

# Supplementary Material, Appendix A

Table 1. Random forest (binary) hyperparameter configuration.

model	n_estimators	max_depth	min_samples split	max_features	optimised threshold	min_samples_leaf
Random Forest (Binary)	100	10	10	sqrt	0.7	1

Table 2. XGBoost hyperparameter configuration.

model	tree_method	num_boost_round	learning rate	max_depth	reg_alpha	min_child_weight
XGBoost	hist	100	0.01	6	L1	1

Table 3. Neural network hyperparameter configuration.

model	learning rate	iterations	activation function	Hidden_layer	neurons	Bias	dropout
Neural Network	0.1	1000	Sigmoid	2	100	Self_bias	0.3

Table 4. Technical glossary terms.

Term	Meaning
Multicollinearity	Describes when two or more predictor variables (e.g., $H_s$ , $U_{10}$ ) are highly correlated. Multicollinearity is the linear relation among two or more variables (Alin 2010).
Variance Inflation Factor (VIF)	Measures the degree of multicollinearity, with higher VIF (>5) indicating greater correlation (Alin 2010).
Homoscedasticity	Means the variance of errors is consistent across all levels of an independent variable (Yang et al 2019).
Akaike Information Criterion (AIC) & Bayesian Information Criterion (BIC)	AIC focuses on minimising information loss, adding a penalty for additional parameters, and essentially favouring a model with fewer variables. BIC penalises the model complexity more stringently for models exhibiting more variables (Jayaweera and Trindade 2024).
Random Forest Variable Importance Metric	Examines the feature importance by comparing and examining how each feature (e.g., $H_s$ , $U_{10}$ , $U_{10Dir}$ ) reduces the error (Gini Impurity for binary) (Gregorutti et al 2016).
Shapiro-Wilk Test	Assessed the normality of the different predictor variables. A lower p-value (<0.05) indicates the dataset is not normally distributed (Estrada and Cosmes 2019).

## References

- Alin, A. (2010). Multicollinearity. *Wiley interdisciplinary reviews: computational statistics*, 2(3), 370-374. <https://doi.org/10.1002/wics.84>
- Yang, K., Tu, J., Chen, T. (2019). Homoscedasticity: An overlooked critical assumption for linear regression. *General psychiatry*, 32(5), 1-21. <https://doi.org/10.1136/gpsych-2019-100148>
- Jayaweera, I. N., Trindade, A. A. (2024). How Certain are You of Your Minimum AIC or BIC Values?. *Sankhya A*, 1-40. <https://doi/10.1007/s13171-024-00344-y>
- Gregorutti, B., Michel, B., Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(1), 659-678. <https://doi/10.1007/s11222-016-9646-1>
- González-Estrada, E., Cosmes, W. (2019). Shapiro–Wilk test for skew normal distributions based on data transformations. *Journal of Statistical Computation and Simulation*, 89(17), 3258-3272. <https://doi.org/10.1080/00949655.2019.1658763>