**Response to Reviewers’ Comments**

|  |  |  |  |
| --- | --- | --- | --- |
| **Reviewer** | **Reviewer Comment** | **Response** | **Section Updated** |
| 1 | How did you identify the number of principal components to keep during PCA? How did you set t-SNE hyperparams (perplexity, learning rate, number of iterations)? Can you show that your results are not just a consequence of the particular PCA / t-SNE settings you used? | We retained 95\% of the variance through PCA to minimize information loss before applying t-SNE.  We employed the default hyperparameters for TSNE(n\_components=2, perplexity = 30, learning\_rate = "auto", n\_iter=300) | Methodology |
| 1 | Fig.1 mentions MultiHead Self Attention and use of a Standard Scaler. Neither are mentioned in the manuscript. | We acknowledge the inconsistency and such information has been added | Methodology |
| 1 | There is no information on how embeddings are constructed for each dataset, nor a description of the contents of the datasets or the expected clustering results. You mention using a sentence embedding model from Sentence-BERT. Is this applied to all five datasets? If so, please specify how you format the text for embedding in each case. If you use tabular features as embeddings please specify the sources of each feature. | Yes the Sentence BERT was used in all datasets. This part has been exaplained in detailed | Methodology |
| 1 | Table 4 does not specify the dataset used in the experiment | MIMIC-III was the dataset used in this Tabel 4. This dataset was created by us and was the most complex among those used, due to its structure, nature, and the presence of longer sentences. It was derived from clinical notes of patients and was designed to detect whether a patient was experiencing an adverse drug reaction. | Experiment Setup  -Comparative Analysis |
| 1 | Table 5 repeats the same exact DBI values for all four clustering algorithms. If this is a copying error, the actual DBI values are missing. Also, the constant data size used in the experiment is missing | I completely agree. There was an oversight in the DBI results due to a copying error, and the data size for the datasets was missing. These issues have been corrected, and the revisions are reflected in Table 5. | Experiment Setup  -Comparative Analysis |
| 1 | Figure 7 is missing the dataset used to run the experiment | MIMIC-III was the dataset used | Experiment Setup  -Comparative Analysis |
| 1 | Figure 8 is missing the constant data size used in the experiment | The constant data size used in Figure 8 is 4000 | Experiment Setup  -Comparative Analysis |
| 1 | The discussion section refers to SS-DBSCAN having better noise sensitivity than the baselines, but does not specify how SS-DBSCAN handles noise differently from the default DBSCAN behavior. Are noise points excluded from the computation of Silhouette when doing the fast grid search, or are you referring to something else? | We now clarified this. | Discussion  -Noise Sensitivity |
| 1 | No discussion of limitations as typically found in a limitations section | We have not yet identified specific limitations; however, we recommend further research to explore the applicability of our approach to other types of data, such as images, audio, and other modalities. | Discussion  -Future work |
| 1 | The Silhouette metric is used for both MinPts selection and experimental evaluation when comparing to baselines. However, it has been shown that Silhouette is not an appropriate metric for density-based clustering algorithms such as DBSCAN, specifically when dealing with irregularly (non-spherical) shaped clusters. Instead, the DBCV [1] metric should be used. | We have used both Silhouette and Davies-Bouldin Index metrics. Other metrics, such as Normalized Mutual Information (NMI) and Homogeneity Score, are available but are designed for labeled datasets. Since our study involves unlabeled datasets, these metrics were not applicable. | Experiment Setup |
| 1 | No justification is given that dimensionality reduction is needed for the SS-DBSCAN hyperparameter selection process. It is intuitive since dimensionality reduction is commonly used with clustering in general, but this needs to be shown either empirically or theoretically if you are claiming that PCA + t-SNE is a critical piece of your contribution. | We now justified the necessity of dimmentionality reduction | Methodology |
| 1 | Why PCA + t-SNE and not UMAP [2]? UMAP is a dimension reduction algorithm commonly used with clustering - would SS-DBSCAN work with UMAP as well as it does with t-SNE? If not, this would be a critical limitation that potential users would want to know about. | Our decision to use PCA followed by t-SNE was not meant to suggest that UMAP or other dimensionality reduction methods are inferior; instead, it was simply the approach that yielded the best results for our specific experiment. Therefore, we performed another experiment to prove this point. Along with this response matrix, we will attach a documents that shows the results. Yes SS-DBSCAN can work with UMAP as well, is not a limitation in this case. | Discussion and  A separate document. |
| 2 | Unclear how the contributions under Section 3 relate to each other. Is't (1) the same as (4)? Are (2) and (3) part of (1)? | The interconnection between the contributions has been demonstrated. Contribution (1) primarily focuses on complexity and high-dimensional data, while Contribution (4) emphasizes scalability and adaptability. | Contribution |
| 2 | Typo: "Contibution" > "Contribution" | This typo has been corrected. | Contribution |
| 2 | Methodology and Evaluation are not sufficiently separated. The Methodology section starts with "4.1 Data Preprocessing" detailing what datasets were used, which should be part of the evaluation (or maybe I am misunderstanding the purpose of this, in which case it should be clarified). First the conceptual approach should be fully described (e.g. in a "Methodology" section), before the details about the scientific evaluation are introduced (e.g. in a "Evaluation" section). | This has been addressed in both sections. We believe it is essential to mention the dataset used and its preprocessing in both sections to ensure clarity from a methodological perspective and within the experimental context. | Methodology and  Experiment Setup |
| 2 | Therefore it's unclear whether the pre-processing in Section 4.1 is part of the Approach or only done for the evaluation. | Data preprocessing is now clearly stated as a crucial step in SS-DBSCAN and all other baseline algorithms and not just for evaluation. This process is not uniform in all research, depends on the data used and more details is given to respond this. | Methodology |
| 2 | These datasets comprise sequences with lengths ranging from a minimum of 50 to a maximum of 500, with an average length of 250.": What datasets are these? Why were they selected? What's their role? What are "sequences" here? How are these minimums/maximums relevant and how were they chosen? | We now explicitly state the rationale behind dataset selection.  Sequences refer to text segments in the dataset.  The limits were chosen based on an initial exploratory analysis, ensuring that a majority of sequences retained essential information without unnecessary complexity. The average length of 250 reflects the natural distribution within the selected datasets | Experiment Setup and Methodology |
| 2 | Missing motivation why PCA, t-SNE, etc. are part of the approach. What purpose do they serve in the bigger picture. | Already addressed | Methodology |
| 2 | Overall, the approach isn't well introduced on a general level. Section 4 dives right into the details. It should first give a general and intuitive overview and motivation. | An overview and motivation subsection has been added at the beginning of the Methodology section to provide context before delving into the technical details. | Methodology  - overview and motivation |
| 2 | A novel stratified sampling technique": This should be better motivated too. What's the intuition behind this? Why could we expect this to work better than the alternatives? A conceptual diagram or something like that could help too | This has been addressed | Methodology  - overview and motivation |
| 2 | Section 5, "across multiple datasets, including...": It's unclear how these datasets were selected. | This has been explained | Experiment Setup |
| 2 | Section 5: How was the dataset used in "varying sizes"? What kind of sampling was applied? This should be explained better. | There was no predefined sample size for this type of research. Instead, experiments were conducted to demonstrate the scalability of our SS-DBSCAN algorithm. The comparative algorithms struggled with increasing data size, highlighting the robustness of our approach. | Experiment Setup |
| 2 | In Section 5, the paper relies too much on the visualizations. They are nice, but don't give scientific answers. I would only show a couple of the visualizations to give an impression for the reader. But the real results are in the tables, and they should get more prominence. | We apologize if the extensive visualizations were distracting. The intention was to illustrate how cluster formations evolve as data size increases for each algorithm. Additionally, these visualizations were supported by results presented in the tables. | Experiment Setup |
| 2 | The tables should better explain what they show. E.g. "more is better" and "less is better". The best results could be shown in bold. The caption should give us more indication what we are looking at. | The best results are now bolded to show the difference |  |
| 2 | For the subsections like "5.1.2. Clustering Results with DBSCAN", it should be made clearer whether this is used as a competing approach or baseline we are comparing your contribution against, OR whether this is a variant of showing the performance of your approach. Specifically, does 5.1.2 applying your approach under the hood too or not? | Our experimental setup focused on comparing four algorithms: SS-DBSCAN, DBSCAN, HDBSCAN, and OPTICS. The objective was to evaluate the performance of SS-DBSCAN against other DBSCAN variants. | Experiment Setup  Results Interpretation |
| 2 | Overall Section 5 is lacking structure. The short subsections and many images disturb the text flow. | We acknowledge that the images disrupt the text flow. This is a common challenge when using LaTeX, as it adjusts the text to fit around empty spaces where images cannot be placed. However, we have ensured that all images and tables are properly cited in the text for easy reference. | Experiment Setup |
| 2 | The datasets included Emotion-Sentiment, Coronavirus-Tweets, Cancer-Doc, and Sonar.": Again, why these? This shold be justified | This has been justified | Experiment Setup |
| 2 | Given that there are still many kinds of datasets out there that this approach was not tested on (which is natural and fine), I think the Discussion section should include a part where the generalization to other kinds of datasets/domains is discussed. Can we expect this to work there too? A statement like "we don't know; needs further work" is perfectly fine here, but I think this should be addressed. | We have explicitly mentioned this in the Discussion section, noting that future work is needed to assess its performance across a broader range of datasets. | Discussion  -Future work |
| 2 | Conclusion could be a bit more elaborate. Maybe picking up some points from the introduction again, and quickly summarizing the discussion points from Section 7. | We have expanded the conclusion as suggested | Conclusion |