enabled healthcare framework," Sensors, vol. 25, no. 4, p. 1223, 2025.

[332] G. Bargshady, C. Joseph, N. Hirachan, R. Goecke, and R. F. Rojas, "Acute pain recognition from facial expression videos using vision transformers," in 2024 46th Annual International Conference of the IELE Engineering in Medicine and Biology Society (EMBC), pp. 1–4, IEEE, 2024.

[333] E. Holden et al., "From facial expressions to algorithms: A narrative review of animal pain assessment using artificial intelligence," Frontiers in Veterinary Science, vol. 11, p. 1436795, 2024.

14.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

13.01.2026

Dr. Saiyed Umer
Assistant Professor
Dept. of CSE, Aliah University
Kolkata-700156

[334] E. Holden et al., "Extensions of computer-vision-based facial pain detection for postoperative and clinical monitoring." Discussed in narrative review, 2024.

[335] C. W. Tan, T. Du, J. C. Teo, D. X. H. Chan, W. M. Kong, and B. L. Sng, "Automated pain detection using facial expression in adult patients with a customized spatial temporal attention long short-term memory (sta-lstm) network," Scientific Reports, vol. 15, no. 1, p. 13429, 2025.

[336] J.-T. Zhang, X.-Y. Hu, W. Duan, M.-H. Ji, and J.-J. Yang, "Application of deep learning-based facial pain recognition model for postoperative pain assessment," Journal of Clinical Anesthesia, vol. 105, p. 111898, 2025.

Overall, the encouraging feedback on the breadth, depth, and theoretical grounding of the literature review is gratefully acknowledged.

(The references have been incorporated into the 'Literature Review' sections of Chapters 2, 3, and 5.)

Comment-2: Material and methods :

However, the thesis would benefit from clearer experimental design details specifically regarding dataset composition, training/testing splits, evaluation metrics (accuracy, F1-score, etc.), and cross-validation strategy. Explicit quantitative comparisons between models are essential to validate claims of performance improvement. Overall, the materials and methods are robust, comprehensive, and systematically presented, though more experimental transparency would strengthen reproducibility.

Response-2: We appreciate the reviewer for the comment. The dataset is mentioned chapterwise; F-1 score and other metrics, along with the train-test split mechanism, are mentioned in **Chapter-1** and within the **'Experiments and Results'** section of each contributing chapter.

14.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

13.01.2026

Dr. Saiyed Umer
Assistant Professor
Dept. of CSE, Aliah University
Kolkata-700156

We sincerely thank the reviewer for the detailed and encouraging evaluation of the methodological framework. The recognition of the well-structured and comprehensive design encompassing multimodal preprocessing, feature extraction, classification, deep learning architectures, and multimodal integration is greatly appreciated.

The acknowledgement that the contribution of each modality is clearly articulated, the algorithmic choices are well justified, and the methodological explanation demonstrates technical competence validates the rigor of the proposed framework. The positive feedback on the clarity provided by the system architecture and block diagrams is also gratefully noted.

The constructive suggestions of the reviewer regarding more precise experimental design details like dataset composition, training and testing splits, evaluation metrics, cross-validation strategy, and explicit quantitative model comparisons are well addressed. Most of the datasets are collected by signing agreements with dataset providers, and one text dataset has been created using LLM. Comparisons are provided for each modality, along with the existing models for that modality.

Overall, the reviewer's assessment that the materials and methods are robust, comprehensive, and systematically presented is sincerely appreciated.

Comment-3: Results and Discussions :

However, the quantitative presentation could be enhanced through tabulated and graphical summaries of performance metrics, ablation studies, and error analyses. Further, statistical significance testing or confidence intervals would add credibility to the reported improvements.

Response-3: We are thankful to the reviewer for the positive assessment of the Results and Discussion section. The recognition of the comparative analysis across unimodal and multimodal systems, and the

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

Dr. Saiyed Umer
Assistant Professor
Dept. of CSE, Aliah University
Kolkata-700156

validation that multimodal fusion improves accuracy and robustness in pain detection, are greatly appreciated. The quantitative presentation is enhanced through tabulated and graphical summaries of performance metrics, ablation studies, error analyses and statistical significance testing in each contributing chapter of this thesis under the section **'Experiments and Results'**.

14.01.26

PROFESSOR
Deptt. of Information Technology
JADAVPUR UNIVERSITY
Block -LB, Plot-8, Sector-3
Salt Lake, Kolkata-700106, India

13.01.2026

Assistant Professor
Dept. of CSE, Alliah University
Kolkata-700156