

Table S1: **Low-level audio feature-based classification statistics.** The results of paired t-tests comparing cross-validations between trained models and corresponding majority-class baseline models. Each model type was calculated at the level of 30-second segments as well as tracks. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$.

Level	Sampling	Macro F1 (mean $\pm$ SD)	Model acc. (mean $\pm$ SD)	Baseline Acc. (mean $\pm$ SD)	t	p
Segment	Default	0.483 ± 0.089	0.507 ± 0.056	0.449 ± 0.045	2.419	0.073
Segment	SMOTE	0.502 ± 0.054	0.509 ± 0.039	0.449 ± 0.045	3.063	0.038*
Segment	UNDER	0.478 ± 0.022	0.482 ± 0.018	0.196 ± 0.056	8.719	0.001***
Segment	Clipping	0.376 ± 0.179	0.417 ± 0.155	0.211 ± 0.122	2.327	0.081
Segment	Boundaries	0.430 ± 0.030	0.480 ± 0.020	0.481 ± 0.077	-0.020	0.985
Track	Default	0.554 ± 0.032	0.583 ± 0.054	0.434 ± 0.017	7.426	0.002**
Track	SMOTE	0.553 ± 0.043	0.569 ± 0.057	0.203 ± 0.012	11.356	0.000***
Track	UNDER	0.514 ± 0.039	0.531 ± 0.051	0.203 ± 0.012	12.399	0.000***
Track	Clipping	0.498 ± 0.174	0.580 ± 0.133	0.427 ± 0.090	1.572	0.191
Track	Boundaries	0.520 ± 0.062	0.568 ± 0.066	0.452 ± 0.030	2.661	0.056