Appendix to Neonatal Sepsis Detection Using Decision Tree Ensemble Methods: Random Forest and XGBoost

Marwan Al-Bardaji and Nahir Danho

APPENDIX A - UNMODIFIED DATA

TABLE I
CAPTION

	Count	Mean	Std	Min	25%	50%	75%	Max
feats btb mean	134668	-28439.0	45112.0	-99999.0	-99999.0	-0.00010418	0.00037728	0.023877
feats rf mean	134668	-28407.0	45133.0	-99999.0	-99999.0	44.812	54.54	97.018
feats spo2 mean	134668	-28373.0	45154.0	-99999.0	-99999.0	91.844	94.871	100.0
feats btb std	134668	-28439.0	45112.0	-99999.0	-99999.0	0.012186	0.02	0.24101
feats rf std	134668	-28433.0	45116.0	-99999.0	-99999.0	11.001	14.686	38.089
feats spo2 std	134668	-28437.0	45114.0	-99999.0	-99999.0	2.739	4.5539	27.82
feats btb max	134668	-28439.0	45112.0	-99999.0	-99999.0	0.053329	0.13255	1.6416
feats rf max	134668	-28379.0	45150.0	-99999.0	-99999.0	81.389	96.0	163.22
feats spo2 max	134668	-28368.0	45157.0	-99999.0	-99999.0	100.0	100.0	100.0
feats btb min	134668	-28439.0	45112.0	-99999.0	-99999.0	-0.047756	-0.031711	-0.0013959
feats rf min	134668	-28429.0	45119.0	-99999.0	-99999.0	18.111	22.778	55.333
feats spo2 min	134668	-28386.0	45146.0	-99999.0	-99999.0	70.722	81.878	100.0
feats btb skew	134668	-28438.0	45113.0	-99999.0	-99999.0	0.13312	2.4165	14.834
feats rf skew	134668	-28443.0	45110.0	-99999.0	-99999.0	-0.092083	0.33747	12.711
feats spo2 skew	134668	-28441.0	45112.0	-99999.0	-99999.0	-1.608	-0.81382	2.5072
feats btb kurtosis	134668	-28425.0	45122.0	-99999.0	-99999.0	1.0974	14.668	239.58
feats rf kurtosis	134668	-28443.0	45110.0	-99999.0	-99999.0	-0.64149	-0.084224	341.28
feats spo2 kurtosis	134668	-28437.0	45114.0	-99999.0	-99999.0	0.52569	2.8307	591.91
feats btb sampAs	134668	-28437.0	45114.0	-99999.0	-99999.0	0.95311	2.5452	275.82
feats btb sampEn	134668	-28439.0	45113.0	-99999.0	-99999.0	0.34335	0.48042	1.0138
feats cirk vikt	128073	1.2967	0.49423	0.4958	0.92919	1.2139	1.5765	5.253
feats bw	134668	844.64	257.06	400.0	636.0	802.0	1013.0	1498.0
feats sex	134668	1.5655	0.49569	1.0	1.0	2.0	2.0	2.0
feats pnage days	134668	31.543	23.368	-0.85139	12.372	26.915	47.071	251.16

APPENDIX B - PREPROCESSING

A. Data description after removal of erroneous data points

TABLE II
CAPTION

	Count	Mean	Std	Min	25%	50%	75%	Max
feats btb mean	95849	0.00020301	0.0010439	-0.028372	-0.00018921	0.0001638	0.00056228	0.023877
feats rf mean	95849	50.194	11.525	8.0081	43.223	50.542	57.607	97.018
feats spo2 mean	95849	93.491	3.4743	35.618	91.363	93.521	95.971	100.0
feats btb std	95849	0.020122	0.014444	0.00055649	0.011286	0.016114	0.023798	0.24101
feats rf std	95849	13.479	4.3061	0.0	10.486	13.064	16.066	38.089
feats spo2 std	95849	4.3072	2.7283	0.0	2.5023	3.679	5.358	27.82
feats btb max	95849	0.14566	0.14853	0.001649	0.047296	0.085492	0.18946	1.6416
feats rf max	95849	89.722	16.519	15.25	79.111	89.889	100.56	163.22
feats spo2 max	95849	99.681	0.91062	79.078	99.944	100.0	100.0	100.0
feats btb min	95849	-0.042934	0.026538	-0.80757	-0.051221	-0.037783	-0.027746	-0.0013959
feats rf min	95849	20.076	6.851	0.39394	17.111	21.0	24.139	55.333
feats spo2 min	95849	74.461	14.721	0.033333	68.122	77.822	84.444	100.0
feats btb skew	95849	2.1813	2.9255	-5.2424	-0.0014337	0.92294	3.8908	14.834
feats rf skew	95849	0.15149	0.62641	-17.822	-0.16126	0.15826	0.47455	12.711
feats spo2 skew	95849	-1.3403	1.1861	-23.228	-1.7823	-1.1137	-0.61499	2.5072
feats btb kurtosis	95849	20.382	31.826	-1.3333	0.74752	4.5817	27.817	239.58
feats rf kurtosis	95849	0.056369	4.3335	-3.0	-0.70815	-0.34926	0.16646	341.28
feats spo2 kurtosis	95849	4.0127	10.32	-3.0	0.31891	1.5552	4.227	591.91
feats btb sampAs	95849	3.6363	5.7169	0.065211	0.84322	1.5616	3.8936	275.82
feats btb sampEn	95849	0.41042	0.15026	0.012408	0.3141	0.42854	0.5171	1.0138
feats cirk vikt	91550	1.2757	0.47036	0.49672	0.92301	1.194	1.549	3.724
feats bw	95849	839.31	258.12	400.0	636.0	789.0	1011.0	1498.0
feats sex	95849	1.5601	0.49637	1.0	1.0	2.0	2.0	2.0
feats pnage days	95849	31.256	22.156	0.025035	12.83	26.993	46.612	134.26

B. Data after standard scaling

TABLE III
CAPTION

	Count	Mean	Std	Min	25%	50%	75%	Max
feats btb mean	95849	6.428e-17	1.0	-27.373	-0.37572	-0.037554	0.34416	22.678
feats rf mean	95849	-7.008e-16	1.0	-3.6604	-0.60485	0.030175	0.64319	4.0628
feats spo2 mean	95849	-4.6673e-16	1.0	-16.658	-0.61254	0.0086296	0.71387	1.8734
feats btb std	95849	-1.5383e-16	1.0	-1.3546	-0.61175	-0.27749	0.25455	15.293
feats rf std	95849	-1.7911e-17	1.0	-3.1303	-0.69519	-0.096462	0.60055	5.7151
feats spo2 std	95849	3.4707e-16	1.0	-1.5787	-0.66155	-0.23023	0.38516	8.6184
feats btb max	95849	1.2715e-16	1.0	-0.96959	-0.66226	-0.4051	0.29487	10.072
feats rf max	95849	1.027e-17	1.0	-4.5083	-0.64232	0.010129	0.65585	4.4495
feats spo2 max	95849	4.7662e-15	1.0	-22.625	0.2895	0.35051	0.35051	0.35051
feats btb min	95849	3.9339e-17	1.0	-28.813	-0.31225	0.19411	0.57233	1.5652
feats rf min	95849	-6.2789e-17	1.0	-2.8729	-0.43279	0.13485	0.59302	5.1463
feats spo2 min	95849	8.8555e-16	1.0	-5.0559	-0.4306	0.22832	0.67817	1.7349
feats btb skew	95849	1.1817e-17	1.0	-2.5376	-0.7461	-0.43013	0.58436	4.3249
feats rf skew	95849	4.0558e-17	1.0	-28.693	-0.49929	0.010799	0.51573	20.05
feats spo2 skew	95849	-1.1326e-16	1.0	-18.454	-0.3727	0.19105	0.61148	3.2439
feats btb kurtosis	95849	-7.9114e-17	1.0	-0.68232	-0.61694	-0.49647	0.23361	6.8872
feats rf kurtosis	95849	1.3601e-16	1.0	-0.7053	-0.17642	-0.093604	0.025405	78.743
feats spo2 kurtosis	95849	5.8921e-17	1.0	-0.67956	-0.35794	-0.23814	0.020765	56.969
feats btb sampAs	95849	-9.2738e-17	1.0	-0.62466	-0.48857	-0.3629	0.045008	47.61
feats btb sampEn	95849	-6.2922e-16	1.0	-2.6488	-0.64106	0.12056	0.70993	4.0156
feats cirk vikt	91550	-4.6828e-16	1.0	-1.6562	-0.74984	-0.1737	0.58105	5.2052
feats bw	95849	2.4497e-15	1.0	-1.702	-0.78764	-0.19489	0.66518	2.5519
feats sex	95849	-3.5855e-14	1.0	-1.1285	-1.1285	0.88617	0.88617	0.88617
feats pnage days	95849	3.841e-15	1.0	-1.4096	-0.83166	-0.1924	0.69308	4.6488

C. Data after filling missing data points

This corresponds to the final data used in the machine learning models.

TABLE IV
CAPTION

	Count	Mean	Std	Min	25%	50%	75%	Max
feats btb mean	95849	6.428e-17	1.0	-27.373	-0.37572	-0.037554	0.34416	22.678
feats rf mean	95849	-7.008e-16	1.0	-3.6604	-0.60485	0.030175	0.64319	4.0628
feats spo2 mean	95849	-4.6673e-16	1.0	-16.658	-0.61254	0.0086296	0.71387	1.8734
feats btb std	95849	-1.5383e-16	1.0	-1.3546	-0.61175	-0.27749	0.25455	15.293
feats rf std	95849	-1.7911e-17	1.0	-3.1303	-0.69519	-0.096462	0.60055	5.7151
feats spo2 std	95849	3.4707e-16	1.0	-1.5787	-0.66155	-0.23023	0.38516	8.6184
feats btb max	95849	1.2715e-16	1.0	-0.96959	-0.66226	-0.4051	0.29487	10.072
feats rf max	95849	1.027e-17	1.0	-4.5083	-0.64232	0.010129	0.65585	4.4495
feats spo2 max	95849	4.7662e-15	1.0	-22.625	0.2895	0.35051	0.35051	0.35051
feats btb min	95849	3.9339e-17	1.0	-28.813	-0.31225	0.19411	0.57233	1.5652
feats rf min	95849	-6.2789e-17	1.0	-2.8729	-0.43279	0.13485	0.59302	5.1463
feats spo2 min	95849	8.8555e-16	1.0	-5.0559	-0.4306	0.22832	0.67817	1.7349
feats btb skew	95849	1.1817e-17	1.0	-2.5376	-0.7461	-0.43013	0.58436	4.3249
feats rf skew	95849	4.0558e-17	1.0	-28.693	-0.49929	0.010799	0.51573	20.05
feats spo2 skew	95849	-1.1326e-16	1.0	-18.454	-0.3727	0.19105	0.61148	3.2439
feats btb kurtosis	95849	-7.9114e-17	1.0	-0.68232	-0.61694	-0.49647	0.23361	6.8872
feats rf kurtosis	95849	1.3601e-16	1.0	-0.7053	-0.17642	-0.093604	0.025405	78.743
feats spo2 kurtosis	95849	5.8921e-17	1.0	-0.67956	-0.35794	-0.23814	0.020765	56.969
feats btb sampAs	95849	-9.2738e-17	1.0	-0.62466	-0.48857	-0.3629	0.045008	47.61
feats btb sampEn	95849	-6.2922e-16	1.0	-2.6488	-0.64106	0.12056	0.70993	4.0156
feats cirk vikt	95849	0.0031612	1.0	-1.6562	-0.74468	-0.17128	0.57615	5.2052
feats bw	95849	2.4497e-15	1.0	-1.702	-0.78764	-0.19489	0.66518	2.5519
feats sex	95849	-3.5855e-14	1.0	-1.1285	-1.1285	0.88617	0.88617	0.88617
feats pnage days	95849	3.841e-15	1.0	-1.4096	-0.83166	-0.1924	0.69308	4.6488

APPENDIX C - KNN IMPUTER

KNN Imputing is an algorithm used to fill in missing values. The missing value will be predicted in reference to the mean of the neighbors.

APPENDIX D - HYPERPARAMTER TUNING PROCEUDURE FOR RANDOM FOREST

1) Hyperparameter spaces notation: The following notation is used to define the entire hyperparameter space H of q different parameters:

$$H = P_1 \times \cdots \times P_q = \bigotimes_i P_i \tag{1}$$

Each hyperparameter space P_i corresponds to hyperparameter i and consists of the set of possible hyperparameter values that can be chosen. The following notation for the sets is used:

$$\begin{gathered}[a, b, s] = \text{All real values from a to b (including b)} \\ \text{with step sizes of } s \text{ starting from } a\end{gathered} \tag{2}$$

The size or cardinality of the hyperparameter space is given by $|H|$.

2) Hyperparameter tuning procedure: Random Forest: The hyperparameter tuning procedure started with the following hyperparameter space using the notation in equations (1) and (2)

- The hyperparameter search procedure started with a cross-validated random halving search

$$\begin{aligned} H_{\text{initial}} &= \bigotimes_{i \in \text{Random forest hyperparameters}} P_i \\ \text{where} \quad P_{n \text{ estimators}} &= [10, 1000, 10], \\ P_{\text{maximum depth}} = [1, 30, 1], &P_{\text{maximum features}} = [1, 30, 1], \\ P_{\text{minimum samples split}} = [1, 30, 1], &P_{\text{minimum samples leaf}} = [1, 30, 1] \end{aligned} \tag{3}$$

  After the cross-validated random halving search, the following hyperparameters achieved the best performance.

$$\begin{gathered} n \text{ estimators} = 160, \text{maximum depth} = 4, \text{maximum features} = 2, \\ \text{minimum samples split} = 3, \text{minimum samples leaf} = 29 \end{gathered} \tag{4}$$

  After the cross-validated random halving search, the model performed with a mean ROC AUC of 0.84 with a standard error of 0.11

- First cross-validated grid search iteration - In the next cross-validated grid search iteration the model was tuned for the n estimator, maximum features, and maximum depth hyperparameters. The hyperparameter space consisted of:

$$\begin{aligned} H_{\text{first grid search}} &= \bigotimes_{i \in \text{Random forest hyperparameters}} P_i \\ \text{where} \quad P_{n \text{ estimators}} &= [140, 200, 10], \\ P_{\text{maximum depth}} = [1, 8, 1], &P_{\text{maximum features}} = [1, 5, 1], \\ P_{\text{minimum samples split}} = \{3\}, &P_{\text{minimum samples leaf}} = \{29\} \end{aligned} \tag{5}$$

  After the first iteration of the cross-validated grid search, the following hyperparameters achieved the best performance.

$$\begin{gathered} n \text{ estimators} = 150, \text{maximum depth} = 1, \text{maximum features} = 2, \\ \text{minimum samples split} = 3, \text{minimum samples leaf} = 29 \end{gathered} \tag{6}$$

  After the first iteration of the cross-validated grid search, the model performed with a mean ROC AUC of 0.82 with a standard error of 0.12

- Second iteration of cross-validated grid search - In the next cross-validated grid search iteration, the model was tuned for the minimum samples split and the minimum samples leaf. The hyperparameter space consisted of:

$$\begin{aligned} H_{\text{second grid search}} &= \bigotimes_{i \in \text{Random forest hyperparameters}} P_i \\ \text{where} \quad P_{n \text{ estimators}} &= \{150\}, \\ P_{\text{maximum depth}} = \{1\}, &P_{\text{maximum features}} = \{2\}, \\ P_{\text{minimum samples split}} = [2, 5, 1], &P_{\text{minimum samples leaf}} = [20, 30, 1] \end{aligned} \tag{7}$$

  After the second iteration of the cross-validated grid search, the following hyperparameters achieved the best performance.

$$\begin{gathered} n \text{ estimators} = 160, \text{maximum depth} = 4, \text{maximum features} = 2, \\ \text{minimum samples split} = 3, \text{minimum samples leaf} = 23 \end{gathered} \tag{8}$$

  After the second iteration of the cross-validated grid search, the model performed with a mean ROC AUC of 0.84 with a standard error of 0.11.

3) Hyperparameter Tuning Procedure for XGBoost: The hyperparameter tuning procedure started with the following hyperparameter space using the notation in equations (1) and (2).

- The hyperparameter search procedure started with a cross-validated random halving search.

$$\begin{aligned} H_{\text{initial}} &= \bigotimes_{i \in \text{XGBoost hyperparameters}} P_i \\ \text{where} \quad P_{n \text{ estimators}} &= [10, 500, 10], \\ P_{\text{maximum depth}} = [1, 10, 1], P_{\text{learning rate}} &= [0.01, 1, 0.01], \\ P_{\text{minimum child weight}} = [0.5, 1.5, 0.01], P_{\text{gamma}} &= [0, 0.5, 0.01], \\ P_{\text{column sample by tree}} = [0.5, 1, 0.01], P_{\text{subsample}} &= [0.5, 1, 0.01], \\ P_{\text{lambda}} = \{1\}, P_{\text{alpha}} = \{1\}, P_{\text{scale positive weight}} &= \{232.78\} \end{aligned} \tag{9}$$

After the cross-validated random halving search, the following hyperparameters achieved the best performance:

$$\begin{aligned} n \text{ estimators} = 100, \text{maximum depth} = 1, \\ \text{learning rate} = 0.34, \text{minimum child weight} = 0.86, \\ \text{gamma} = 0.36, \text{column sample by tree} = 0.93, \\ \text{subsample} = 0.6, \text{lambda} = 1, \\ \text{alpha} = 1, \text{scale positive weight} = 232.78 \end{aligned} \tag{10}$$

After the cross-validated random halving search, the model performed with a ROC AUC of 0.84 with a standard error of 0.11

- First cross iteration validated grid search - maximum

$$\begin{aligned} H_{\text{first grid search}} &= \bigotimes_{i \in \text{XGBoost hyperparameters}} P_i \\ \text{where} \quad P_{n \text{ estimators}} &= [50, 150, 10], \\ P_{\text{maximum depth}} = [1, 3, 1], P_{\text{learning rate}} &= [0.24, 0.34, 0.44], \\ P_{\text{minimum child weight}} = [0, 0.96, 0.01], P_{\text{gamma}} &= \{0.36\}, \\ P_{\text{column sample by tree}} = \{0.93\}, P_{\text{subsample}} &= \{0.6\}, \\ P_{\text{lambda}} = \{1\}, P_{\text{alpha}} = \{0\}, P_{\text{scale positive weight}} &= \{232.78\} \end{aligned} \tag{11}$$

After the first iteration of the cross-validated grid search, the following hyperparameters achieved the best performance.

$$\begin{aligned} n \text{ estimators} = 140, \text{maximum depth} = 1, \\ \text{learning rate} = 0.34, \text{minimum child weight} = 0, \\ \text{gamma} = 0.36, \text{column sample by tree} = 0.93, \\ \text{subsample} = 0.6, \text{lambda} = 1, \\ \text{alpha} = 0, \text{scale positive weight} = 232.78 \end{aligned} \tag{12}$$

After the first iteration of the cross-validated grid search, the model performed with a ROC AUC of 0.84 with a standard error of 0.09

- Second iteration of cross-validated grid search - In the next cross-validated grid search iteration, the model was tuned for the minimum samples split, and minimum samples leaf. The hyperparameter space consisted of

$$\begin{aligned} H_{\text{second grid search}} &= \bigotimes_{i \in \text{XGBoost hyperparameters}} P_i \\ \text{where} \quad P_{n \text{ estimators}} &= \{140\}, \\ P_{\text{maximum depth}} = \{1\}, P_{\text{learning rate}} &= \{0.34\}, \\ P_{\text{minimum child weight}} = \{0\}, P_{\text{gamma}} &= [0, 0.41, 0.01], \\ P_{\text{column sample by tree}} = [0.87, 0.98, 0.01], P_{\text{subsample}} &= [0.55, 0.65, 0.01], \\ P_{\text{lambda}} = \{1\}, P_{\text{alpha}} = \{0\}, P_{\text{scale positive weight}} &= \{232.78\} \end{aligned} \tag{13}$$

After the second iteration of the cross-validated grid search, the following hyperparameters achieved the best performance.

$$\begin{aligned} n \text{ estimators} = 140, &\text{maximum depth} = 1, \\ \text{learning rate} = 0.34, &\text{minimum child weight} = 0, \\ \text{gamma} = 0, &\text{column sample by tree} = 0.92, \\ \text{subsample} = 0.6, &\text{lambda} = 1, \\ \text{alpha} = 0, &\text{scale positive weight} = 232.78 \end{aligned} \tag{14}$$

After the second iteration of the cross-validated grid search, the model performed with a ROC AUC of 0.84 with a standard error of 0.09.

- Third iteration of cross-validated grid search - In the third cross-validated grid search iteration, the model was tuned for the alpha and lambda hyperparameters. The hyperparameter space consisted of

$$\begin{aligned} H_{\text{third grid search}} = &\bigotimes_{i \in \text{XGBoost hyperparameters}} P_i \\ \text{where} \quad P_{n \text{ estimators}} = \{140\}&, \\ P_{\text{maximum depth}} = \{1\}, P_{\text{learning rate}} = \{0.34\}&, \\ P_{\text{minimum child weight}} = \{0\}, P_{\text{gamma}} = \{0\}&, \\ P_{\text{column sample by tree}} = \{0.92\}, P_{\text{subsample}} = \{0.6\}&, \\ P_{\text{lambda}} = [0.8, 1.2, 0.01], P_{\text{alpha}} = [0, 0.2, 0.01]&, \\ P_{\text{scale positive weight}} = \{232.78\}& \end{aligned} \tag{15}$$

After the third iteration of the cross-validated grid search, the following hyperparameters achieved the best performance.

$$\begin{aligned} n \text{ estimators} = 140, &\text{maximum depth} = 1, \\ \text{learning rate} = 0.34, &\text{minimum child weight} = 0, \\ \text{gamma} = 0, &\text{column sample by tree} = 0.92, \\ \text{subsample} = 0.6, &\text{lambda} = 0.86, \\ \text{alpha} = 0.17, &\text{scale positive weight} = 232.78 \end{aligned} \tag{16}$$

After the third iteration of the cross-validated grid search, the model performed with a ROC AUC of 0.84 with a standard error of 0.09.

APPENDIX E - MODEL PERFORMANCE BEFORE AND AFTER HYPERPARAMETER TUNING

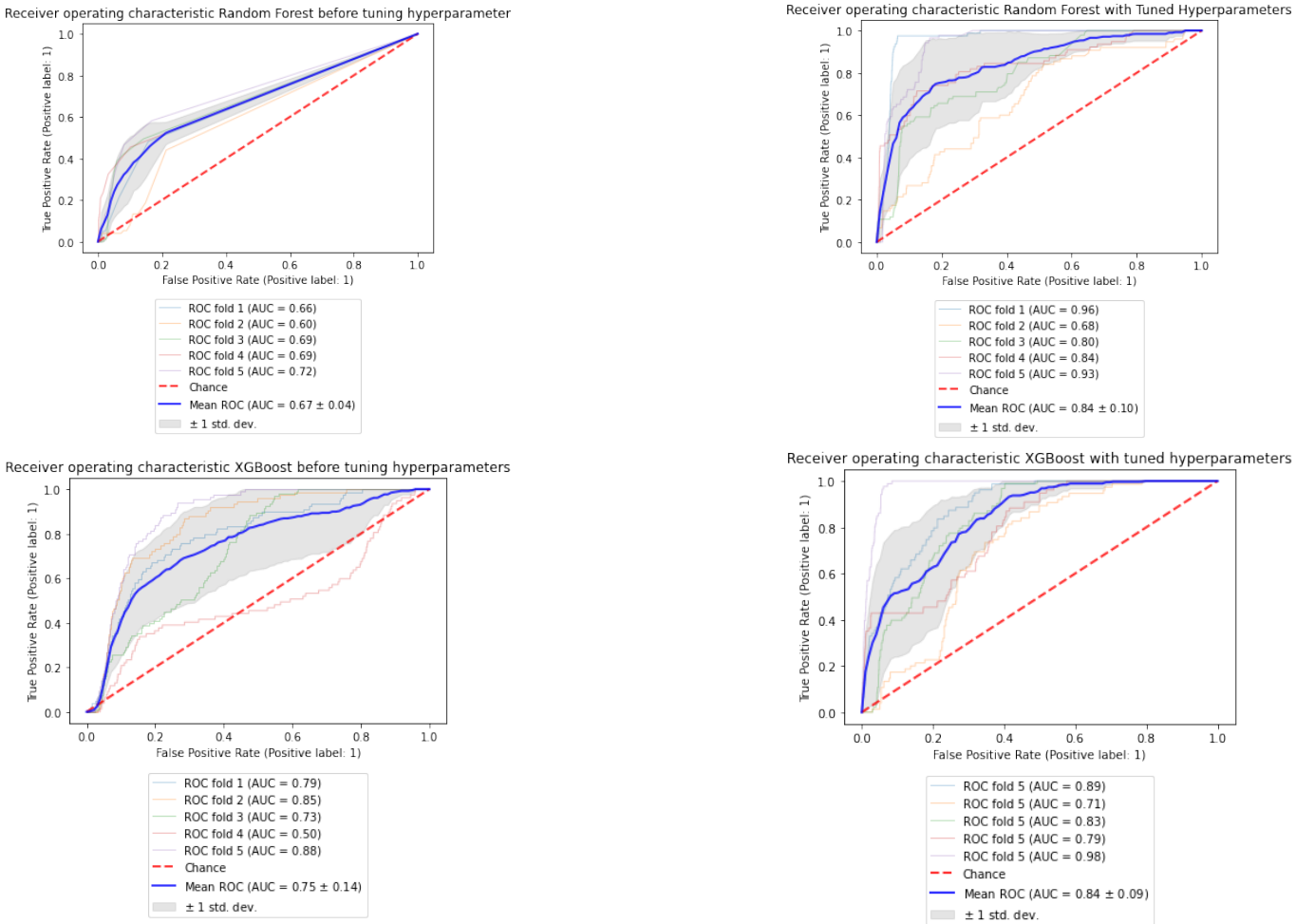

Fig. 1. The results of the AUC ROC for both Random Forests and XGBoost

APPENDIX F - FEATURE IMPORTANCE

.

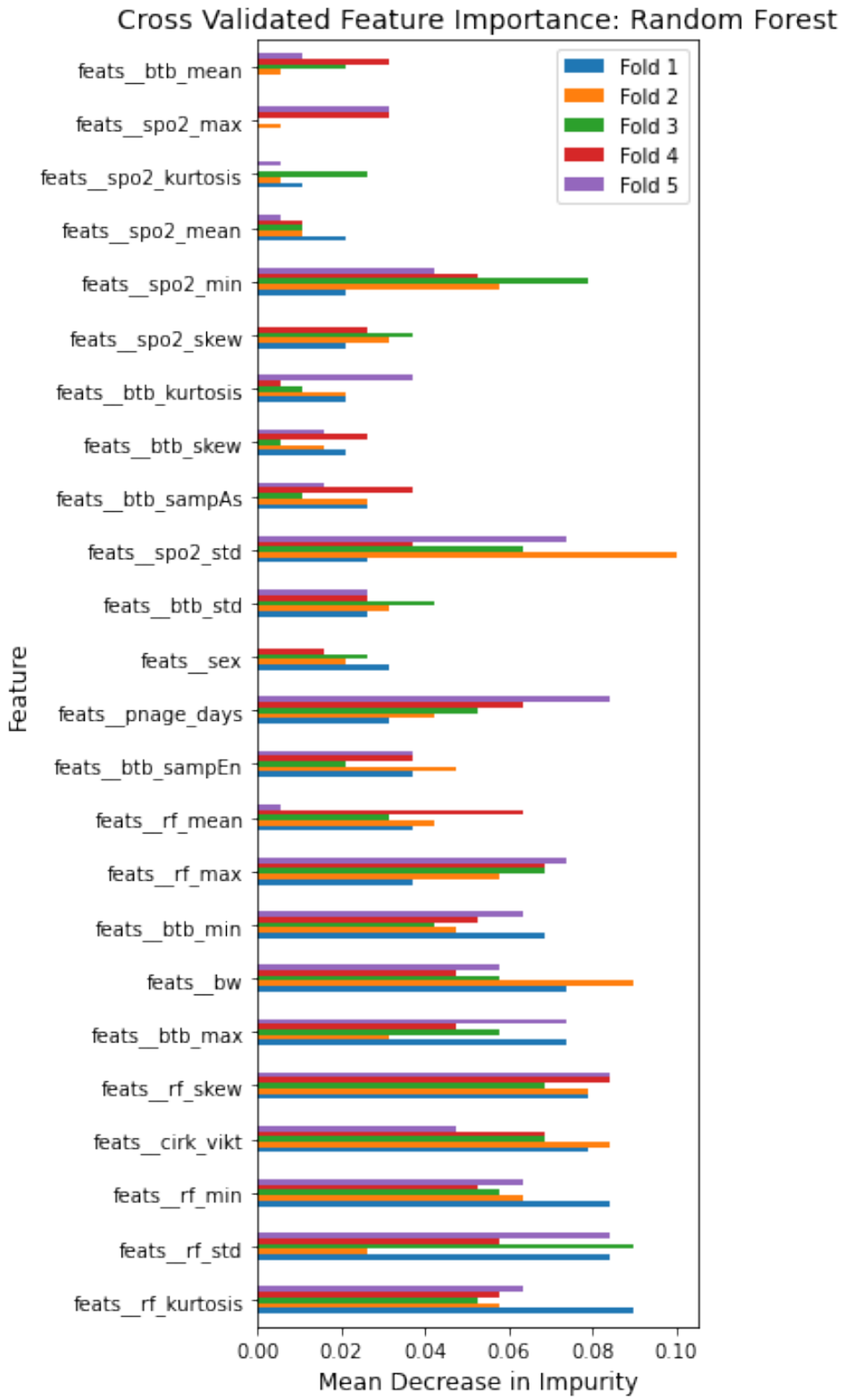

Fig. 2. Feature importance of Random Forest

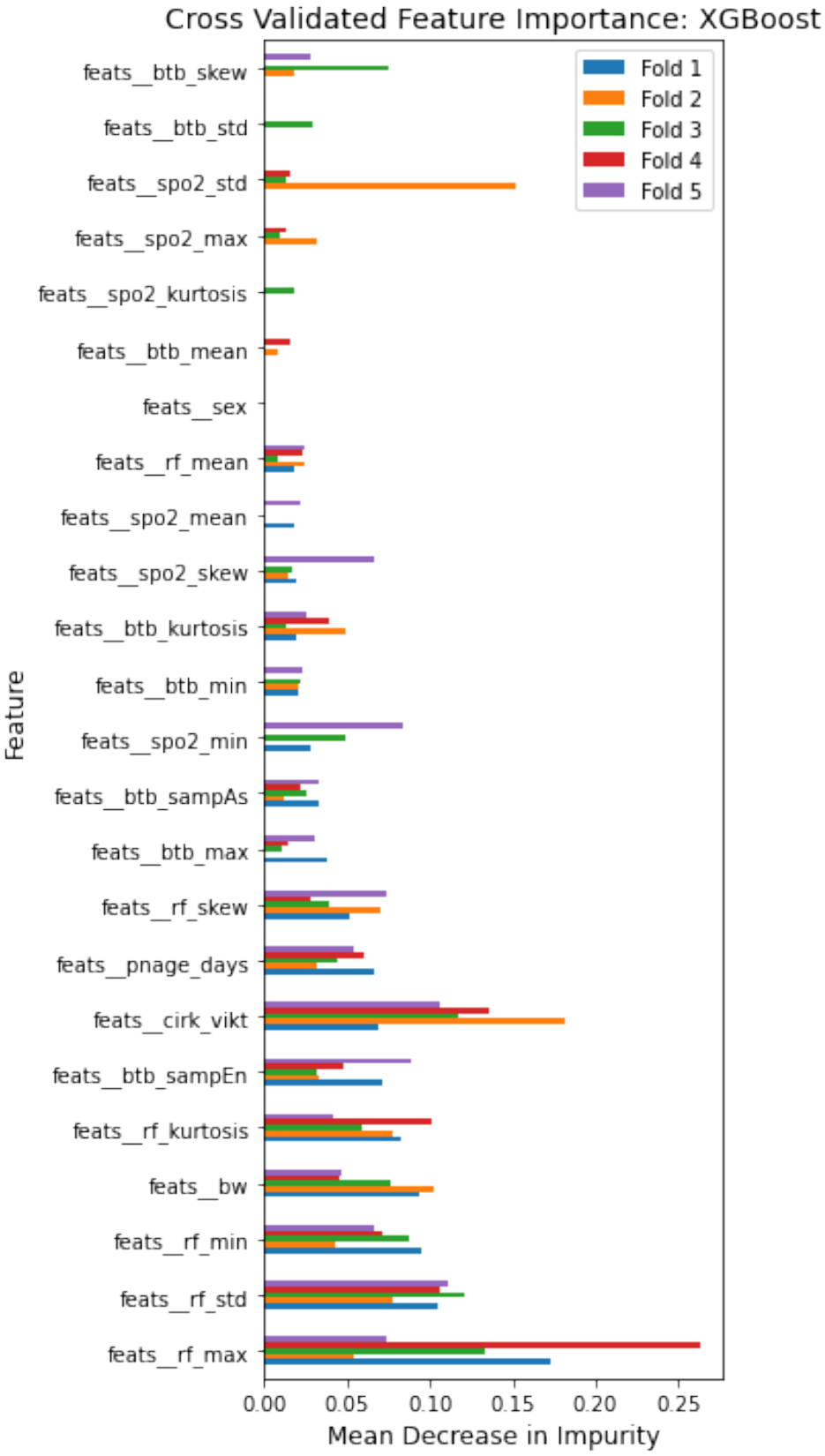

Fig. 3. Feature importance of XGBoost

APPENDIX G - ROC AUC SCORE DEPENDENCE ON HYPERPARAMTERS

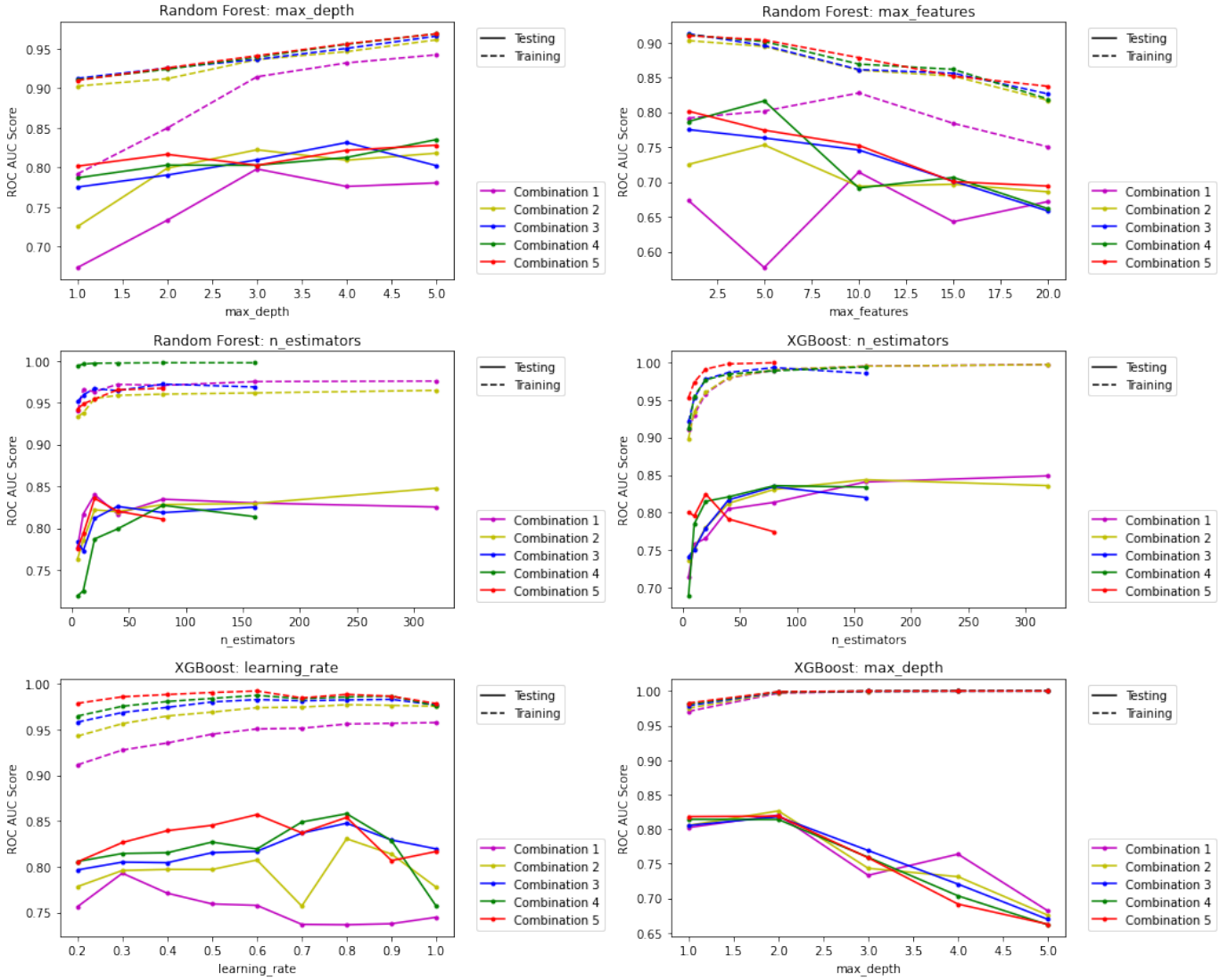

Fig. 4. Parameters impact the ROC AUC score for both Random Forest and XGBoost