COMP 6721 Applied Artificial Intelligence (Fall 2023)

Project Assignment, Part III

Due date (Moodle Submission): Monday, December 4
Counts for 45% of the course project

“A.I.ducation Analytics”: Final Version. Congratulations on reaching this pivotal phase of the project, where you will integrate all you’ve learned — from machine learning fundamentals to deep learning with Convolutional Neural Networks (CNNs), from rigorous training practices to nuanced evaluation methods — into crafting a complete, ethically-aligned AI system. This final part focuses on evaluating and mitigating potential *biases* in your AI, along with addressing any issues identified in the second part’s demo.

Bias in AI. As AI-based applications increasingly become a part of everyday life, the analysis and mitigation of *biases* – often inadvertently introduced through training data – have emerged as critical concerns.[1]

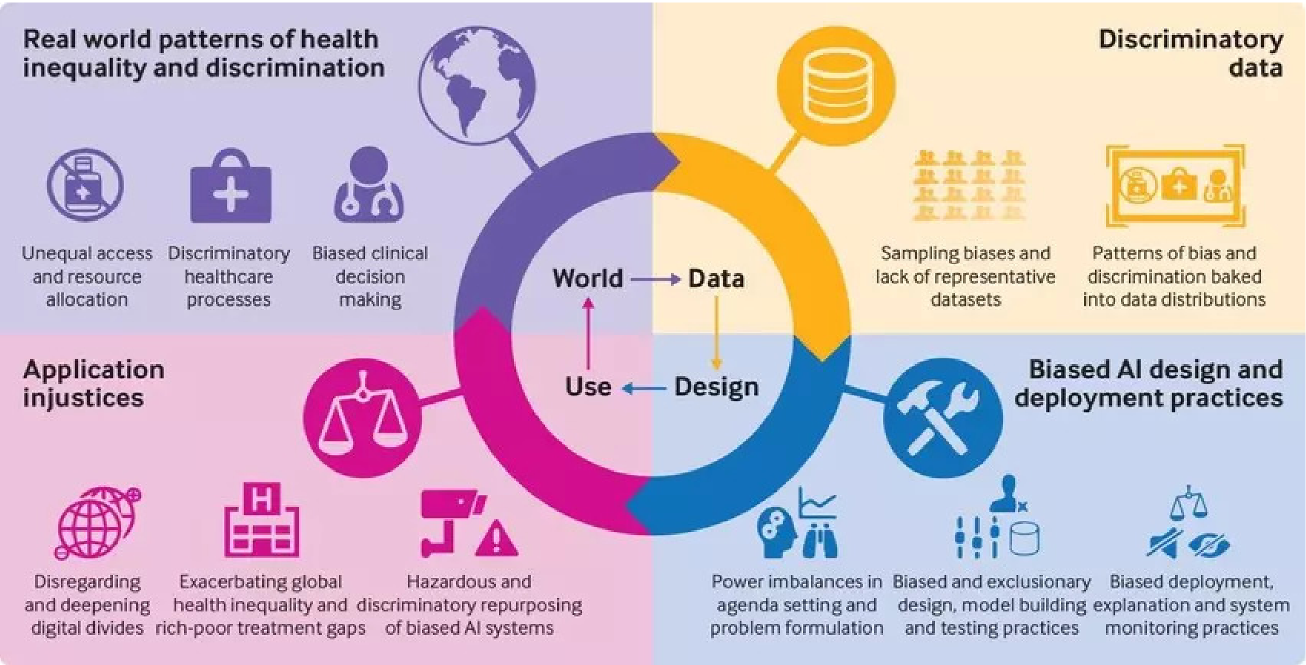
Real world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World
Data
Use
Design
Application injustices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Biased AI design and deployment practices
Power imbalances in agenda setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices

Figure 1: Bias in AI (World Economic Forum[2])

In this part of the project, your challenge is to ensure that your AI system is not only technically proficient but also ethically sound. You are required to analyze your AI for **two** of the following

[1] For instance, see the case of *“Amazon scrapping a secret AI recruiting tool due to bias against women”*, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[2] https://www.weforum.org/agenda/2021/07/ai-machine-learning-bias-discrimination/

categories: *age, race,* or *gender.*[3] For example, using suitably annotated testing data, investigate whether your system's performance across the four classes remains consistent across different genders.[4] Experiment with methods such as re-balancing or augmenting your training dataset to address any identified biases. Your evaluation should encompass the network's performance on both the complete dataset and the subsets representing the chosen biased attributes.

Evaluation: K-fold Cross-validation. Building upon the basic train/test split methodology from Part II, this phase requires you to employ *k-fold cross-validation* to enhance the robustness and reliability of your evaluation, especially across different classes.[5]

You are instructed to perform a 10-fold cross-validation (with random shuffling) on your AI model (the final model from Part II). This method involves dividing your dataset into 10 equal parts, using each part once as a test set while the remaining nine parts serve as the training set. This process is repeated 10 times, each time with a different part as the test set. This technique helps in assessing the model's performance more comprehensively and reduces the bias that can occur in a single train/test split.

Utilize the functionalities provided by *scikit-learn* or *skorch* for implementing the k-fold cross-validation.[6] It is important to avoid a manual, static split of the dataset for this process.

Document the results of your 10-fold cross-validation in the project report, using the format shown in Table 1. This includes the performance metrics (accuracy, precision, recall, F1-score) for each fold, as well as the average across all 10 folds.

In your report (see below), include a summary that discusses the overall performance consistency and any notable variations observed across different folds.

Fold	**Macro**			**Micro**			**Accuracy**
	P	**R**	**F**	**P**	**R**	**F**	
1	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-
...	...	...	...	...	...	...	...
10	-	-	-	-	-	-	-
Average	-	-	-	-	-	-	-

Table 1: How to report the performance metrics for each fold in the 10-fold cross-validation, together with the average over all folds. Use this table format for (i) your model from Part II of the project and (ii) your final, updated model from Part III.

Detecting and Mitigating Bias. This critical phase of your project involves analyzing your AI model for potential biases and implementing strategies to mitigate them. The process is divided into three main steps: data collection, analysis for bias, and retraining the system if needed. Start from the model you developed and saved in Part II of the project for the initial bias evaluation.

[3] If your group has only two members, you have to analyze only one category.

[4] Refer to https://www.media.mit.edu/projects/gender-shades/overview/ for an example of systematic AI product testing.

[5] For an overview of k-fold cross-validation, visit https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-iterators

[6] Refer to https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html for detailed guidance on

Attribute	**Group**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Age	Young	-	-	-	-
	Middle-aged	-	-	-	-
	Senior	-	-	-	-
	Average	-	-	-	-
Gender	Male	-	-	-	-
	Female	-	-	-	-
	Other/Non-binary	-	-	-	-
	Average	-	-	-	-
Overall System Average		-	-	-	-

Table 2: Reporting the result from the bias analysis in this table. For your two selected bias attributes (here "age" and "gender" as an example), report the macro-averaged metrics as shown above. Use this table layout for both (i) your model from Part II and (ii) your updated, final model from Part III. Note: you may have different groups for each attribute (e.g., more age groups), depending on your dataset.

Data Collection for Bias Analysis: Segment your dataset based on the chosen attributes (e.g., age, gender, race). Ensure that these attributes are accurately annotated in your dataset, either through existing labels or through additional annotation efforts as part of this project. Evaluate the saved model from Part II on each group separately using a standard train/test split (like you did in Part II), collecting macro-averaged performance metrics (Accuracy, Precision, Recall, F1-Score) for each demographic group. Fill these metrics into the Bias Analysis Table (Table 2). Repeat for your second chosen attribute.

Analysis of Collected Data: Analyze the filled Bias Analysis Table to identify any significant performance discrepancies across different demographic groups, which may indicate biases. Pay close attention to variations in performance metrics among the groups within each attribute.

Retraining and Mitigation: If biases are detected, consider dataset augmentation to enhance the representation of underrepresented groups. This might include sourcing additional images that accurately represent these groups, exploring different lighting conditions, backgrounds, or responsibly using synthetic data generation techniques to diversify the training data.[7] After augmenting your dataset, retrain your model and reassess its performance using the bias analysis approach. Report the results in a second Bias Analysis Table for the new model. Document any performance changes and the steps taken to mitigate bias in your report.

Ensure that all steps, findings, and changes in model performance after mitigation are thoroughly documented in your project report (see below). This detailed analysis is crucial for demonstrating your understanding of ethical AI development and the importance of creating fairer AI systems.

using k-fold in scikit-learn.

[7] If you are using synthetic data generation techniques, ensure that the generated data is diverse and representative of real-world variations. Avoid creating synthetic data that reinforces stereotypes or neglects minority groups, as this could lead to further biases in your model.

Report. Update your report, correcting any issues from the first two versions and adding news chapter for the bias detection & elimination, as well as new evaluation sections for the cross-validation. You complete report should now look like the following:

Title page: *As shown in Part I, but if any changes in your group happened during Part I–III, update this page accordingly (including what changed and during which part)*

Dataset: *As defined in Part I, including any changes you made for Parts II or II.*

- Describe the changes you made to your dataset for mitigating your system's bias in a new section. That is, detail any additions, removals, or modifications to the dataset specifically for addressing bias, separately from general changes made earlier in the project.

Data Cleaning: *As explained in Part I, including any changes you made for Part II or III.*

Labeling: *As explained in Part I, including any changes you made for Part II.*

- Describe how your dataset was labeled for the two chosen bias attributes.
- Describe how you verified the labels.

Dataset Visualization: *As explained in Part I.*

- Update your three visualizations from Part I for the final dataset used in the project.

CNN Architecture. *As explained in Part II.*

- If you made any changes to your architecture from Part II, describe the changes and the final version here.

Evaluation. *Report your results from Part II.*

- **Confusion matrix:** Add the confusion matrix for your final model from Part III.
- **K-fold cross-validation:**
  - Add the two tables showing the results from the 10-fold cross-validation (one table for your model from Part II, one for the final model), using the template shown in Table 1.
  - Discuss any significant observations or trends noted across the different folds, particularly in relation to the consistency of the model's performance.
  - *Only for the model from Part II:* Contrast the results obtained from the k-fold cross-validation with those from the original train/test split evaluation from Part II. Analyze any differences in performance metrics and discuss possible reasons for these discrepancies. This analysis should focus on understanding how the k-fold cross-validation results might differ from the initial evaluation due to the model's varying performance across different segments of the data.

Bias Analysis. Report the results from your bias analysis and the steps taken for mitigation:

- **Introduction:** Describe which bias attributes (e.g., age, gender, race) you analyzed and your approach for assessing bias in these categories.
- **Bias Detection Results:** Present the results of the bias analysis conducted on the Part II model in the Bias Analysis Table (Table 2). Analyze disparities in model performance across different demographic groups.
- **Bias Mitigation Steps:** Describe the steps taken for bias mitigation, including changes to your dataset and any retraining of your model.

- **Comparative Performance Analysis:** Present the results for the re-trained model from Part III in another Bias Analysis Table, using the template shown in Table 2. Contrast the performance metrics of your original model from Part II with the final, adjusted model and discuss the impact of your bias mitigation efforts.

Length: ca. 2 pages, plus tables

Reference Section: *As explained in Part I.*

- Make sure your reference section is still complete and up-to-date.

Deliverables. This is it: you submit your complete project, including all the work from Parts I–III. Like before, ensure you bundle all the necessary items specified below into a single `.zip` or `.tgz` archive for submission on Moodle:

Python Code. All Python scripts developed for this project:

- This encompasses scripts for data cleaning, data visualization, and dataset processing.
- Your PyTorch code for the CNNs, including the variants, code for evaluation as well as saving, loading, and testing the models.
- Note: Only pure Python code (`.py` files) will be accepted. Jupyter notebooks or other formats will not be considered.
- Your code should be well-commented and modular to facilitate easy understanding and evaluation.
- Ensure that your code is fully functional and runnable. If the markers encounter persistent errors or unresolvable issues, this may impact your grade.

Dataset. A file or document detailing the provenance of each dataset/image:

- For publicly available datasets, incorporate only a reference with the necessary details such as dataset name, source, and licensing type (i.e., do not try to submit your whole dataset content here).
- For custom or modified images, ensure you include them alongside any manually crafted metadata.
- Supply 10 representative images from each class within your archive and incorporate a direct link to the full dataset in your repository (e.g., on Github).

README. A comprehensive `readme.txt` or `readme.md` file:

- It must enumerate the contents and describe the purpose of each file in your submission.
- Clearly outline the steps to execute your code for (a) data cleaning and (b) data visualization.
- Describe the steps for running your code to train, evaluate, and apply the models.
- If your instructions are incomplete and your code cannot be run you might not receive marks for your work.

Report. Your finalized project report:

- Must be structured adhering to the guidelines provided earlier.
- **New:** Update your report with the new sections for Part III, as well as any updates for the other sections.

- Submit your report as a PDF file.

Originality Form. Include a **single** *Expectation of Originality* form:

- Available at https://www.concordia.ca/ginacody/students/academic-services/expectation-of-originality.html
- This form, attesting to the originality of your work, must be electronically signed by **all** team members.
- If the form is missing, your project will not be marked!
- If your group has not changed, you can submit the same form as for Parts I and II.

Submission Procedure: You must submit your code electronically on Moodle by the due date (late submission will incur a penalty, see Moodle for details).

Project Demo. We will schedule demo sessions for your project using the Moodle scheduler. The demos will be on campus and all team members must be present for the demo. Guidelines for preparing for the demo will also be posted on Moodle.

Other Guidelines. Please refer to the Project Part #1 document for the *Academic Integrity Guidelines for the A.I.ducation Analytics Project* as well as the *Project Contribution and Grading Policy.*

Project Grading. Please be aware that this assignment contributes significantly to your overall project grade, accounting for 45% of the total marks. In this final evaluation, we will not only assess the new elements added in Part III but also revisit your deliverables from Parts I and II. This comprehensive approach ensures a holistic assessment of your entire project. Therefore, it is crucial that you address and rectify any issues or shortcomings identified in the earlier parts of the project. Your final submission should reflect a complete, polished, and cohesive project that integrates all three parts effectively. By doing so, you demonstrate not only your technical skills and knowledge gained throughout the course but also your ability to respond to feedback and improve your work progressively.

Closing Words. As we close this chapter of your AI journey, take a moment to reflect on the strides you've made. Starting as explorers in the realm of machine learning, you've navigated through the complexities of deep learning and Convolutional Neural Networks (CNNs). You've grappled with data cleaning, labeling, and turned theoretical concepts like precision, recall, and k-fold cross-validation into practical tools in your arsenal. More than just a project, this has been a transformative learning experience, equipping you with skills directly transferable to both advanced academic pursuits and real-world industry applications. As you step forward, carry these insights and experiences with confidence, ready to make an impact in the dynamic world of AI. 🚀🧠🤖