Technical Audit of an Automated Decision System

Nisha Ramanna nlr10004@nyu.edu
Olivia Marcelin oam8118@nyu.edu

April 18, 2024

1 Background

In recent years, the intersection of data science and healthcare has emerged as a pivotal area with immense potential to revolutionize patient care, diagnosis, and treatment. With vast amounts of data now available, including electronic health records and medical imaging, automated decision systems (ADS) promise to provide valuable insights that can improve healthcare outcomes and reduce costs [5]. However, a critical challenge that must be addressed is the potential for bias in ADSs used in healthcare. Bias in AI can lead to disparities in care, misdiagnoses, and inequities in access to treatment [4].

One attempt at applying ADSs to healthcare is the OSIC Pulmonary Fibrosis Progression Kaggle competition. Participants worked to submit an ADS that predicted the severity of decline in lung function for patients diagnosed with Idiopathic Pulmonary Fibrosis (IPF) and other fibrotic lung diseases [1]. An ADS with such an objective has many stakeholders that must be accounted for. An ADS with such an objective has many stakeholders that must be accounted for, such as the physicians who use the decisions in their diagnosis and treatment as well as insurance companies that may use the decisions to determine a recipient's coverage. There are also entire healthcare institutions that may use an ADS like this to decide how to allocate their resources to their community of patients as a whole. The most important stakeholder in this scenario is the patient, who is directly impacted by the decisions made by the ADS as they guide how their ailments are viewed and treated.

While an ADS like the one described in the competition can vastly improve the way patients with fibrotic lung diseases are diagnosed and treated, there are many harms that can come from it as well. There could be lack of transparency and interpretability due to the complexity of the algorithm, and if physicians aren't able to explain the decisions from the algorithms, patients can grow to distrust the system and their physician for using it. There is also risk of overreliance on the ADS by healthcare providers. If physicians or insurance companies rely on these decisions too much without any room for human input,

they could overlook details the ADS doesn't account for leading to an incorrect course of action for patients. A major harm that can come from an ADS like this is bias in its decision making, which causes discrepancy in care across different groups of patient populations to no fault of their own.

In this paper, we seek to look into the potential of bias and perform a technical audit of the first place submission's ADS from the competition [2]. As the competition asks for an ADS to make predictions for every week in a patient, we will focus our audit on the final week's prediction of every patient instead of the entire dataset to specify our target variable. In our audit, we intend to use both sex and age as sensitive features in order to investigate any difference in treatment between male and female patients as well as between younger and older patients.

2 Input and Output

2.1 Data Collection & Selection

The data used by this ADS consists of a collection of information about several patients. Each patient is identified by a unique ID, labeled "Patient". A "baseline CT scan" of the patient's lungs, taken at week 0, is also included. Then, there are multiple measurements of a patient's "forced vital capacity" (FVC), or the volume of air exhaled by the patient, over the course of 1-2 years. This variable represents the patient's lung function. Each FVC measurement is accompanied by two variables. First is "Weeks", which marks the number of weeks before/after the baseline CT scan when this FVC measurement was taken. Second is "Percent", which represents the patient's FVC as a percentage of the expected FVC for a person with similar traits. Finally, the dataset includes the patient's "Age", "Sex", and their smoking history, represented by "SmokingStatus". This data was collected from 200 real patient's suffering from pulmonary fibrosis. The CT scans have been anonymized to preserve patient confidentiality. The FVC measurements were taken over the course of 1-2 years when patients came in to visit their doctor. Note that since different patients get check-ups at different frequencies, the relative timing of FVC measurements will vary from patient to patient.

2.2 Data Description

The feature "Patient" is a unique string which identifies a patient. The "baseline CT scan" is given as a collection of DICOM (Digital Imaging and Communications in Medicine) formatted images. A CT scan produces a 3D model. The 3D model is represented by a set of 2D slices, which are presented in DICOM format. The 2D slices are stored in a folder named after the patient's unique ID. The feature "FVC" is an integer. This feature has range 827 to 6399. A higher FVC indicates stronger lungs, while a lower FVC indicates weaker lungs. Below, I have included the distribution of this feature:

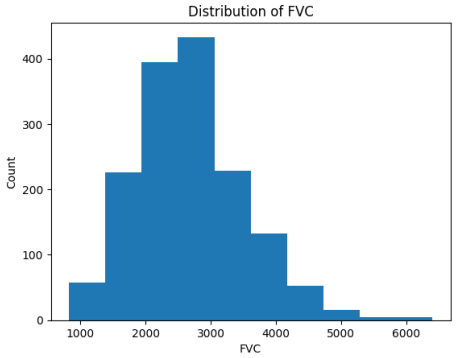

Figure 1: Distribution of Feature "FVC"

The feature "Weeks" is also an integer. Its values range from -5 to 133. A negative value for weeks indicates the FVC was taken before the baseline CT scan. A positive value X indicates the FVC was taken X weeks after the baseline CT scan. This is the distribution of the feature "weeks":

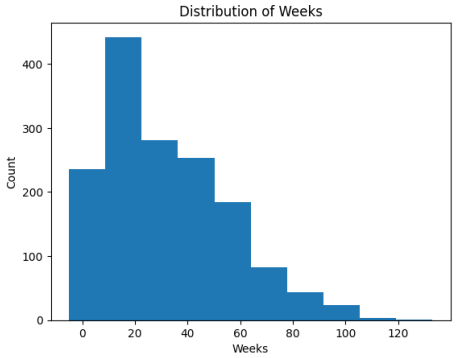

Figure 2: Distribution of Feature "Weeks"

The feature "Percent" is represented as a float. Its values range from 28.88 to 153.15. A percentage below 100 indicates this patient's lungs are functioning worse than a person with similar characteristics. A percent above 100 indicates the patient is doing better than a similar patient. Below is the distribution of "Percent":

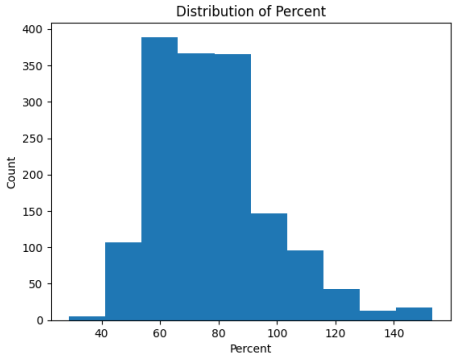

Figure 3: Distribution of Feature "Percent"

The feature "Age" is represented as an integer. It ranges from 49 to 88, indicating all patients in this study are middle to late age. Here is the feature's distribution:

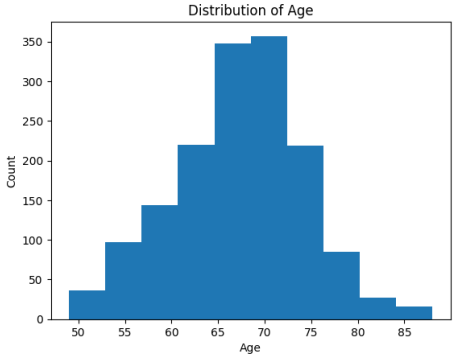

Figure 4: Distribution of Feature "Age"

"Sex" is a string, either Male or Female. I've provided the distribution for the feature below. Clearly, there are far less instances of female patients than male patients in this dataset. This could be explained by the fact that the majority of patients with pulmonary fibrosis are men [6].

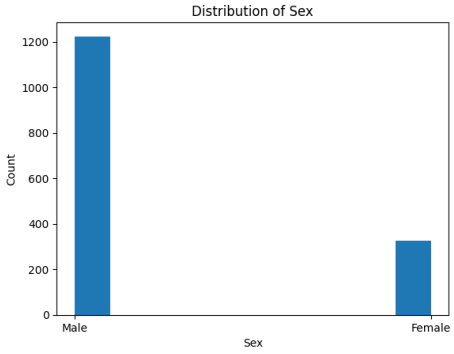

Figure 5: Distribution of Feature "Sex"

"SmokingStatus" is a string, either Ex-smoker, never smoked, or currently smokes. The distribution of this feature illustrates that majority of the patients in this dataset are ex-smokers, and very few currently smoke.

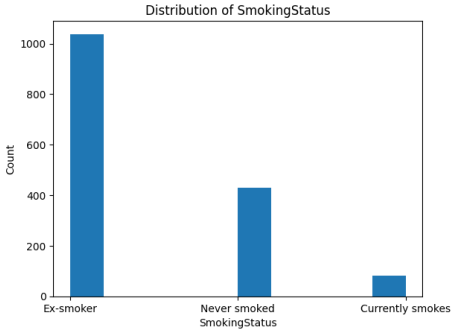

Figure 6: Distribution of Feature "SmokingStatus"

There are no missing values in the train or test set.

2.3 System Output

In the training dataset, the ADS is provided with all FVC measurements or a given patient. In the test dataset, it is only given the first FVC measurement. The output of the system is the predicted final 3 FVC measurements for each patient and a confidence value for the prediction. Specifically, three features are outputted. The first is "Patient_Week", which is a combination of the unique Patient ID and the week in which this predicted FVC would have been measured. The second feature is "FVC". The third feature is "Confidence", which is the confidence value for the given prediction.

3 Implementation and Validation

3.1 Data Cleaning and Pre-Processing

The first pre-processing step taken was converting tabular data, such as 'sex; and 'smoking status', into categorical binary variables. Additionally, the variable 'age' was normalized. The lung scans, given in DICOM format originally, were resized to a standard 512x512 size. Additionally, the pixel values of the images were normalized. Additionally, the ADS excluded some patient IDs from the training process, referring to them as 'BAD_ID'. While the reasoning behind the selection of these 'BAD_IDs' isn't explained, we can assume there were some issues with the data quality of these patients.

3.2 System Implementation

This ADS blends two models together—an EfficientNet B5, which is a kind of Convolutional Neural Network (CNN) and a Quantile Regression Dense Neural Network. The EfficientNet was trained on both the CT scans and the tabular data, while the Quantile Regression model was only trained on the tabular data. Both models were trained from scratch and then blended, creating diverse predictions and improving overall accuracy.

3.3 Validation

This ADS was validated using the same methods as the organizers of the Kaggle competitions. Organizers scored submissions based on their predictions of the FVC values for the last 3 weeks for each patient. Specifically, they used a log likelihood metric to assess the accuracy of predictions, which was mimicked in this ADS.

4 Outcomes

4.1 Chosen Metrics

Since the target variable 'FVC' we are studying is continuous and not categorical, many metrics that are typically used in medical contexts, such as False Negative Value, are not applicable. Therefore, we chose to focus on regression metrics that emphasize the difference in predicted value from true value. For performance analysis, we will look at the submission's Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The MAE measures the average magnitude of error in predictions while the RSME measures the square root of the average of squared differences between predictions and true values. With both of these metrics, a smaller values implies a more accurate model. For our fairness analysis of age and sex, we chose to use Mean Prediction Bias and Coefficient of Variation Ratio (CVR) for each group. Both a positive and negative Mean Prediction Bias value indicates the model has predicted poorly for

a subpopulation. Whether that value is positive or negative indicates if there is an over or underestimation, respectively, for the group. For this metric, a value closer to 0 indicates lack of favor to one group over another. The CVR is the ratio of the coefficient of variation (standard deviation / mean) of the predictions. For this metric, a value closer to 1 indicates similar uncertainty across groups. Using these metrics, let's take a closer look at the ADS.

4.2 Overall Performance

We calculated that the ADS had a MAE of 162.632 and an RMSE of 249.934. Considering that the value of 'FVC' is typically in between 2000 and 4000, we feel that the MAE is fairly acceptable. However, there certainly is room for improvement. The RMSE value being significantly larger than the MAE value indicates that there are several outliers for which the model cannot predict well for. In a healthcare setting, this is unacceptable. The model must be able to perform well on all patients.

4.3 Age Fairness Outcomes

To analyze the difference in outcome by age group, we will look at the 60-70 subpopulation and the 70+ subpopulation in the dataset. Note that the test set only includes patients 60+, so these are the only subpopulations we can analyze. The Mean Prediction Bias for 70+ is 129.428- this positive bias indicates an overestimation for the age group. This may lead to excessive concern and unnecessary interventions in older patients that don't need it, preventing care from going to patients that could actually benefit from it. Conversely, the Mean Prediction Bias for 60-70 is -87.702- this negative bias indicates an underestimation for this age group. This could lead to lack of concern or intervention for patients that do need it earlier on simply because of their age. Overall, the ADS predicts a more severe lung decline in older patients.

The CVR for 70+ is 1.0, which suggest the variability in predictions for this group is the reference level. This implies that there is a high degree of dispersion in predictions for the 70+ population. Therefore the ADS may have difficulty with accurate predictions within this age group. In comparison, the CVR for the 60-70 subpopulation is 0.160, which is quite low and implies much less relative variability in predictions for this age group. These CVR values emphasize a need for additional work to be done for the 70+ age group to improve their accuracy.

4.4 Sex Fairness Outcomes

As a disclaimer, we encountered a limitation in our audit pertaining to our study of outcomes between gender. The test data was severely limited and only included male patients. We attempted to expand by measuring the accuracy of our model on training and testing data. However, we encountered another barrier. The Kaggle competition provided a file listing which predictions it

wanted the ADS to make. For most predictions, the actual FVC value was withheld, preventing us from measuring fairness. The few predictions for which we did have the corresponding actual FVC value were only for male patients. We attempted to generate predictions for female patients for which we had the corresponding actual FVC value by modifying the file listing the desired predictions. However, this did not work for unknown reasons. Therefore, we were unable to properly measure fairness across genders.

We still generated the Mean Prediction Bias and CVR for the male subgroup. Our calculation for Mean Prediction Bias shows that the predicted FVC for males is 81.659 ml higher on average compared to the actual FVC. This conveys a systematic overestimation for males in the dataset. The CVR for males was 1.0 for males. This indicates a high degree of dispersion in predictions and therefore there is potential difficulty for accurate predictions among males. This data isn't too informative without the female metrics for comparison. The bias for overestimation in males does raise concern for the possibility of a similar or opposite bias among females patient.

4.5 Additional Performance Methods

To analyze the stability and robustness of this ADS, we compared its performance metrics when the model was trained using four different random seeds. Below is the result of this experiment: Note that there are minimal changes in

Table 1: Performance of ADS Across Random Seeds

Seed	Mean Absolute Error	Root Mean Squared Error
42	162.63175590837	249.93350409133015
1234	162.91310722672944	236.91335207789606
2020	156.7184370473623	238.02636004934135
5555	159.3333060487509	245.1238546166675

the performance of the ADS across all tested random seed values. This indicates the ADS is highly stable and robust as its predictions are consistent for all seed values tried.

5 Summary

In conclusion, we do not believe this data was appropriate for this ADS. Lack of representation of certain subpopulations resulted in different biases across demographic groups. Biases in this ADS could potentially prevent someone from receiving healthcare treatment they need or could result in precious medical resources being allocated poorly.

Based on the accuracy and performance metrics selected, we believe this model was not robust. While the ADS generated very similar MAE and RMSE for different random seeds, showing some amount of robustness, you also have to

consider the relatively large difference between the MAE and RMSE. The higher RMSE suggests that there are several instances where the ADS's prediction for FVC was significantly off from the actual FVC, potentially indicating a lack of robustness.

The ADS's MAE of 162 indicates that on average, the predicted FVC is 162 units off from the actual value. Given that the value of FVC can range from around 2000 to 4000, we can interpret that our model is fairly accurate. We selected MAE and RMSE as accuracy measures for our ADS because it predicted a continuous variable. Many traditional accuracy measures are better suited for categorical target variables. Additionally, we believe that both healthcare providers and patients will benefit from the selection of these accuracy measures. MAE indicates the average error of the ADS, showing whether or not it generally performs well. RMSE is especially beneficial for healthcare providers and patients because even if the average error is low, this doesn't mean the ADS isn't making some predictions with high error- RMSE will highlight these occurrences, though. This is especially important in a healthcare setting because even one poor prediction can have catastrophic results for a patient.

We also do not believe this ADS is fair. We weren't even able to properly examine the fairness of the ADS across gender and sex. There were only 5 test instances provided and all were for male patients who were 65+. Therefore, we couldn't compare the performance of the ADS on male vs. female patients or across the full age range of patients. However, using the limited test data provided, we did find that the ADS overestimated significantly for patients over 70 and underestimated for those aged 60-70. This suggests age-related discrepancies in the model's predictive accuracy. We selected mean prediction bias as a fairness metric as it would help ensure that no demographic group was being preferred or discriminated against by the ADS's predictions. This specifically helps women and younger patients, as historically pulmonary fibrosis has been diagnosed in majority elder men [3]. Therefore, these subpopulations may be more likely to be misdiagnosed or left undiagnosed. We selected CVR as our second fairness metric as it ensures consistency and reliability in the ADS's predictions for groups. This can be especially helpful for healthcare providers who want to ensure that the model's predictions are reliable.

Based on this assessment of the ADS, we would not be comfortable deploying it in the public sector due to its lack of robustness and its pattern of making predictions with high errors. Additionally, the fact that it has only been tested on such a limited patient demographic (males over 65), and that it still showed biases across the limited age range tested on, indicates that this ADS potentially has major fairness issues. To improve this ADS, we would begin by diversifying the data collection, specifically for the test set. we would vary the genders and ages of the patients to ensure that the model is trained and tested on a dataset more representative of the real world population. Additionally, we would choose to remove gender and age from the dataset when training the ADS. We acknowledge that historically, pulmonary fibrosis has been diagnosed in older men. This could indicate a correlation between age/gender and the disease. However, scientists have not yet identified the correlation or its underlying

causes. Therefore, it is still possible that a medical bias is influencing healthcare professionals, leading to hesitancy in diagnosing this disease in younger or female patients.

References

[1] David Elizabeth Estes Julia Elliott Justin Zita SimonWalsh Slepetys Will Cukierski Ahmed Shahin, Carmela Wegworth. Osic pulmonary fibrosis progression, 2020.

[2] artkulak. Osic pulmonary fibrosis progression 1st place solution. https://www.kaggle.com/competitions/osic-pulmonary-fibrosis-progression/discussion/189346, 2015.

[3] American Lung Association. Symptoms of pulmonary fibrosis, Nov 2022.

[4] Katherine J. Igoe. Algorithmic bias in health care exacerbates social inequities - how to prevent it, Mar 2021.

[5] Institute of Data. The impact of data science in healthcare: Revolutionizing the industry, Jan 2024.

[6] Lucile et al Sesé. Gender differences in idiopathic pulmonary fibrosis: Are men and women equal?, Aug 2021.