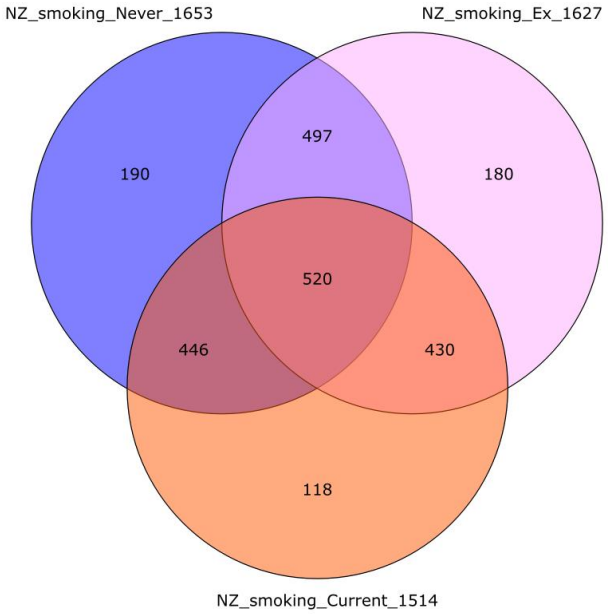
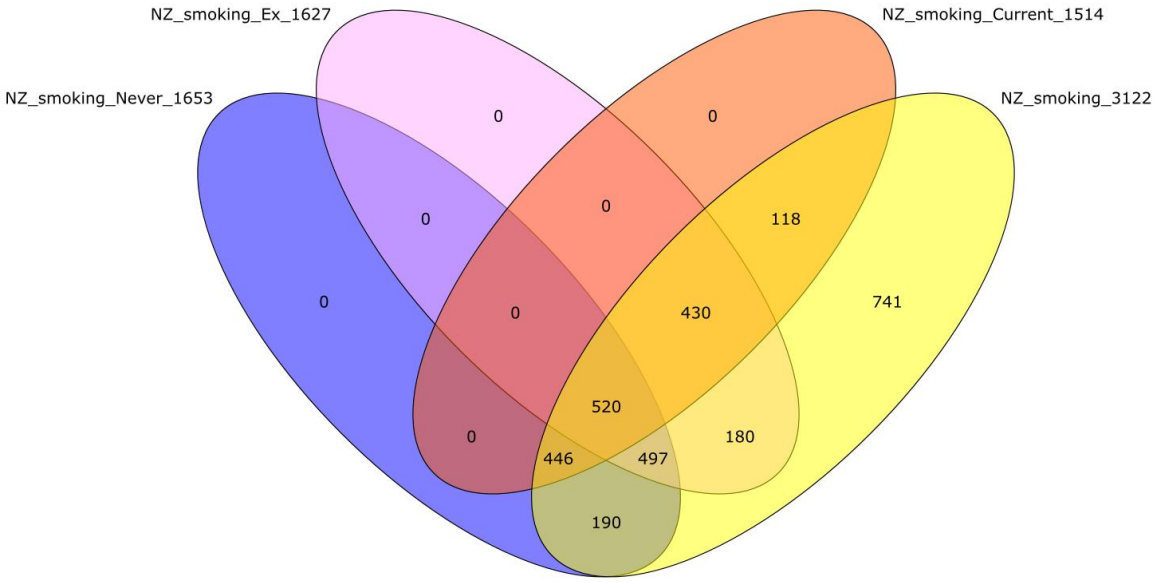
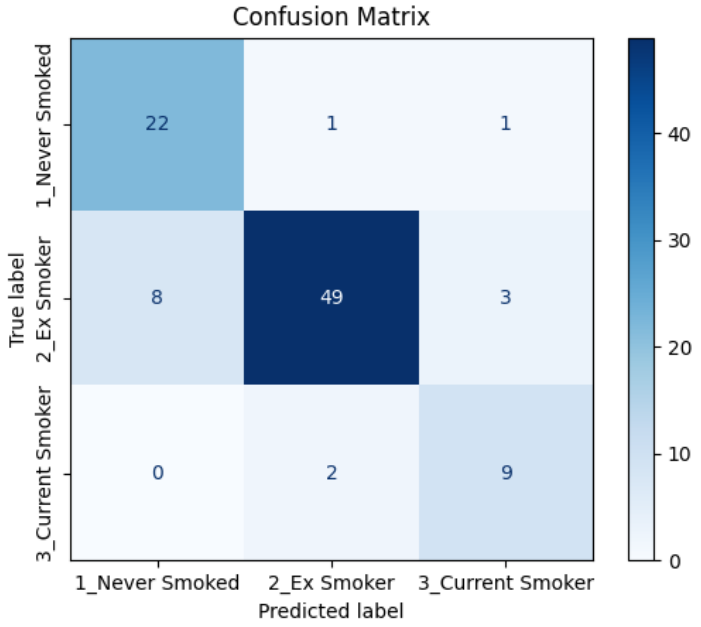


(a) Intersection of sets of CpG sites used to predict never-, ex- and current-smokers. While a total of 2381 CpG sites are used in the entire model, only 520 CpG sites are used in all three predictions, and each sub-predictor only uses approximately 1600 CpG sites.

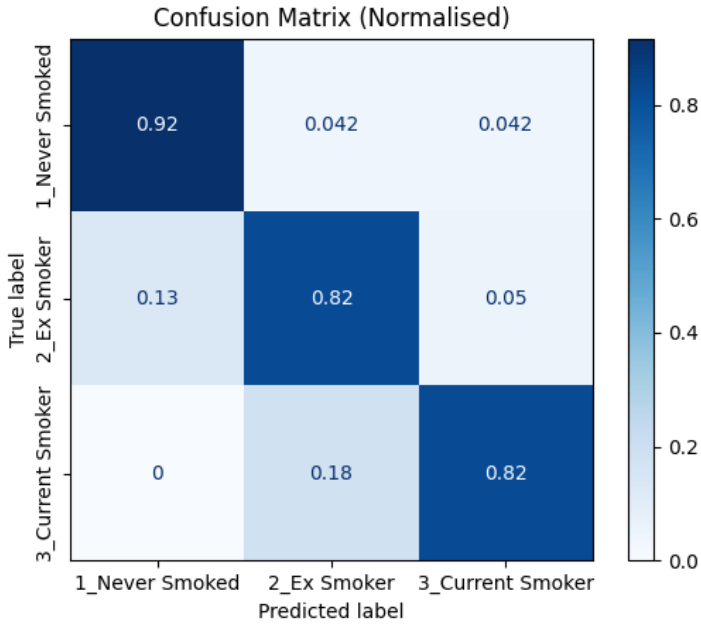


(b) Same diagram above intersected with the original 3122 CpG sites returned from Kruskal-Wallis.

Figure 1: Intersections of sets of CpG sites used in model

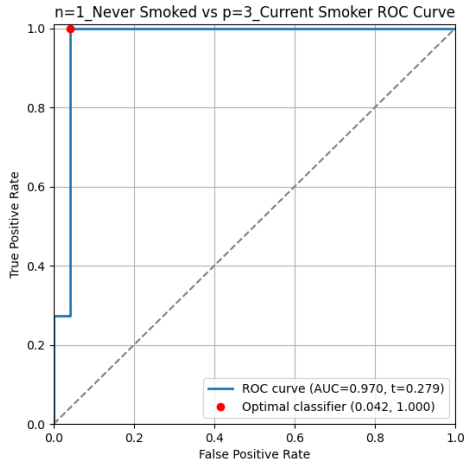


(a) Confusion Matrix for Cohort 1 test set. The model makes very few incorrect predictions.

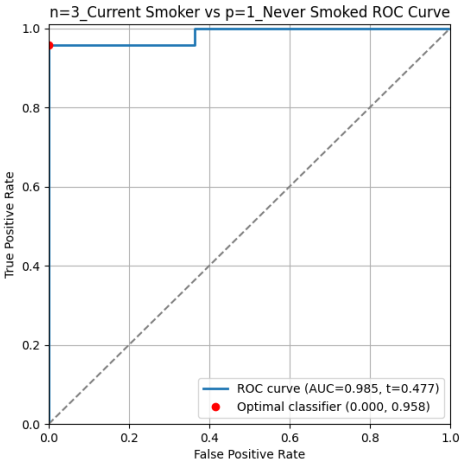


(b) Confusion Matrix (normalised) for Cohort 1 test set. As before, we can see the model achieves near-perfect predictive performance.

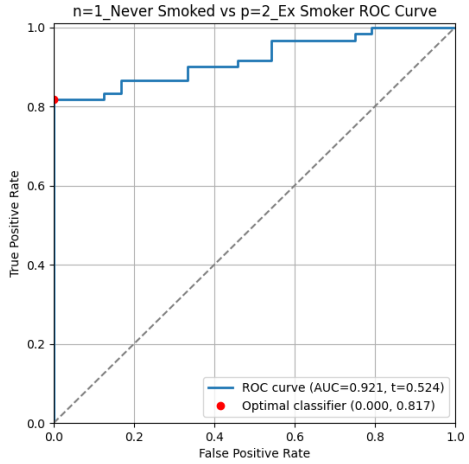
Figure 2: Confusion Matrices (Cohort 1 test set)



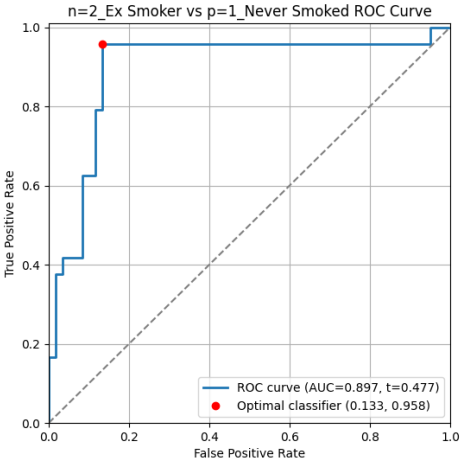
(a) Separation of never- vs current-smokers by current-smoker sub-predictor



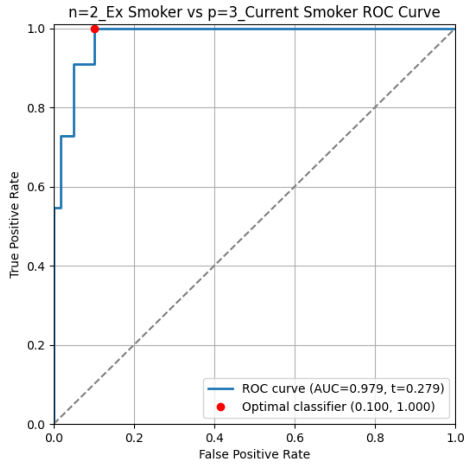
(b) Separation of never- vs current-smokers by never-smoker sub-predictor



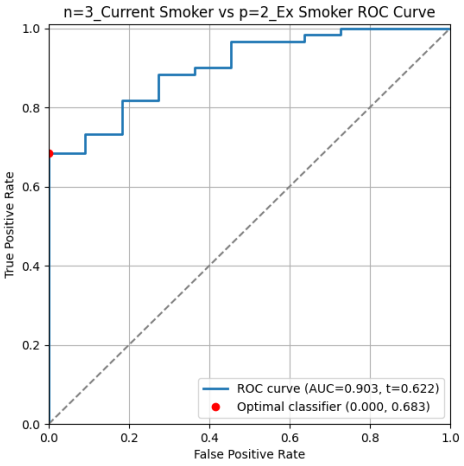
(c) Separation of never- vs ex-smokers by ex-smoker sub-predictor



(d) Separation of never- vs ex-smokers by never-smoker sub-predictor



(e) Separation of ex vs current-smokers by current-smoker sub-predictor



(f) Separation of ex vs current-smokers by ex-smoker sub-predictor

Figure 3: One vs one sub-predictor performance (Cohort 1 test set)

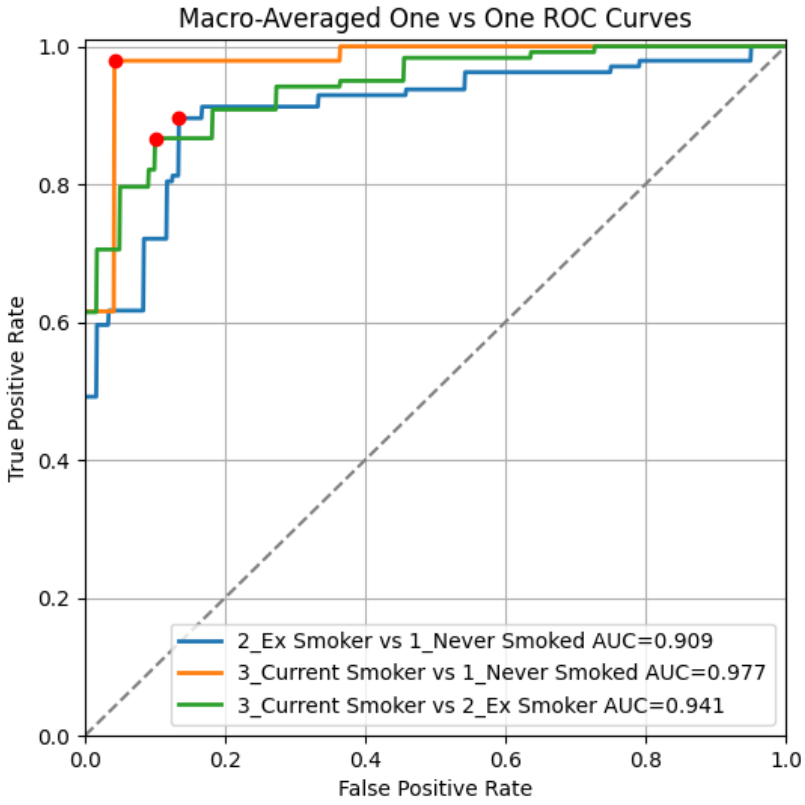
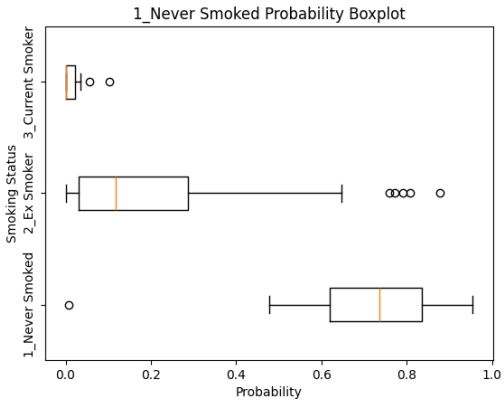
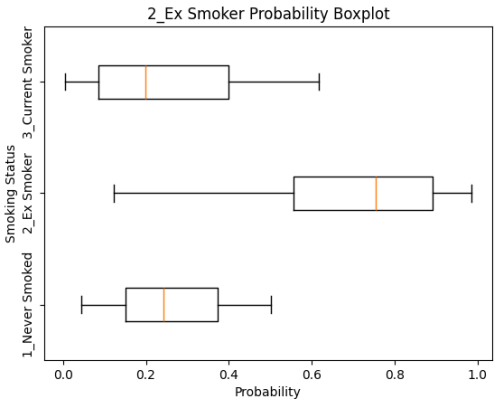


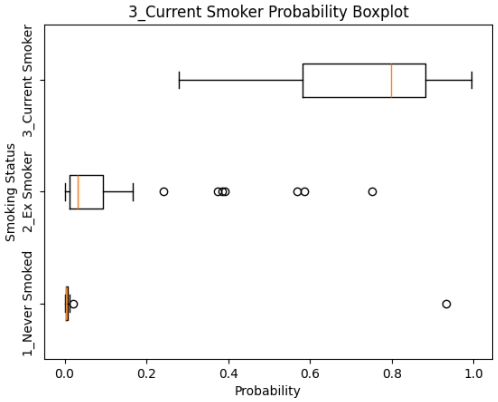
Figure 4: Class separation of classifier (Cohort 1 test set)



(a) Never-smoker sub-predictor probabilities across all three classes

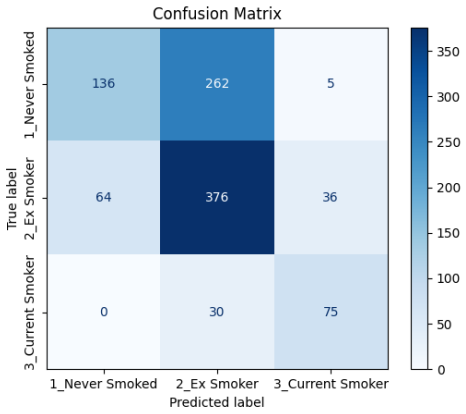


(b) Ex-smoker sub-predictor probabilities across all three classes

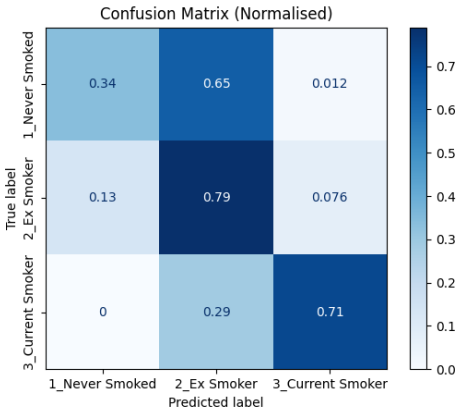


(c) Current- smoker sub-predictor probabilities across all three classes

Figure 5: Boxplots of probability distributions (Cohort 1)

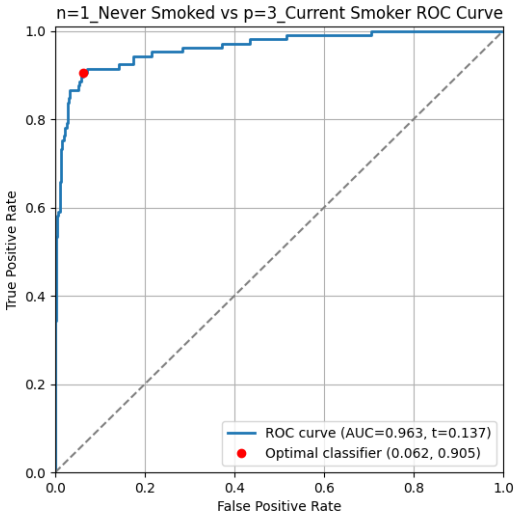


(a) CM

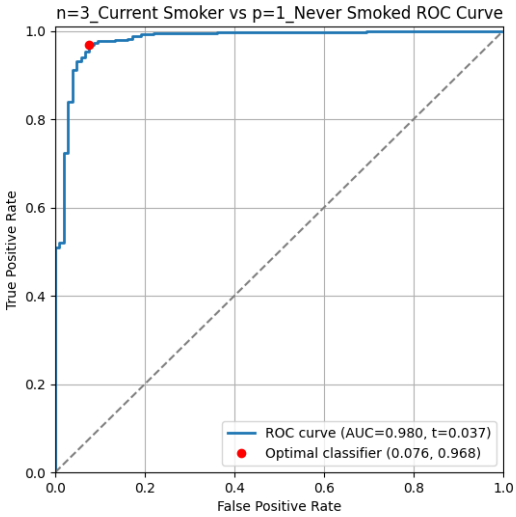


(b) CM Normalised

Figure 6: Confusion Matrices (Cohort 2)

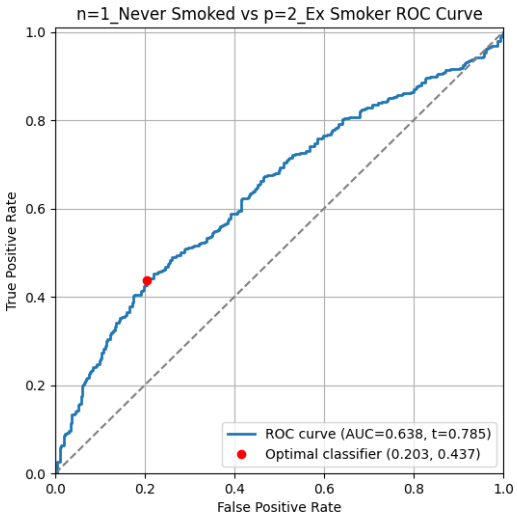


(a) Seperation of never vs current smokers by current smoker sub-classifier

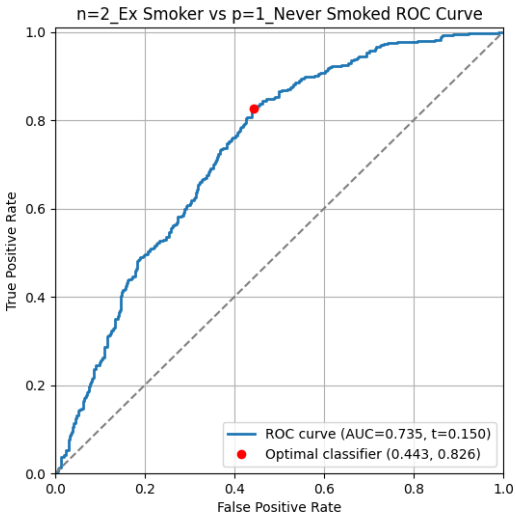


(b) Seperation of never vs current smokers by never smoker sub-classifier

Figure 7: Never smoker vs current smoker sub-classifiers performance (Cohort 2)

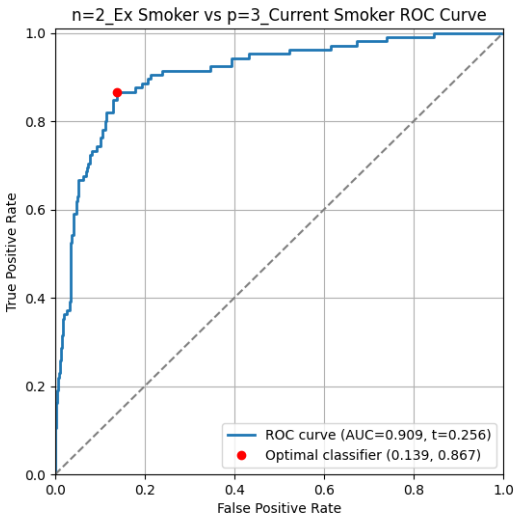


(a) Seperation of never vs ex-smokers by ex-smoker sub-classifier

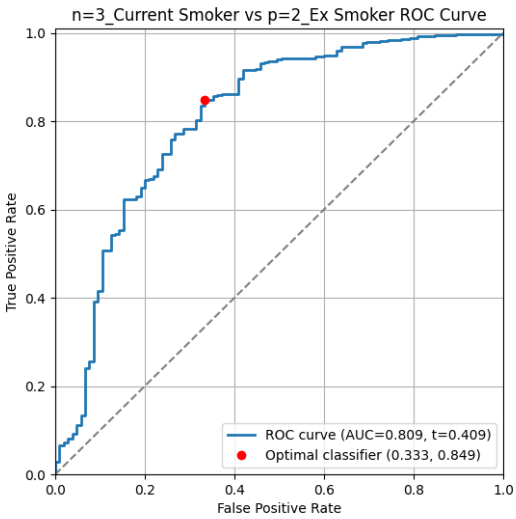


(b) Seperation of never vs ex-smokers by never smoker sub-classifier

Figure 8: Never smoker vs ex-smoker sub-classifiers performance (Cohort 2)



(a) Seperation of ex vs current smokers by current smoker sub-classifier



(b) Seperation of ex vs current smokers by ex-smoker sub-classifier

Figure 9: Ex-smoker vs current smoker sub-classifiers performance (Cohort 2)

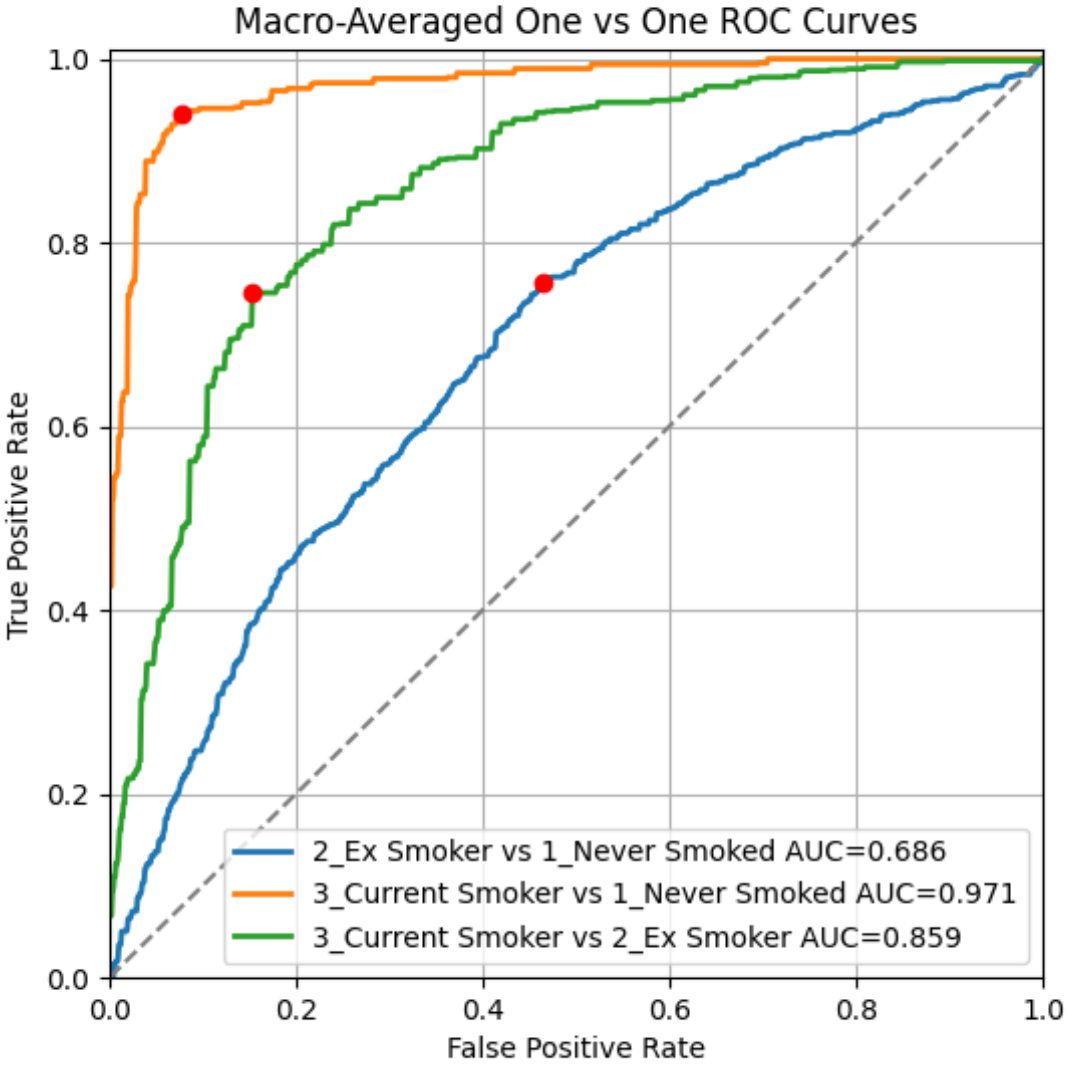
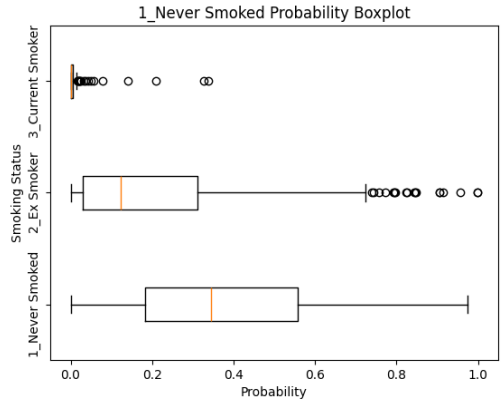
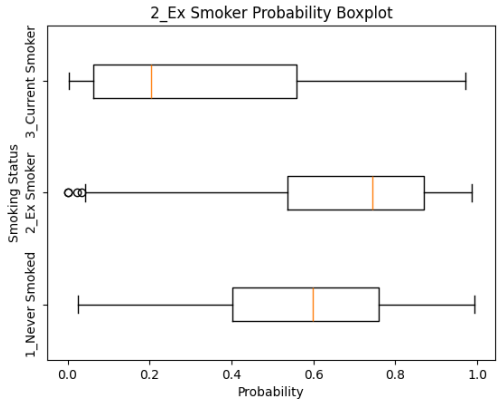


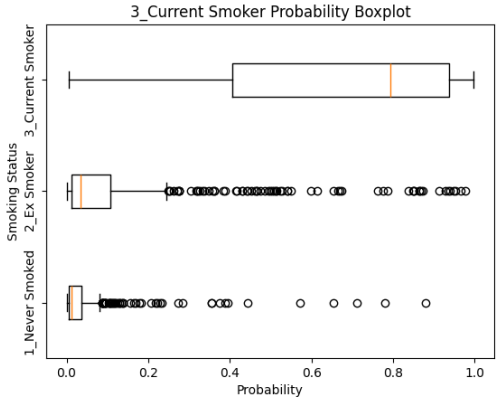
Figure 10: Class seperation of classifier (Cohort 2)



(a) Never smoker sub-classifier probabilities across all three classes

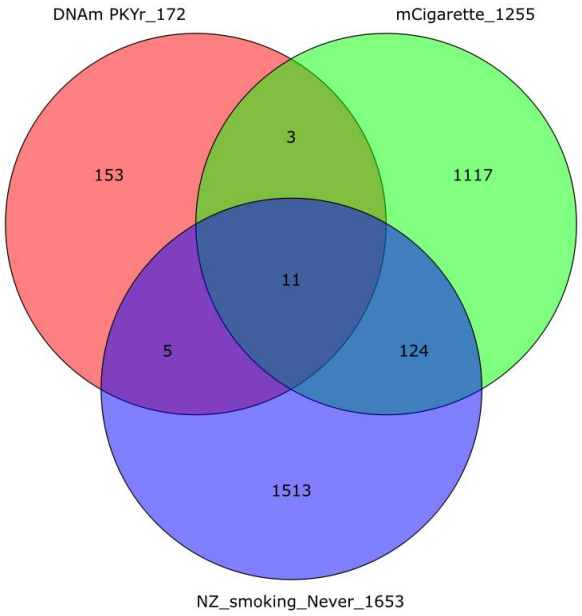


(b) Ex-smoker sub-classifier probabilities across all three classes



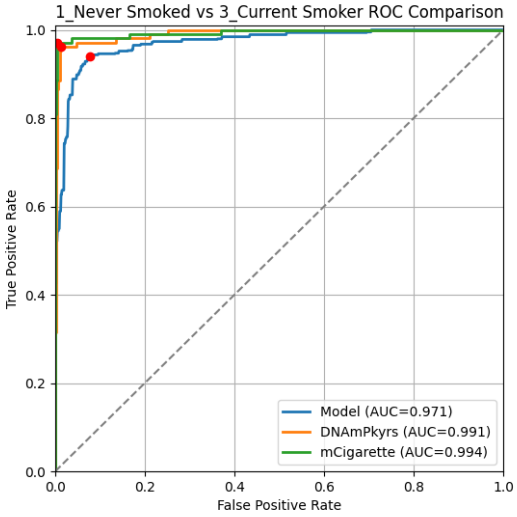
(c) Current smoker sub-classifier probabilities across all three classes

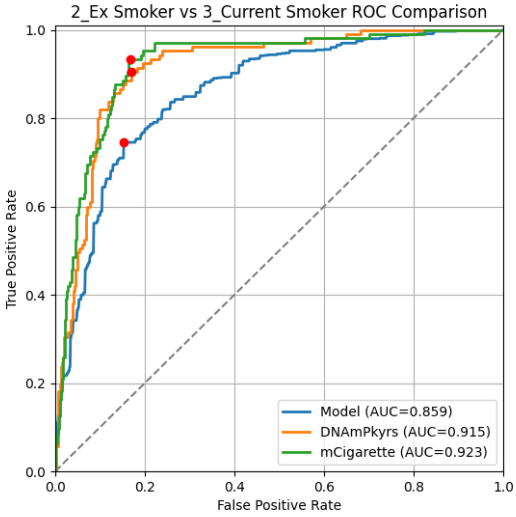
Figure 11: Boxplots of probability distributions (Cohort 2)

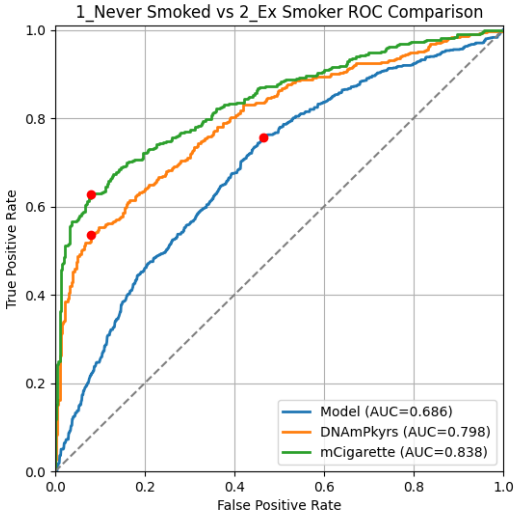



DNAm PKYr_172
mCigarette_1255
154
4
1145
10
4
96
1517
NZ_smoking_Ex_1627


DNAm PKYr_172
mCigarette_1255
153
4
1137
10
5
104
1395
NZ_smoking_Current_1514


1_Never Smoked vs 3_Current Smoker ROC Comparison
True Positive Rate
False Positive Rate
Model (AUC=0.971)
DNAmPkyrs (AUC=0.991)
mCigarette (AUC=0.994)


2_Ex Smoker vs 3_Current Smoker ROC Comparison
True Positive Rate
False Positive Rate
Model (AUC=0.859)
DNAmPkyrs (AUC=0.915)
mCigarette (AUC=0.923)


1_Never Smoked vs 2_Ex Smoker ROC Comparison
True Positive Rate
False Positive Rate
Model (AUC=0.686)
DNAmPkyrs (AUC=0.798)
mCigarette (AUC=0.838)