CITY UNIVERSITY OF LONDON EST 1894

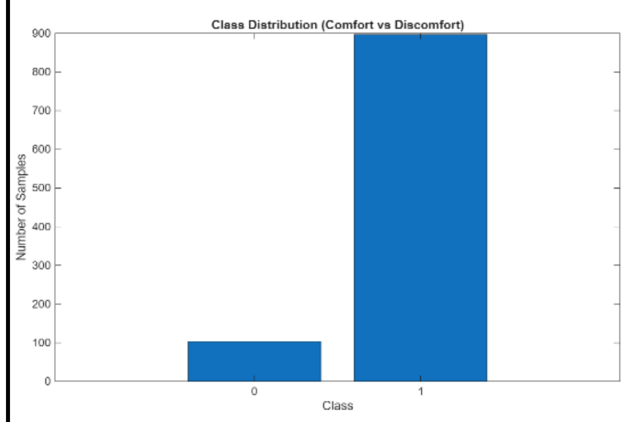
A Comparison of Logistic Regression and Random Forest for Patient Comfort Classification in Infusion Rooms

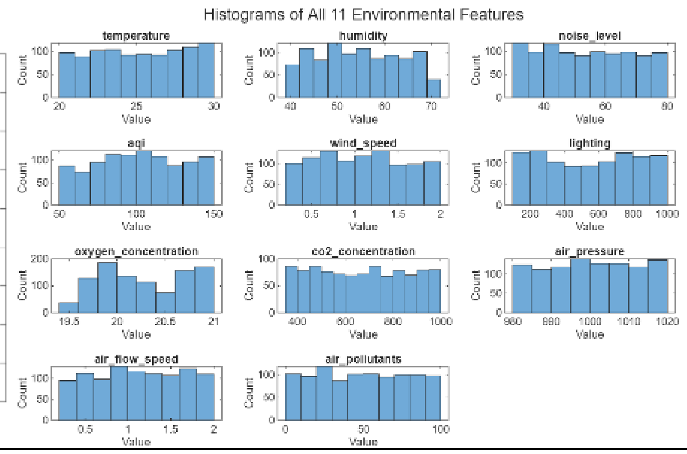
Description and motivation

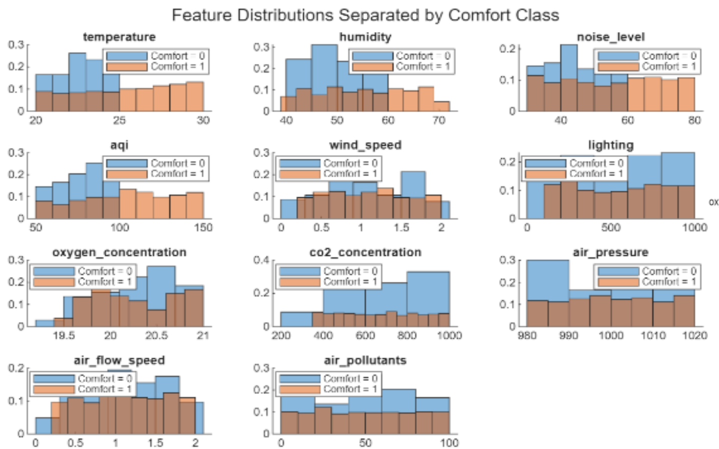
We aim to solve the binary classification problem of predicting excessive patient discomfort in hospital infusion rooms using only environment sensor measurements recorded during each infusion session. Our objective is to compare the performance of a linear model (Logistic Regression) and a non-linear ensemble method (Random Forest) in identifying how environmental conditions contribute to discomfort and determine which model provides more reliable predictions for real-world healthcare settings.

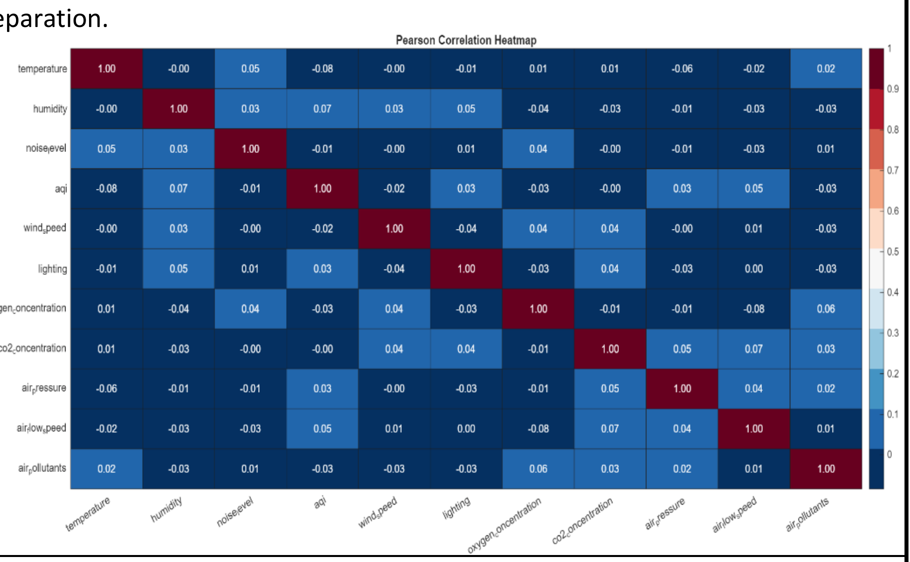
Exploratory Analysis

- Medical Environment Comfort Prediction Dataset from Kaggle, containing 1,000 hospital infusion-room sessions with 11 numeric environmental features and a binary target indicating excessive discomfort.
- The dataset has no missing values, so no imputation was required; only basic datatype checks and preparing the numeric feature matrix.
- Additional feature engineering was not performed. All predictors are numeric environmental features, including temperature, humidity, noise level, air quality index (AQI), C0₂ and O₂ concentration, air pressure, lighting intensity, wind speed, air flow speed and particulate pollutants. The models rely completely on ray sensor data.
- The target column has class imbalance problem, with discomfort cases making up almost 90% of "excessive discomfort" (class 1) and only 10% of "no excessive discomfort" (class 0), requiring the use of class weighting to prevent bias during model training.
- The histograms show mixed variabilities in features: temperature, humidity, noise, lighting and CO₂ vary widely, while oxygen concentration and air pressure remain tightly regulated.
- Discomfort cases tend to appear under higher temperature, humidity, noise and CO₂ conditions, whereas regulated features like O₂ and air pressure show almost no class separation.
- The Pearson correlation results show very weak linear relationships across all environmental features, showing no strong multicollinearity in the dataset.
- All features were standardized using z-score normalization to keep them on a similar scale, which is particularly important for Logistic Regression.









Logistic Regression

- Logistic Regression is a supervised classification algorithm that predicts the probability of a binary outcome — in our case, whether an infusion session is comfortable (0) or uncomfortable (1).
- The model assumes that there is a linear relationship between the environmental features (like temperature, humidity, CO₂, noise) and the likelihood of discomfort.
- It produces coefficients that show how much each feature pushes the probability up or down. For example, a positive coefficient for noise would mean that higher noise levels make discomfort more likely.
- The model separates the two classes using a linear decision boundary, and a prediction is made by checking whether the probability of discomfort goes above a chosen threshold (usually 0.5).
- Logistic Regression is easy to train and interpret, and parameters are typically found using methods like Maximum Likelihood Estimation. Regularization (L1/L2) can also be applied to prevent overfitting and improve generalisation.

Advantages

- Easy to interpret-coefficients show how each feature affects discomfort.

-Trains extremely quickly and is simple to implement, making it a strong baseline model.

-Works well when a linear separation between classes is reasonable.

-Less prone to overfitting when regularization is applied.

Disadvantages

- Assumes the relationship between environmental features and discomfort is linear, which is often not realistic.
- Cannot easily model interactions like "high CO₂ + high temperature = higher discomfort."
- Needs feature scaling (z-score) to perform well.
- Performance drops when data is complex or not linearly separable.

Random Forest

- Random Forest is a supervised ensemble classification model made up of many individual decision trees where each tree is trained on a slightly different subset of the data, and the final prediction (comfort vs discomfort) comes from the majority vote of all trees.
- Unlike Logistic Regression, it naturally learns non-linear patterns and feature interactions-useful for environmental sensor data where comfort depends on combined effects (e.g., temperature + humidity + AQI).
- Because each tree sees different parts of the data, the model captures different patterns, making it good at recognizing subtle relationships that a linear model would miss.
- Random Forest provides built-in feature importance that identifies which environmental variables have the strongest influence on predicting patient discomfort.
- It is robust to noise, outliers and weak correlations, which aligns well with this dataset where environmental variables show weak linear relationships but still influence comfort through complex interactions.

Advantages

- Captures non-linear patterns and feature interactions automatically.
- Typically achieves higher accuracy for complex environmental data.
- Provides useful feature importance insights.
- Robust to noise, outliers, and redundant features.

Disadvantages

-Not as easy to interpret- it's difficult to explain exactly how many trees arrive at the final prediction.
-More computationally expensive than linear models.
-Larger and harder to summarize compared to a simple linear model.
-Harder to visualize than a single decision boundary.

Hypothesis Statement:

- Since the goal is to predict whether patients experience excessive discomfort based on room conditions, we expect both models to perform meaningfully better than chance.
- As the dataset likely contains non-linear interactions between features that a linear model cannot fully capture, Random Forest is expected to outperform Logistic Regression.
- LR is expected to offer clear interpretability and faster training, while RF is likely to provide stronger predictive performance, higher accuracy and higher recall for discomfort.

Experimental results, parameter choices and feature selection:

Logistic Regression:

- Used as a baseline linear classifier for modelling.
- Applying frequency-based class weighting reduced bias toward the majority class and improved recall and F1-score for the minority (non-discomfort-Class 0) class, while maintaining high overall accuracy.
- The Ridge-regularized Logistic Regression was tuned using a grid search over λ ∈ {1e-5 … 10}, with F1-score (class 1: over-discomfort) used as the optimization metric due to class imbalance.

Choice of parameters

- The regularization parameter λ = 0.00001 (1e-5) achieved a CV F1 ≈ 0.912, with very similar performance on the test set.
- Larger λ values reduced model flexibility and slightly degraded recall, while smaller values did not improve performance further.

Random Forest:

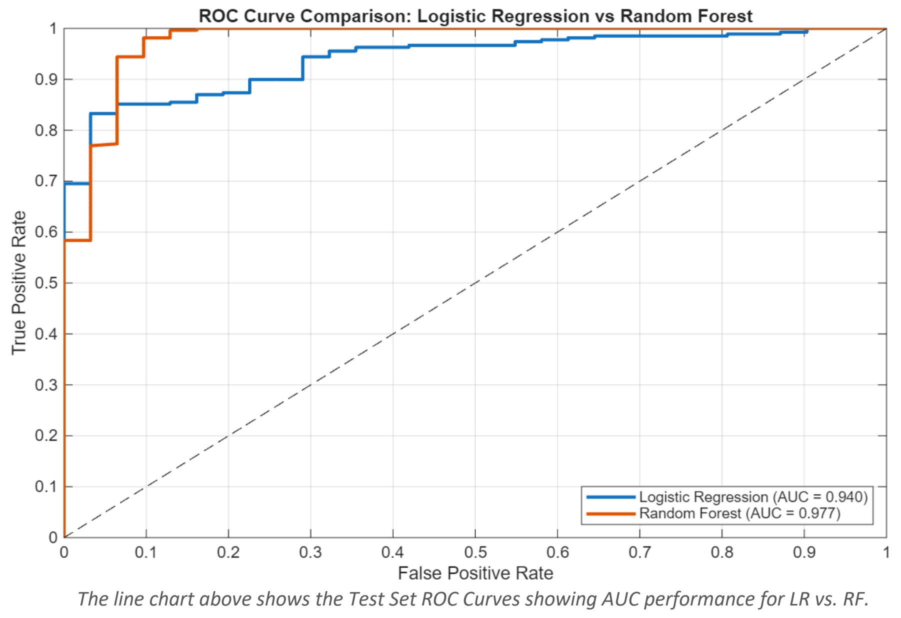
- Hyperparameter tuning using only 10-fold cross-validation showed that increasing the number of trees improved stability up to a point, after which gains saturated.
- Frequency-based class weighting reduced bias toward the dominant discomfort class and improved recall and F1-score, ensuring that discomfort events were consistently identified across validation folds.
- The best model achieved near-perfect cross-validated F1 (≈ 0.995) and perfect recall on the unseen test set.

Choice of parameters

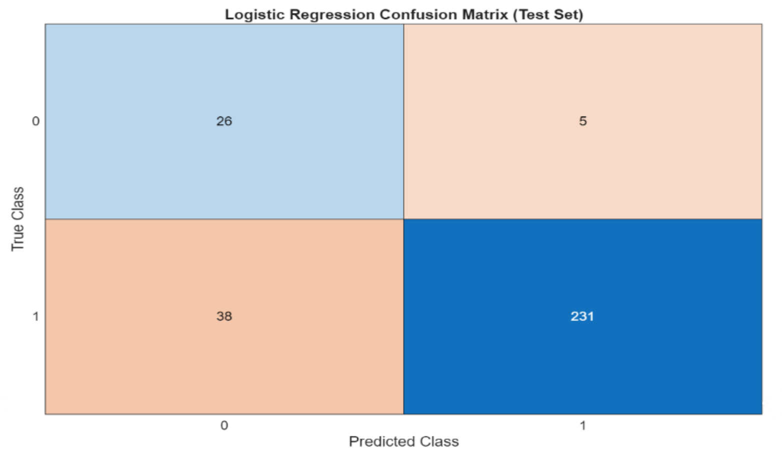
- The optimal configuration used 20 trees with a minimum leaf size of 10.
- Increasing the number of trees beyond 20 did not improve cross-validated performance, while smaller forests showed higher variance.
- A larger minimum leaf size reduced overfitting and improved generalisation on the unseen test set.

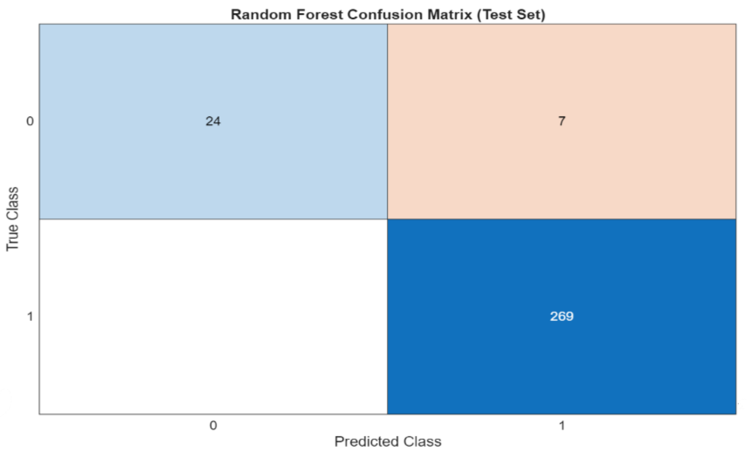
	Logistic Regression	Random Forest
Precision (Test)	0.979	0.975
Recall (Test)	0.859	1.000
F1-score (Test)	0.915	0.987
Avg Train AUC	0.939	0.998
Test AUC	0.940	0.977
Avg Train Error	0.146	0.006
Test Error	0.143	0.023
Train Time	0.014	0.043
Prediction time	0.016	0.015

The table above shows the Training and Testing Performance Metrics for LR vs RF.



The line chart above shows the Test Set ROC Curves showing AUC performance for LR vs. RF.





Methodology:

1. Split data into a 70: 30 split for train and test data, with the test set kept completely unseen.
2. 10 - fold cross validation was applied on the training dataset only to estimate model generalization and to compute average AUC, precision, recall, F1-score, error and other performance metrics.
3. Inverse-frequency class weights were used during training to reduce bias caused by severe class imbalance(Class 1).
4. Grid search combined with 10-fold cross-validation to tune model hyperparameters: the regularisation parameter (λ) for Logistic Regression, and the number of trees and minimum leaf size for Random Forest.
5. Best hyperparameter settings selected using 10-fold cross-validated F1-score
6. Retrained on the full training set and evaluated on unseen test data using precision, recall, F1-score, AUC, confusion matrices, and training time.

Analysis and Evaluation of results:

- The average training AUC of the Logistic Regression (0.939) was consistent with the expected range for linear models on weakly correlated environmental data, but lower than the Random Forest (0.998). This difference aligns with findings by Couronné et al. (2018), who showed that "Random Forest often outperforms Logistic Regression when relationships are nonlinear or involve interactions among predictors". The moderate AUC gap here indicates true nonlinear dependencies between environmental factors influencing discomfort.
- Class imbalance (≈ 90 % discomfort vs. 10 % comfort) initially biased the models toward always predicting "discomfort." To address this, class weighting was applied instead of resampling, ensuring the minority "comfort" class was not ignored. This approach follows van den Goorbergh et al. (2022), who cautioned that "resampling may distort model calibration in medical predictive contexts". The correction slightly reduced raw accuracy but improved model fairness and calibration across classes.
- After applying balanced training and standardization, the LR model achieved Precision = 0.979, Recall = 0.859, F1 = 0.915, and an AUC of 0.940. While highly precise, it failed to identify 38 true discomfort cases, highlighting sensitivity limitations of linear boundaries. Its low train–test AUC difference (Δ = 0.001) confirmed excellent stability with minimal overfitting.
- The RF model achieved near-perfect performance on the unseen test set (Precision = 0.975, Recall = 1.000, F1 = 0.987, Test AUC = 0.977). The confusion matrix revealed only seven false positives and no false negatives, meaning all patient discomfort cases were correctly detected. In clinical settings where missed alarms are unacceptable, this full recall is ideal despite a minor increase in false positives.
- ROC comparison demonstrated that the Random Forest curve closely hugged the top-left corner, confirming superior discrimination between "comfort" and "discomfort" sessions. This agrees with Muchlinski et al. (2016) and Daghistani & Alshammari (2020), both of whom found ensemble methods consistently achieving higher AUC and F1 scores than Logistic Regression in healthcare settings.
- Training error (RF = 0.0057 | LR = 0.146) and test error (RF = 0.023 | LR = 0.143) further illustrate RF's strong generalisation and resilience to variance. Both models trained extremely quickly (≈ 0.02 s), indicating suitability for real-time clinical integration. RF's slightly longer prediction time (0.0108 s vs 0.0008 s for LR) is negligible in practice.
- Cross validation vs test performance demonstrates good generalisation for both models. LR F1-score remains stable from 0.912 (10-fold CV) to 0.915 (Test), while RF changes only slightly from 0.995 to 0.987. The small gaps between training/CV and test metrics suggest that neither model is severely overfitting; RF's high training AUC (0.998) combined with strong test AUC (0.977) is an example of "benign overfitting", where the model fits complex patterns without losing test performance.
- From a clinical perspective, Recall is more important than overall accuracy, because missing a true discomfort episode may leave a patient untreated. LR's Recall of 0.859 means roughly 1 in 7 discomfort sessions would be missed, whereas RF's Recall of 1.000 ensures that no excessive-discomfort events go undetected on our test data.
- In terms of model bias–variance trade-off, Logistic Regression exhibited lower variance but higher bias, underfitting the complex relationships among environmental variables. Random Forest, in contrast, captured these interactions naturally, attaining low bias and acceptable variance.

Lessons Learned

- Severe class imbalance taught me to focus on F1-score, recall, confusion matrices, not only accuracy, ensuring fairer performance, since every best model is not selected based on accuracy. We must see every factor that affects the model.
- Predicting models by training only once or twice on the whole dataset is not at all sufficient. We must separate the original dataset into train and test set(unseen), operate on different MATLAB files for each set.
- We must use correct and same resampling techniques while comparing two ML models.
- We must calculate average AUC, test error, time etc. for both training and testing dataset, not taking only single set.

Future Work:

- Apply SMOTE to better balance discomfort class and improve minority class recall.
- Try other classifiers (e.g., SVM, Gradient Boosting) and use feature selection to improve performance and interpretability.

References:

(1)Couronné, R., Probst, P. & Boulesteix, AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 19, 270 (2018). https://doi.org/10.1186/s12859-018-2264-5

(2) van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022 Aug 16;29(9):1525-1534. doi: 10.1093/jamia/ocac093. PMID: 35686364; PMCID: PMC9382395.

(3) Muchlinski, David & Siroky, David & He, Jingrui & Kocher, Matthew. (2015). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. Political Analysis. 24. 1-17. 10.1093/pan/mpv024.

(4) Tahani Daghistani and Riyad Alshammari, "Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes," Journal of Advances in Information Technology, Vol. 11, No. 2, pp. 78-83, May 2020. doi: 10.12720/jait.11.2.78-83

(5) Zhang, C., Liu, L. Machine learning prediction model for medical environment comfort based on SHAP and LIME interpretability analysis. *Sci Rep* 15, 39269 (2025). https://doi.org/10.1038/s41598-025-22972-6

(6) Levy, J.J., O'Malley, A.J. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC Med Res Methodol* 20, 171 (2020). https://doi.org/10.1186/s12874-020-01046-3