

Table A. P-values for significance tests (t-test) for Tasks 1-12, where $p < 0.05$ are highlighted. The cases where OPERA models outperform the best baseline are underlined.

Dataset	ID	Best baseline	OPERA-CT	OPERA-CE	OPERA-GT
UK COVID-19	T1	VGGish	<u>0.0001</u>	0.0002	<u>0.4230</u>
	T2	CLAP	<u>0.0000</u>	<u>0.0022</u>	<u>0.0000</u>
COVID-19 Sounds	T3	CLAP	0.0075	0.2155	<u>0.9161</u>
	T4	CLAP	<u>0.0558</u>	0.0000	0.0011
CoughVID	T5	CLAP	0.0003	0.0003	0.0000
	T6	CLAP	<u>0.0000</u>	<u>0.0000</u>	<u>0.0000</u>
ICBHI	T7	CLAP	0.0000	0.0000	0.0000
Coswara	T8	CLAP	<u>0.1586</u>	0.8547	0.0000
	T9	Opensmile	<u>0.0000</u>	<u>0.0000</u>	<u>0.0000</u>
KAUH	T10	CLAP	<u>0.0183</u>	<u>0.0003</u>	<u>0.9875</u>
Respiratory@TR	T11	CLAP	0.7182	<u>0.0439</u>	0.4200
SSBPR	T12	Opensmile	<u>0.0027</u>	0.9944	0.0000