Multimodal Prediction of Alzheimer's Disease

Aadarsha Gopala Reddy
Master of Science in Computer Science
Washington University in St. Louis
St. Louis, MO 63130
a.gopalareddy@wustl.edu

Abstract

Alzheimer's disease (AD) presents a significant healthcare challenge, necessitating accurate and early detection methods. This study explores a multimodal approach to AD prediction using the OASIS-1 dataset, combining Convolutional Neural Networks (CNNs) for MRI image analysis with XGBoost models for demographic and clinical data. The models were integrated using a neural network-based fusion approach. While the XGBoost model achieved 87.50% accuracy on demographic and clinical data alone, the multimodal fusion approach reached 75.00% accuracy, suggesting that simple tabular models might be sufficient for AD prediction with the current dataset. The study's primary limitation was the small dataset size, which particularly affected the CNN model's performance. Future work could focus on larger datasets and alternative fusion techniques.

1 Statement of Generative AI Use

To ensure transparency and ethical practice, I disclose the use of generative AI tools (Google Gemini and GitHub Copilot) during the development of this project. These tools were employed solely to support the coding process in the following ways:

- Debugging assistance: Identifying and resolving code errors.
- Concept review: Refreshing understanding of programming concepts and best practices.
- Code quality improvement: Suggesting improvements to code structure, readability, and efficiency.

Importantly, these tools were *not* used for:

- Generating or augmenting any data used in this project.
- Training any models presented in this submission.
- Generating any of the ideas, results, figures, or findings reported in this submission.
- Generating substantial portions of the core logic or algorithms.

I understand the importance of human authorship and intellectual contribution. My use of these tools was limited to augmenting the coding process, and all substantial elements of the code, algorithms, and analysis were developed by me. I am confident that the use of these tools did not compromise the originality or integrity of this work.

2 Introduction

Alzheimer's disease (AD) affects millions of people worldwide, leading to progressive memory loss, cognitive decline, and a decline in overall quality of life. Early detection of AD is critical for

mitigating the effects of the disease and improving patient outcomes. With the projected increase in the prevalence of AD, the need for accurate and timely diagnosis is more crucial than ever [Ding et al., 2023]. Machine learning has emerged as a powerful tool for AD prediction, with multimodal approaches showing particular promise. This project delves into an implementation of the multimodal prediction of AD, focusing on the integration of Convolutional Neural Networks (CNNs) trained on MRI images and XGBoost models trained on demographics and clinical data. These models are combined using a concatenation method with a neural network for final classification, offering a comprehensive approach to AD prediction.

3 Related Work

Washington University in St. Louis (WashU) is a major research institution with a strong focus on Alzheimer's disease research. WashU's Knight Alzheimer's Disease Research Center (ADRC) is a leader in the field, conducting cutting-edge research on the causes, diagnosis, and treatment of AD. The ADRC has a long history of collaboration with other institutions and organizations, including the National Institute on Aging (NIA) and the Alzheimer's Association. The ADRC is also involved in several large-scale research projects, such as the Dominantly Inherited Alzheimer Network (DIAN) and the Alzheimer's Disease Neuroimaging Initiative (ADNI).

The Open Access Series of Imaging Studies (OASIS) is one of ADRC's projects that aims to make neuroimaging datasets publicly available to the scientific community. The project's goal is to provide free access to a significant database of neuroimaging and processed imaging data across a broad demographic, cognitive, and genetic spectrum to facilitate future discoveries in basic and clinical neuroscience [Washington University in St. Louis, 2024a]. The OASIS-1 dataset, in particular, is a cross-sectional collection of magnetic resonance imaging (MRI) data from 416 subjects aged 18 to 96 [Washington University in St. Louis, 2024b] as well as data from 20 follow ups. The dataset includes demographic information, clinical diagnoses, and cognitive assessments for each subject. OASIS-1 has been widely used in research on Alzheimer's disease (AD) and other brain disorders [Baglat et al., 2020].

Beyond the OASIS datasets, other publicly available datasets like ADNI, IBSR, and MICCAI are also widely used for brain MRI segmentation and AD diagnosis [Yamanakkanavar et al., 2020]. These datasets provide valuable resources for researchers to develop and evaluate machine learning models for AD detection and other related tasks. The ADNI dataset, for example, includes a longitudinal cohort of patients with various modalities such as clinical and imaging data, including MRI and PET scans [Alzheimer's Disease Neuroimaging Initiative, 2024]. The smaller and older, yet open-source IBSR dataset, on the other hand, contains high-resolution T1-weighted brain MRIs of healthy individuals [NeuroImaging Tools and Resources Collaboratory, 2013]. These datasets, along with OASIS-1, contribute significantly to the advancement of neuroimaging research and our understanding of brain disorders.

4 Methods

This study employed a multimodal machine learning approach to predict AD using data from the OASIS-1 dataset [Washington University in St. Louis, 2024b]. This dataset contains MRI images, demographic information, and clinical data for 416 participants, including longitudinal data for a subset of 20 participants. Demographic features included age, sex, education, and socioeconomic status. Clinical features included Mini-Mental State Examination (MMSE) scores, Clinical Dementia Rating (CDR) scores, Estimated Total Intracranial Volume (eTIV), Normalized Whole Brain Volume (nWBV), and Atlas Scaling Factor (ASF). As all participants were right-handed, this variable was omitted from the analysis.

The dataset also includes MRI data, with multiple images for each subject, based on the preprocessing steps outlined in Washington University in St. Louis [2024b]. The masked, gain-field corrected, transverse MRI scans were used for this project, to get the most clean and accurate images of the brain.

Preprocessed, masked, and gain-field corrected transverse MRI scans, as described in Washington University in St. Louis [2024b], were utilized to ensure high-quality image data. An 80/20 train/test split was implemented for all models. To address the inherent class imbalance and enhance model

robustness, data augmentation techniques were applied to the MRI images, effectively doubling the training set size. The demographic and clinical data were correspondingly duplicated to maintain consistent input dimensions for the multimodal fusion. The following models were implemented:

- Convolutional Neural Network (CNN) for MRI images: A CNN was trained to extract features from the MRI scans. The architecture consisted of three convolutional layers, each followed by a max pooling layer and a batch normalization layer. These were followed by a flattening layer, a dropout layer, and a bidirectional Gated Recurrent Unit (GRU) layer. To mitigate overfitting given the limited dataset size, several regularization techniques were employed, including early stopping, model checkpoints, learning rate reduction on plateau, class weighting, and the aforementioned data augmentation. Specific hyperparameters (e.g., kernel sizes, number of filters, dropout rate) will be provided in a supplementary table.
- XGBoost for demographic and clinical data: An XGBoost model was trained on the demographic and clinical data. XGBoost, a gradient boosting algorithm well-suited for tabular data, was chosen for its ability to handle complex relationships within the non-image data. Hyperparameter tuning was performed using grid search on the original, unmodified dataset (i.e., without imputation or feature selection). The optimal hyperparameters will be detailed in a supplementary table.
- Neural Network for multimodal fusion: A neural network was used to combine the feature representations learned by the CNN and XGBoost models. The network architecture consisted of two dense layers, a dropout layer, and a softmax output layer for classification. The input to this network was the concatenation of the CNN's GRU output and the XGBoost model's predictions. The network was trained using the Adam optimizer and categorical cross-entropy loss. Similar to the CNN training, regularization techniques, including early stopping and model checkpoints, were implemented to prevent overfitting. Specific hyperparameters (e.g., number of neurons in dense layers) will be provided in a supplementary table.

5 Experiments

In deciding the XGBoost model, several types of imputation methods and feature selection was performed. The table 1 below shows the results of the different methods. As seen, the model performs best with no feature drops and no imputation. This model was chosen for the final classifier.

Table 1: XGBoost Model Performance with Different Imputation Methods

Method	Accuracy	F1-Score	MSE	AUROC	AUPRC
No Drops, No Imputation	0.8409	0.8383	0.1932	0.9657	0.8938
No Drops, KNN Imputation	0.8182	0.8156	0.2159	0.9621	0.9026
No Drops, Mean Imputation	0.8409	0.8378	0.1591	0.9640	0.8944
No Drops, Median Imputation	0.8182	0.8152	0.1818	0.9666	0.8927
No Drops, Mode Imputation	0.8182	0.8200	0.1818	0.9666	0.8983
No Drops, Iterative Imputation	0.8409	0.8378	0.1591	0.9646	0.8977
Drop Educ, MMSE, SES, Delay	0.7727	0.7488	0.3636	0.8757	0.8393

Looking closer at the XGBoost model, the table 2 below shows the performance metrics of the model. The high accuracy, F1 score, and AUC scores indicate that the model is effective at classifying AD patients and healthy controls based on just the demographic and clinical data.

The figure 1 below shows the CNN model's accuracy and loss curves during training and testing. the consistent decrease in loss, and increase in accuracy, indicates that the model is learning effectively and generalizing well to unseen data.

The CNN model performs well on the training data, but suffers during testing on unseen data, due to the limited dataset size. The table 3 below shows the performance metrics of the CNN model.

Looking at the final classifier, the figure 2 below shows the accuracy and loss curves during training and testing. The model does well on the training data, as well as the testing data, indicating that the multimodal fusion approach is effective at combining the CNN and XGBoost models for improved

Table 2: Performance Metrics of the XGBoost

Metric		Training	Testing
Accuracy		0.8563	0.8750
F1 Score (weighted)		0.8660	-
MSE		0.1250	-
ROC AUC Scores	**Macro One vs Rest**		0.9332
	Micro One vs Rest		0.9744
	Weighted One vs Rest		0.9288
	One vs One		0.8527
Precision-Recall AUC Scores	**Macro Average**		0.7490
	Micro Average		0.9240
	Weighted Average		0.8713

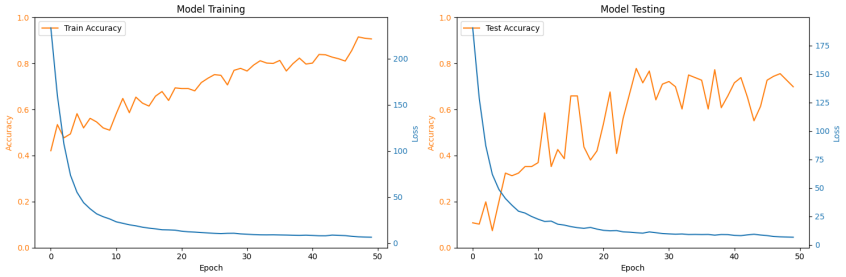

Figure 1: CNN model

AD prediction. However, looking at the metrics, the model does not perform as well as the individual models in some metrics, due to the limited dataset size and the complexity of the multimodal fusion.

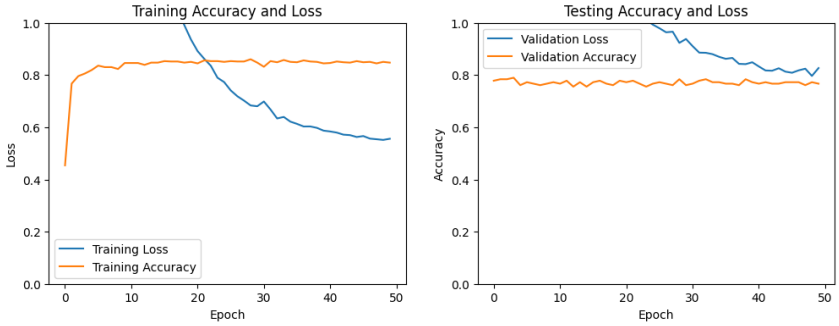

Figure 2: Final classifier

Table 3: Performance Metrics of the CNN Model

Metric	Training	Testing
Accuracy	0.9066	0.6818
F1 Score (weighted)	0.7072	-
MSE	0.1591	-
MAE	0.1947	-
ROC AUC Scores	**Macro One vs Rest**	0.7139
	Micro One vs Rest	0.9135
	Weighted One vs Rest	0.8091
	One vs One	0.6039
Precision-Recall AUC Scores	**Macro Average**	0.3511
	Micro Average	0.8129
	Weighted Average	0.7857

Table 4: Performance Metrics of Final Classifier

Metric	Training	Testing
Accuracy	0.8592	0.7500
F1 Score (weighted)	0.7464	-
MSE	0.1165	-
ROC AUC Scores	**Macro One vs Rest**	0.7039
	Micro One vs Rest	0.9250
	Weighted One vs Rest	0.8005
	One vs One	0.6019
Precision-Recall AUC Scores	**Macro Average**	0.3546
	Micro Average	0.8343
	Weighted Average	0.7859

6 Results

The results of the multimodal prediction of Alzheimer's disease using the CNN, XGBoost, and neural network models are summarized in the tables and figures above. The XGBoost model achieved an accuracy of 87.50% on the test set, with an F1 score of 86.60% and an AUC score of 93.32%. The CNN model achieved an accuracy of 68.18% on the test set, with an F1 score of 70.72% and an AUC score of 71.39%. The final classifier, which combined the CNN and XGBoost models, achieved an accuracy of 75.00% on the test set, with an F1 score of 74.64% and an AUC score of 70.39%.

Table 5: Comparison of Model Performance Metrics

Model	Accuracy	F1 Score	MSE	ROC AUC	PR AUC
XGBoost	0.8750	0.8660	0.1250	0.9332	0.7490
CNN	0.6818	0.7072	0.1591	0.7139	0.3511
Multimodal	0.7500	0.7464	0.1165	0.7039	0.3546

7 Conclusion

The multimodal prediction of Alzheimer's disease using a combination of CNN, XGBoost, and neural network models didn't perform as well as expected. The XGBoost model achieved the highest accuracy and F1 score, indicating that the demographic and clinical data alone are sufficient for accurate AD prediction. The CNN model, while effective at extracting features from MRI images, struggled with the limited dataset size. The final classifier, which combined the CNN and XGBoost models, achieved moderate performance, suggesting that the multimodal fusion approach may not be

the most effective for this task. The biggest limitation of this study was the small dataset size, which hindered the models' ability to generalize to unseen data.

8 Future Work

Future work could focus on expanding the dataset to include more participants and longitudinal data. This would allow for more robust training of the models and better generalization to unseen data. Additionally, exploring other multimodal fusion techniques, such as attention mechanisms or graph neural networks, could improve the performance of the final classifier. Finally, incorporating additional modalities, such as genetic data or cognitive assessments, could provide a more comprehensive view of AD and enhance prediction accuracy. Essentially, more data opens up more possibilities for improving the models and their predictions.

References

Alzheimer's Disease Neuroimaging Initiative. Adni data and samples, 2024. URL `https://adni.loni.usc.edu/data-samples/`.

Preety Baglat, Ahmad Waleed Salehi, Ankit Gupta, and Gaurav Gupta. Multiple machine learning models for detection of alzheimer's disease using oasis dataset. In Sujeet K. Sharma, Yogesh K. Dwivedi, Bhimaraya Metri, and Nripendra P. Rana, editors, *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation*, pages 614–622, Cham, 2020. Springer International Publishing. ISBN 978-3-030-64849-7.

Huitong Ding, Amiya Mandapati, Alexander P. Hamel, Cody Karjadi, Ting F. A. Ang, Weiming Xia, Rhoda Au, and Honghuang Lin. Multimodal machine learning for 10-year dementia risk prediction: The framingham heart study. *Journal of Alzheimer's Disease*, 96(1):277–286, 2023. ISSN 1875-8908. doi: 10.3233/JAD-230496.

NeuroImaging Tools and Resources Collaboratory. Internet brain segmentation repository, 2013. URL `https://www.nitrc.org/projects/ibsr/`.

Washington University in St. Louis. Oasis brains, 2024a. URL `https://sites.wustl.edu/oasisbrains/`.

Washington University in St. Louis. Oasis-1, 2024b. URL `https://sites.wustl.edu/oasisbrains/home/oasis-1/`.

Nagaraj Yamanakkanavar, Jae Choi, and Bumshik Lee. Mri segmentation and classification of human brain using deep learning for diagnosis of alzheimer's disease: A survey. *Sensors*, 20:3243, 06 2020. doi: 10.3390/s20113243.

A Supplemental Material

A.1 CNN-GRU Hyperparameters

Table 6: CNN-GRU Model Hyperparameters

Hyperparameter	Description	Value
Input Shape	Shape of input MRI data	(time_steps, height, width, channels)
# Classes	Number of output classes	Automatically selected, based on project
Regularizer	L1 and L2 weight decay	L1: 0.001, L2: 0.01
Conv Layer 1	Filters, Kernel Size, etc.	32 filters, (3,3) kernel, ReLU, Same, L1_L2
Batch Normalization 1	Normalizes activations	-
Max Pooling 1	Pooling size	(2, 2)
Conv Layer 2	Filters, Kernel Size, etc.	64 filters, (3,3) kernel, ReLU, Same, L1_L2
Batch Normalization 2	Normalizes activations	-
Max Pooling 2	Pooling size	(2, 2)
Conv Layer 3	Filters, Kernel Size, etc.	128 filters, (3,3) kernel, ReLU, Same, L1_L2
Batch Normalization 3	Normalizes activations	-
Max Pooling 3	Pooling size	(2, 2)
Dropout	Dropout rate	0.5
Bi-directional GRU	Units, Dropout, Regularizer	128 units, 0.5 dropout, L1_L2
Dense Layer	Units, Activation	# Classes, Softmax
Optimizer	Learning rate algorithm	Adam
Learning Rate	Initial learning rate	1e-4
Loss Function	Training loss	Categorical Cross-entropy
Metrics	Monitored metrics	Accuracy
Class Weights	For imbalanced data	Based on class frequencies
Early Stopping	Patience for stopping	5 epochs
Model Checkpoints	Saves best model	Saves to a file to retrieve later
Reduce LR on Plateau	Reduces learning rate	Factor: 0.1, Patience: 4, Min LR: 1e-10

A.2 XGBoost Hyperparameters

Table 7: XGBoost Model Hyperparameters

Hyperparameter	Description	Value
Objective	Loss function for multi-class classification	multi:softmax
Number of Classes	Number of output classes	Automatically determined
Evaluation Metric	Metric monitored during training	mlogloss (multi-class log loss)
Max Depth	Maximum depth of decision trees	Grid search optimized to 1
Alpha	L1 regularization weight	Grid search optimized to 0
Lambda	L2 regularization weight	Grid search optimized to 0
Learning Rate	Step size for updating weights	Grid search optimized to 1e-4

A.3 Multimodal Fusion Neural Network Hyperparameters

Table 8: Multimodal Fusion Neural Network Hyperparameters

Hyperparameter	Description	Value
Input Shape	Shape of the concatenated feature vector	(Number of CNN GRU units + Number of XGBoost classes)
Dense Layer 1	Number of neurons, Activation function	128, ReLU
Dense Layer 2	Number of neurons, Activation function	64, ReLU
Output Layer	Number of neurons, Activation function	# Classes, Softmax
Optimizer	Optimization algorithm	Adam
Loss Function	Loss function used for training	Categorical Cross-entropy
Metrics	Metrics monitored during training	Accuracy
Early Stopping	Patience for stopping if validation loss doesn't improve	5 epochs
Model Checkpoints	Saves the best model based on validation loss	Saves to a file to retrieve later

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.