Classification of Melanoma Skin Cancer with Ensemble Learning and Stratified K-Fold Validation

Anonymous Author(s)
Affiliation
Address
email

Abstract

In recent years, many AI models have been developed to aid physicians in the diagnosis of different types of skin cancer. However, little progress has been made in providing accurate diagnoses before meeting a physician, which could potentially save large amounts of time for all parties involved. In this work, we demonstrate the potential of using large model ensembles to provide highly accurate estimations for the presence of skin cancer from a given image. Our best ensemble reached a peak pAUC-above-80 score of 0.171. In addition, we showcase the significant improvement that can be made through various augmenting and preprocessing techniques. Our work also has the novel use of Quadruple Stratified Leak-Free KFold Cross-Validation in medical areas.

1 Introduction

1.1 Backround

Skin cancer, which comprises both melanoma and non-melanoma types, is one of the most common cancers globally, with millions of cases diagnosed annually. Around 92,000 new cases of melanoma and 2,750,000 cases of nonmelanocytic skin cancer are estimated to occur worldwide each year, with a large number of cases going unreported each year [1]. Early detection of skin cancer is crucial for improving survival rates, as the treatments are more effective during the early stages of the disease when the cancer has not metastasized to other parts of the body. However, accurate diagnosis requires specialized dermatologists, which is problematic in areas with limited healthcare access. In recent years, the application of artificial intelligence in medical imaging has garnered a significant amount of attention in aiding the early diagnosis of diseases.

Deep learning models, particularly convolutional neural networks (CNNs), have been shown to be capable of learning complex patterns and features from large datasets of dermoscopic images [2]. Such models have also shown the ability to classify skin lesions with accuracy that sometimes can exceed that of experienced dermatologists. In a study by Esteba et al.in 2017, a deep neural network was trained on more than 129,000 images of skin lesions, achieving performance on par with a group of 21 board-certified dermatologists. [3]. Despite these advancements, however, AI models especially in clinical settings are generally employed as decision-support systems rather than a standalone diagnostic tool [4]. This is mainly because of concerns about the generalizability and reliability of AI models across various different datasets and medical environments.

One significant limitation currently in the development of predictive models for skin care diagnosis is the variability of dermoscopic image datasets as they often contain images with inconsistent lighting, different image resolutions, and a wide range of skin tones and lesion types. Furthermore, skin lesions can present differently depending on the patient's age, skin type, and the type of cancer [5]. Additionally, these medical images are affected by details such as hair, shadows, and reflections, which

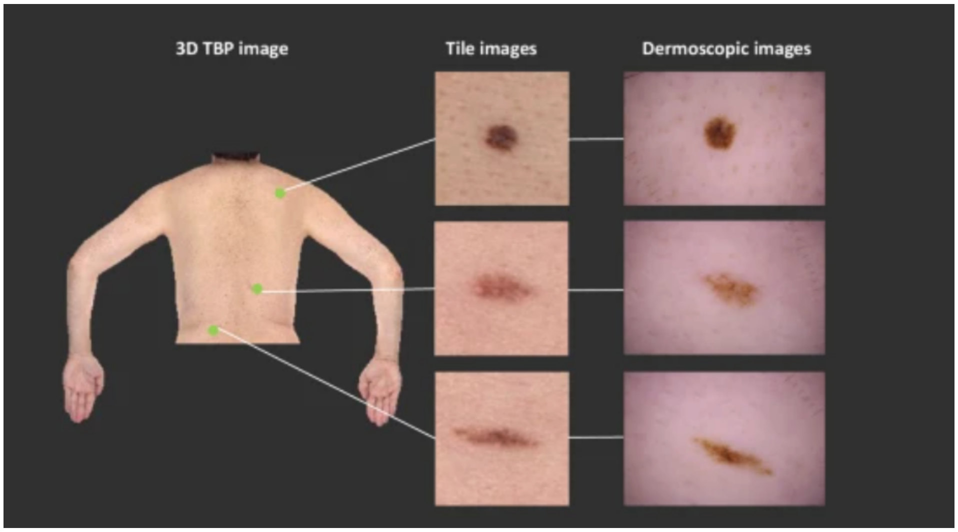

Figure 1: Examples of Extracted Lesions

further complicates such accurate diagnosis. Addressing these issues requires strong preprocessing techniques and data augmentation tools to ensure model robustness.

Ensemble learning, however, has emerged as a promising approach to solving these issues. Ensemble methods combine the predictions of multiple different models to improve the overall accuracy and robustness of models [6]. Particularly in the field of medical image analysis, ensemble methods have shown superior performance compared to approaches with a single model. Ensemble methods can decrease the weaknesses of individual models, which leads to improved diagnostic accuracy by aggregating predictions from multiple models. For example, Swin Transformers and ConvNeXt models have been applied to medical images to capture both local and global features of skin lesions [7].

1.2 Dataset

For this task, we utilized the SLICE-3D dataset [8], a set of over 400,000 cropped images of skin lesions from dermatologic centers across the world. Skin lesions are parts of one's skin that differ from the surrounding area, and should be classified as benign (non-cancerous) or malignant (cancerous). The images used in this dataset were extracted from 3D total body photographs. Through the use of AI software to identify all lesions on a patient, data was collected from thousands of patients across the world from 2015-2024. An example of such is shown in Fig. 1.

Along with the provided images, each lesion also had labels such as unique patient IDs (there were multiple lesions per patient), sex, approximate age, location of the lesion, and maximum diameter of the lesion. Every lesion image was 140x140 pixels and had an assigned probability score for whether it was cancerous or not.

2 Preprocessing

2.1 Data Augmentation

We found that our model's accuracy greatly increased when using a series of data augmentations to manipulate the data in a more usable manner. The first group of transformations that significantly improved model accuracy was rotations and flips: Each image was randomly transposed by rotating it 90 degrees. Additionally, there were random vertical and horizontal flips for each image along their vertical and horizontal axes. Finally, the images were randomly shifted and scaled.

The next group of augmentations altered the content of the images themselves. The contrast was randomly adjusted for every image. Each of the following blur/noise transformations was randomly

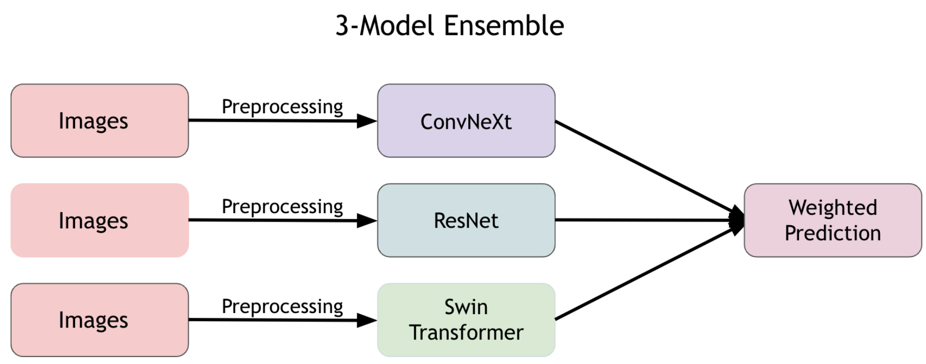

Figure 2: Model Architecture

selected and had an 80% chance of being applied: motion blur, median blur, Gaussian blur, and Gauss noise. Each of the following geometric distortions was randomly selected and had an 80% chance of being applied: optical distortion, grid distortion, and elastic transform.

Regarding the image's color, its hue, saturation, and brightness of the image were randomly shifted. This was done to mimic a change in lighting.

2.2 Data Preprocessing

We applied Contrast Limited Adaptive Histogram Equalization (CLAHE) [9] to improve contrast in localized areas of each image. This was especially helpful because of the subtle details that could indicate the presence of melanoma. Some other basic preprocessing techniques were used, such as resizing the images, randomly cutting out portions of the image, and normalizing the pixel values.

In order to highlight the major improvements all these various preprocessing techniques had, we created the table below to show the changes in pAUC-above-80 [10] scores (which will be elaborated on later) for our model.

Table 1: pAUC-above-80 scores with Different Types of Preprocessing

Ensemble with	pAUC-above-80
No Preprocessing	0.152
Only Basic Preprocessing	0.155
Only Rotations and Flips	0.167
Only Content Augmented	0.169
Only Color Augmented	0.169
All Augmentations and Preprocessing	0.171

3 Architecture

The final ensemble of models used in training was a ConvNeXt [11], a Residual Neural Network (ResNet) [12], and a Swin Transformer [13](pictured above in Fig. 2).

Prior studies have shown that ConvNeXts have a very high performance on image data, and their integration of CNNs and transformers allows for advanced designs. Their hierarchical feature extraction lets them learn high-level and low-level features, which is especially important for classifying melanoma. Additionally, these models have efficient learning and robust generalization.

We also used ResNets because of their deep architectures–the residual connections they have is very helpful for training deep networks. Additionally, in recent image classification benchmarks, ResNets have been shown to have state-of-the-art accuracy, making them a good fit for the ensemble.

Finally, the third model in the ensemble was a Swin transformer, known for efficiently handling high-resolution images. We thought this would be an effective complement to the other two models because of this focus. Additionally, they have advanced feature extraction and provide a unique set of features that are less likely to be captured by regular CNNs.

The different accuracies achieved by each model alone and together are shown in the table below.

Table 2: Performance by Individual Models

Model	**pAUC-above-80**
ConvNeXt	0.159
ResNet	0.166
Swin Transformer	0.162
Full Ensemble	0.171

4 Training

In training, we used a batch size of 32 and a learning rate of 0.001. The optimizer was Adam and the weight decay was 0.001. We found that the adaptive learning rate from using Adam [14] helped greatly with the model's accuracy. It is also especially useful for this case because of how it ensures that parameters related to crucial features are updated more effectively than less important features. The ReLU [15] activation function was also used, because of its ability to model complex relationships between input features and output labels. We found that it had higher accuracies than other popular activation functions like tanh and sigmoid.

4.1 Quadruple Stratified Leak-Free K-Fold Cross-Validation

To counter some of the issues we thought would be an issue during training, we utilized triple stratified cross-validation for the best results.

In this dataset, each patient had many images of skin lesions, meaning the same patient could appear multiple times. If some images of patients were in the training set while others were in the validation set, this could have caused data leakage, so we implemented the "patient isolation" strategy by ensuring all images from a single patient were in the same fold.

Another issue with the dataset was that there was a large class imbalance between benign and malignant lesions (malignant lesions only made up around 2% of the images). In response, we stratified the training so each fold had the same proportion of malignant images.

There were also many different image counts for different patients. As mentioned, most patients had more than one image in the dataset, but the distribution of the number of images they had was relatively large. To address this issue, we grouped patients based on the number of images they had in the dataset. In the end, each fold had a similar distribution of patients with different image counts.

The final stratification had to do with the diameter of each lesion. Because the distribution of diameter sizes was relatively large, we decided to stratify them, ensuring that each fold had a similar distribution.

4.2 Learning Rate Scheduler

We coupled the Adam optimizer together with a learning rate scheduler, which started the rate low and increased it quickly, finally decreasing it slowly at the end. This was done so the model could quickly learn in the beginning, and stabilize later on so it would converge to its maximum accuracy. When detecting small differences in skin lesions, stability was crucial—noisy gradients would cause instability and misclassification.

4.3 pAUC-above-80 Score

To evaluate the performance of the ensemble, we opted to use the partial area under the ROC curve above 80% true positive rate. We thought this would be a better indicator of performance than

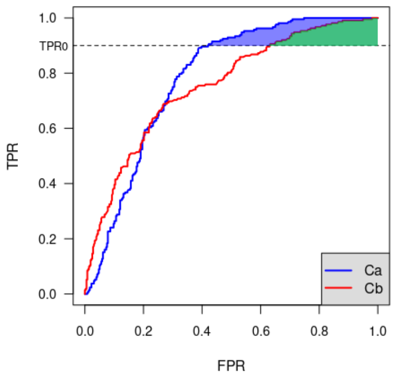

Figure 3: Example of pAUC curve

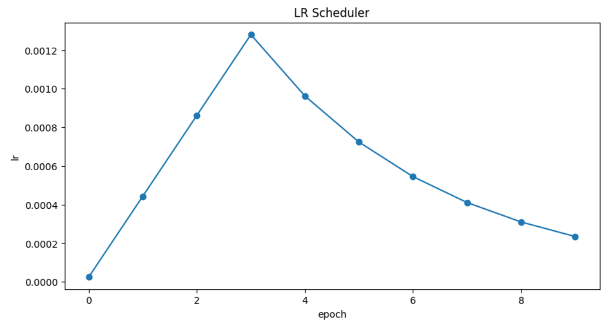

Figure 4: Learning Rate Scheduler

accuracy because of the specific part of the ROC curve we wanted to assess. In actual clinical practice, certain parts of the ROC curve are not as important, and in the case of diagnosing cancer, systems are required to be highly sensitive, which is why only the area above the 80% true positive rate is evaluated. This would mean the maximum score is 0.20, which means our final ensemble with a score of 0.171 performed very well, considering other models trained on this dataset. An example of a partial area under the ROC curve is shown in Fig. 3.

4.4 Results

After all the optimizations we made, our final ensemble with all preprocessing and stratification had an accuracy of 0.171, which was extremely high compared to the benchmark models trained on this data, which had a score of 0.168. We also implemented some basic hyperparameter tuning at the end to polish the model. We also tried other models in our ensemble, mainly tree algorithms, but they didn't yield much results. XGBoost, LGBM, and CatBoost were all implemented. Their results are shown below.

Table 3: Performance by Other Models

Model	**pAUC-above-80**
XGBoost	0.145
LGBM	0.132
CatBoost	0.142

5 Conclusion

To conclude, in our work, we explored the possibility of creating an AI model for classifying skin lesions as cancerous or noncancerous. Although such models have helped physicians in recent years, there has not been much progress in developing models for diagnosing skin cancer before even visiting a physician.

We discussed the novel application of many techniques in medical imagery, such as all the data preprocessing and data augmentation we did to increase the accuracy of our model. We also proposed the use of Quadruple Stratified Leak-Free K-Fold Cross-Validation to address any flaws there might have been within the dataset. Our final model performed very well compared to other benchmark models for this dataset, scoring a 0.171 pAUC-above-80 score.

5.1 Future Work

In the future, we hope to test our theories with preprocessing and augmentation with other medical image datasets. Additionally, we would like to utilize more hyperparameter tuning in our model to improve it further. Finally, we are still looking into other models that may be able to improve our overall ensemble.

References

[1] B. K. Armstrong and A. Kricker, “Skin Cancer,” Dermatologic Clinics, vol. 13, no. 3, pp. 583–594, Jul. 1995, doi: https://doi.org/10.1016/s0733-8635(18)30064-0.

[2] T.-C. Pham, C.-M. Luong, V.-D. Hoang, and A. Doucet, “AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function,” Scientific Reports, vol. 11, no. 1, p. 17485, Sep. 2021, doi: https://doi.org/10.1038/s41598-021-96707-8.

[3] A. Esteva et al., “Dermatologist-level Classification of Skin Cancer with Deep Neural Networks,” Nature, vol. 542, no. 7639, pp. 115–118, Jan. 2017, doi: https://doi.org/10.1038/nature21056.

[4] P. Tschandl et al., “Human–computer collaboration for skin cancer recognition,” Nature Medicine, vol. 26, no. 8, pp. 1229–1234, Aug. 2020, doi: https://doi.org/10.1038/s41591-020-0942-0.

[5] T. J. Brinker et al., “Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review,” Journal of Medical Internet Research, vol. 20, no. 10, p. e11936, Oct. 2018, doi: https://doi.org/10.2196/11936.

[6] T. G. Dietterich, “Ensemble Methods in Machine Learning,” Multiple Classifier Systems, vol. 1857, pp. 1–15, 2000, doi: https://doi.org/10.1007/3-540-45014-9$_1$.

[7] P. Vasuki, “A survey on image preprocessing techniques for diverse fields of medical imagery,” IEEE Proceedings, 2014.

[8] International Skin Imaging Collaboration. SLICE-3D 2024 Challenge Dataset. International Skin Imaging Collaboration https://doi.org/10.34970/2024-slice-3d (2024).

[9] G. Yadav, S. Maheshwari and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Delhi, India, 2014, pp. 2392-2397, doi: 10.1109/ICACCI.2014.6968381.

[10] Wikipedia Contributors, “Partial Area Under the ROC Curve,” Wikipedia, Jul. 27, 2024. https://en.wikipedia.org/wiki/Partial$_{A}rea_{U}nder_{t}he_{R}OC_{C}urve(accessedSep.10, 2024).$ °

[11] S. Woo et al., "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 16133-16142, doi: 10.1109/CVPR52729.2023.01548.

[12] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[13] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002, doi: 10.1109/ICCV48922.2021.00986.

[14] D. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Dec. 2014.

[15] A. M. Javid, S. Das, M. Skoglund and S. Chatterjee, "A ReLU Dense Layer to Improve the Performance of Neural Networks," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 2810-2814, doi: 10.1109/ICASSP39728.2021.9414269.

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: Yes, all claims made in the abstract reflect the scope of the paper. We discuss the preprocessing techniques we used and the stratification done.

   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: We talk about how there are still flaws with the ensemble and that it can be improved further.

   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: No theoretical results.

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: [NA]

Justification: No experiments.

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: [Yes]

Justification: Open access to the dataset is provided. All the instructions required to replicate results are elaborated on fully. Although we can't release the code at this time, everything is made clear.

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: We talk about hyperparameters, how they were chosen, the optimizers, the learning rates, etc.

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: No experiments.

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: [NA]

   Justification: No experiments.

   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: All anonymity is preserved and everything in the code of ethics is followed.

   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: [Yes]

    Justification: Yes, all the societal impacts of the research are discussed fully.

    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: The paper poses no such risks.

    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: [Yes]

    Justification: Everyone is properly credited.

    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: The paper does not release new assets.

    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: No crowdsourcing nor research with human subjects.

    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: No crowdsourcing nor research with human subjects.

    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.