1 Part 1

1.1 Paper Selection

Positive Paper: *Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose CT scans for lung cancer screening: accuracy and prognostic value* [1].
Negative Paper: *Hospital quality classification based on quality indicator data during the COVID-19 pandemic* [2].

1.2 Positive Paper Evaluation

The positive paper presents a robust evaluation of an AI model for detecting lung nodules and coronary artery calcium. They use heart scaling to ensure consistent heart size in images, and the models are well-documented. Additionally, the model is validated by two expert radiologists, providing a high level of confidence in the results. Results are supported by confidence intervals and p-values, and the paper includes extensive statistical testing. They use multiple tests and test assumptions for these tests, enhancing the reliability of their findings. The correlation biplot is a useful visualization tool that helps to understand the relationship between model predictions and various attributes of the patients. The paper uses a train/test split for model evaluation.

1.3 Negative Paper Evaluation

The negative paper suffers from several critical issues that undermine its findings. The main problem is the bias introduced by inadequate data submissions from hospitals, leading to a non-representative dataset. The authors attempt to classify hospital quality during the COVID-19 pandemic. However, they do not account because these hospitals don't suddenly change their care quality. Then after the pandemic, they will find that all the hospitals are suddenly substantially better, which is not the case. The article uses data transformation techniques, such as averaging per year and collapsing months to a single value, thus removing all trends and seasonality in the data. In addition, the paper replaces NULL values with 0 this lowers the possible score and introduces further bias. The Neural Network (NN) and Decision Tree (DT) models are overfitted to the data, as they do not use early stopping or two-level cross-validation. The authors then conclude that the linear discriminant analysis is the best model, but this conclusion is flawed because the data is not representative of the population and the models are overfitted. When they report results, they do not provide confidence intervals or statistical significance for their results, without which it is impossible to assess the reliability of their findings. There is no standardization or normalization of the data, which helps some models to perform better. Finally, the paper lacks a sufficient description of the models used, making it difficult to replicate their findings.

1.4 Recommendations for Improvement of the Negative Paper

To improve the negative paper, the authors should address the following issues:

- Ensure that the dataset is representative of the population by addressing the bias introduced by inadequate data submissions from hospitals.
- Avoid replacing NULL values with 0, as this can introduce bias and lower the possible score.
- Use early stopping and two-level cross-validation to prevent overfitting of the models.
- Provide confidence intervals and statistical significance for the results to assess the reliability of the findings.

- Standardize and normalize the data to improve model performance.
- Include a more detailed description of the models used to allow for replication of the findings.
- Consider using a more appropriate problem formulation, such as predicting future hospital quality rather than classifying past data.

References

[1] Chamberlin, J., Kocher, M. R., Waltz, J., Snoddy, M., Stringer, N. F. C., Stephenson, J., Sahbaee, P., Sharma, P., Rapaka, S., Schoepf, U. J., Abadia, A. F., Sperl, J., Hoelzer, P., Mercer, M., Somayaji, N., Aquino, G., & Burt, J. R. (2021). *Automated detection of lung nodules and coronary artery calcium using artificial intelligence on low-dose CT scans for lung cancer screening: accuracy and prognostic value.* **BMC Medicine**.

[2] Nurhaida, I., Dhamanti, I., Ayumi, V., Yakub, F., & Tjahjono, B. (2024). *Hospital quality classification based on quality indicator data during the COVID-19 pandemic.* **International Journal of Electrical and Computer Engineering**

2 Part 2: Cross-Validation

Variable	Description
HR features	Mean, median, std, min, max, AUC of the HR signal
Round	Puzzle round (1–4)
Phase	Phase within each round (1–3)
Individual	Participant ID
Role	Puzzler or instructor
Frustration	Self-rated frustration (0–10)
Cohort	Cohort ID (D11, D12, D13)

Table 1: Key variables in the HR_data.csv subset.

2.1 Feature Selection and Preprocessing

As specified in the task description, only the hart rate data will be used for the classification task. This the following attributes will be used: HR Mean, HR Median, HR std, HR Min, HR Max, HR AUC. The data will be standardized and normalized to ensure that all features contribute equally to the model training.

The Frustration variable will be used as the target variable having values from zero to eight, however, the values 6 and above will be grouped into one class (high frustration) due to the small number of samples in these classes.

2.2 Machine Learning Models

Two machine learning models will be used for the classification task: Logistic Regression, XGBoost and a baseline classifier that predicts the most common label. Logistic Regression is chosen for its simplicity and interpretability, while XGBoost is selected for its ability to handle non-linear relationships and its strong performance in classification tasks. These models will output class probabilities for each frustration level, allowing for a multi-class classification task.

2.3 Cross-Validation Scheme

To ensure that the models generalize well to unseen data, a two-level cross-validation scheme will be employed. The inner fold will use stratified group 3-fold cross-validation, ensuring that each fold contains a representative sample of each frustration level. The outer fold will use Leave-One-Group-Out Cross-Validation, where each group is defined by the Individual ID. This approach ensures that no individual is present in both the training and test sets, preventing data leakage and ensuring that the model can generalize to new individuals.

In the inner fold the models will be tested on different hyperparameters to find the best performing model. The hyperparameters for the Logistic Regression model will be the regularization strength. For the XGBoost model, the hyperparameters will include the learning rate, maximum depth of trees, and n_ estimators (number of trees) and reg_lambda (L2 regularization term).

2.4 Statistical Comparison of Models

To compare the performance of the models across the folds, a repeated-measures ANOVA will be conducted. Repeated-measures ANOVA is appropriate here because it accounts for the non-independence of observations within each fold. The ANOVA can be seen in Table 2. Even

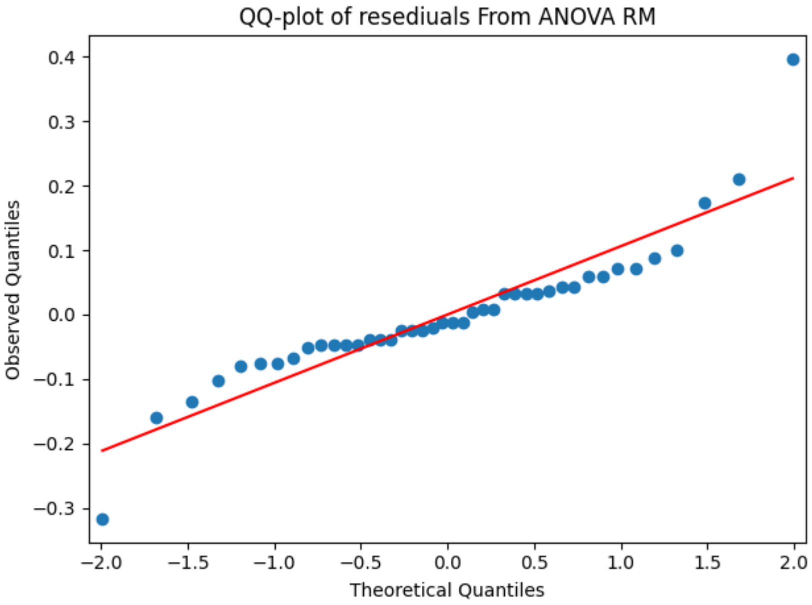

Figure 1:

though the p-value is not significant, it is close to the threshold of 0.05, indicating that there may be a chance of a difference in performance between the models.

Source	**F**	**Num DF**	**Den DF**	**Pr > F**
Model	3.3137	2	26	0.0522

Table 2: Results of the repeated-measures ANOVA comparing model performance across folds.

To check the normality of the residuals, QQ plots will be used. In Figure 1 the QQ plot shows that the residuals are somewhat normally distributed, which is a key assumption for the ANOVA test. However, there are some deviations from normality, particularly in the tails of the distribution.

To control the family-wise error rate when performing multiple comparisons, Holm's step-down correction will be applied. This method is a stepwise procedure that adjusts the p-values of multiple tests to reduce the likelihood of false positives.

Comparison	T	df	p_{unc}	p_{Holm}	BF_{10}	g
Baseline vs. LogReg	1.794	13	0.096	0.192	0.962	0.685
Baseline vs. XGB	2.261	13	0.042	0.125	1.816	0.674
LogReg vs. XGB	−0.354	13	0.729	0.729	0.285	−0.107

Table 3: Pairwise post-hoc comparisons (paired t-tests, Holm-corrected).

3 Conclusion

From the analysis Table 3 it can be seen that the XGBoost model outperforms the Logistic Regression model and the baseline classifier in terms of accuracy however, the difference is not statistically significant additionally the baseline classifier performs better than the other two models but not significantly.