Electronic Supplementary Material (ESI) for Analytical Methods.

Supplementary materials

Bury et al_revised_24.101.2022

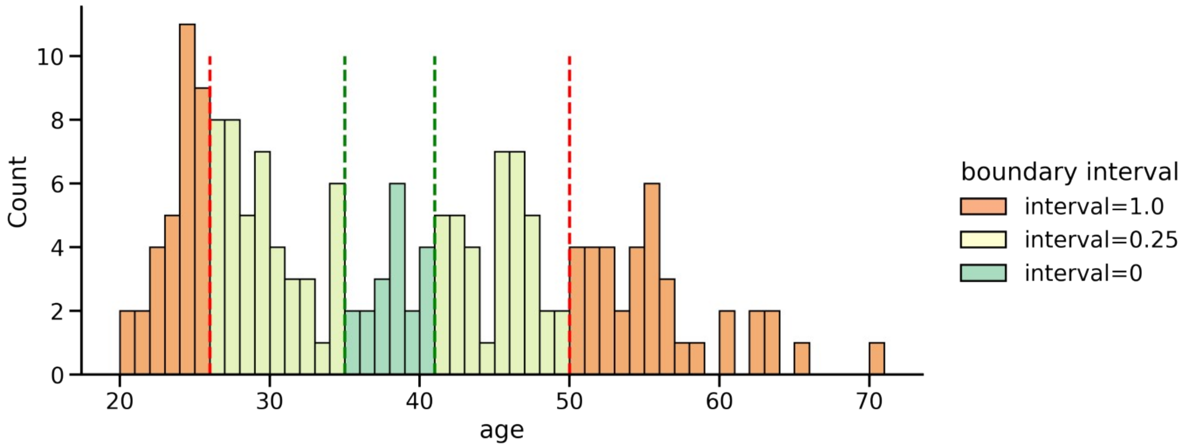


Fig. S1. Illustration of the ε masking strategy applied to the full male & female sample set. Vertical dashed lines are overlaid over the age distribution across fingerprint samples, to indicate specific intervals around the mean, μ_{age}, at distinct ε boundary values, illustrating the proportion of data samples that will be masked during model training/evaluation at each fixed $\varepsilon > 0$ value. At an $\varepsilon = 1$ value, only fingermark samples corresponding to individuals aged <26 or >50 are included for model building/validation. The practical usage of binary age classification is limited, particularly for models trained at high $\varepsilon \geq 1$, whereby the model is trained to only distinguish between samples at the two extremities of the sample age distribution.

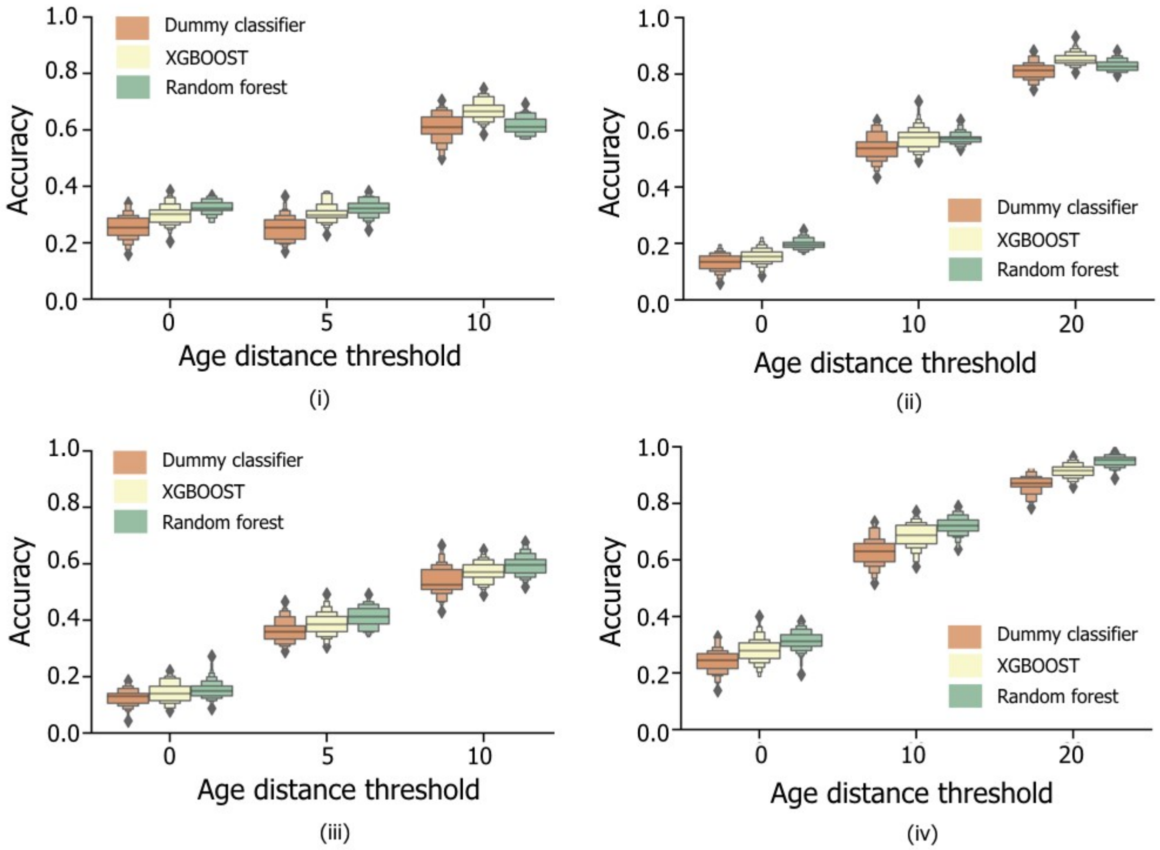
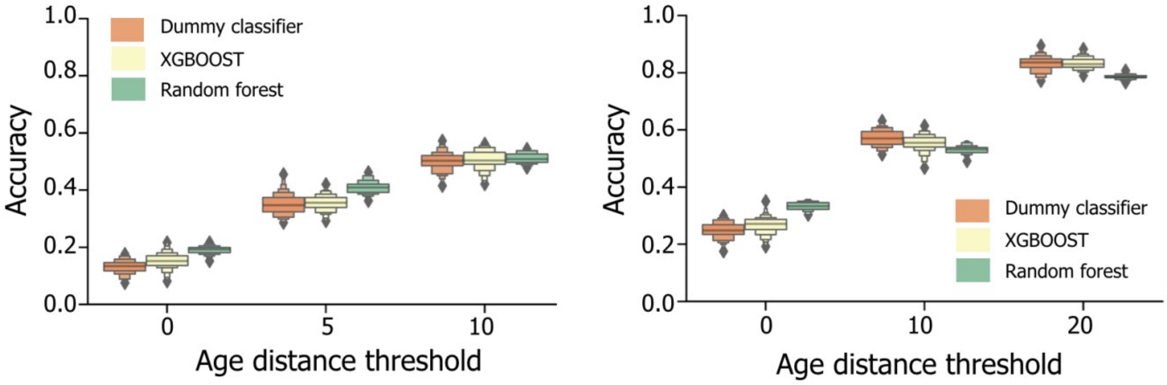


Fig. S2. Distribution of categorical age prediction scores across XGBOOST, random forest and baseline dummy classification model types, with the inclusion of under sampling of the most represented age bins during the CV-training strategy (Fig. 2). In (i) and (iii) the age bin width of 5 is used, whereas in (ii) and (iv) 10 year age bins are used. In (iii) and (iv), a further sample filtering step has been included, by omitting 149 individual fingerprint spectra from CV model evaluation, which had been identified as containing significant polymer contamination (Table S1). For both bin widths, consistent with Fig. 5, the effect of counting predictions made in neighboring bins to the true age bin are included, corresponding to the modified accuracy scores discussed in Section 2.3.

Sample Sex	# non-contaminated samples	# samples with minor contamination	# highly contaminated samples
Male	172	12	62
Female	176	10	87

Table S1. Contamination states identified for each individual fingerprint spectra. Three spectra were collected per individual, over a set of 172 individuals.



(i) (ii)

Fig. S3. Distribution of categorical age prediction scores across XGBOOST, random forest and baseline dummy classification model types, for *enhanced* fingermark samples. For age bin widths (i) 5 and (ii) 10. For both bin widths, the effect of counting predictions made in neighboring bins that within a specified age-difference of the true age (*x*-axis) to the true age bin are included, corresponding to the modified accuracy scores discussed in Section 2.3.