

Figures

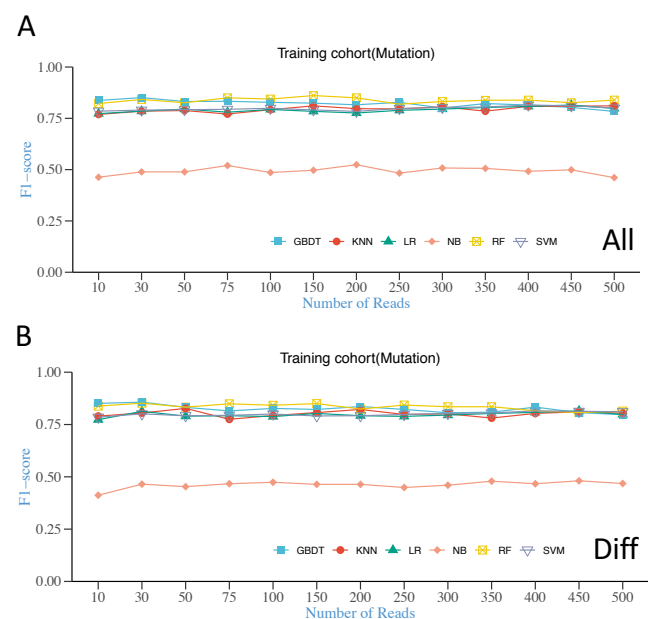


Figure S1. Performance of six models at different minimum reads thresholds

The x-axis represents the set minimum reads number, and the y-axis represents the F1 score.

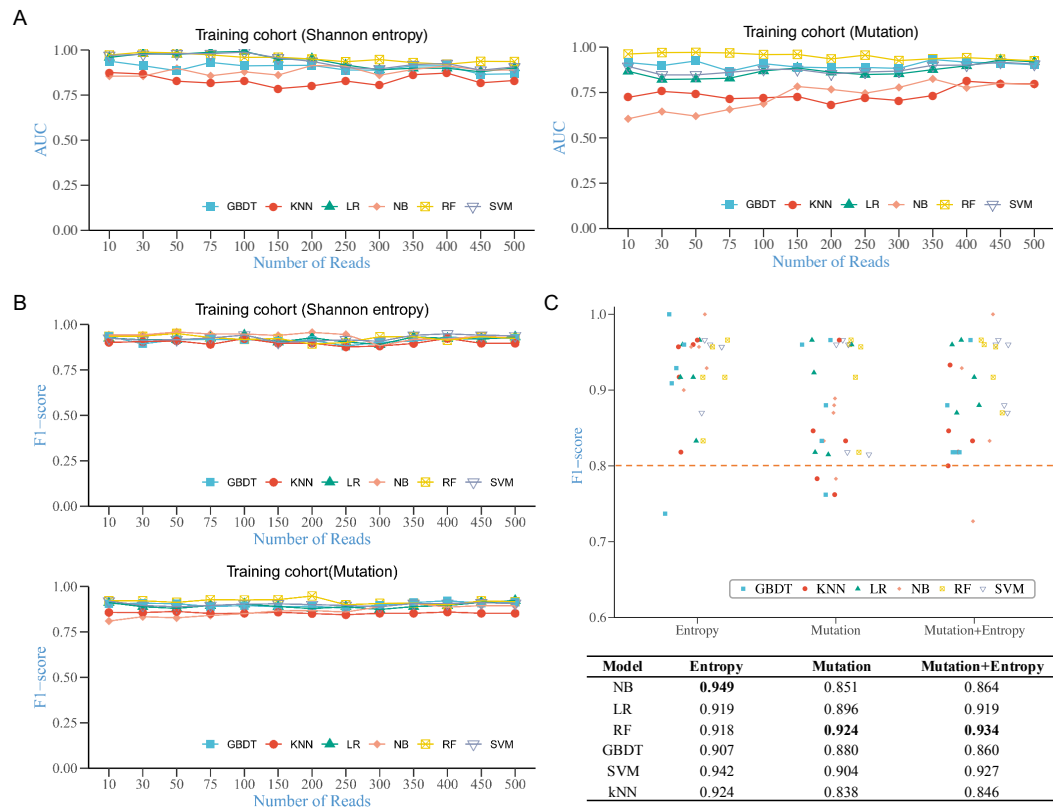


Figure S2. Impact of different read counts at different loci on the performance of six models

A,B. Performance of the six models at different read count threshold. The x-axis represents the minimum read counts set for each nucleotide position, and the y-axis represents the AUC or F1 scores.

C. F1 scores of the six ML classification models on three sequence features in five-fold cross-validation. The numerical values in the table within the figure represent the average F1 scores of each model in five-fold cross-validation.

Tables

Table S1. Baseline information of clinical samples

Baseline characteristics	NPC (n=13)	Non-NPC (n=13)	P-value
gender,n(%)			1
Male	8(64.3%)	7(60.0%)	
Female	5(35.7%)	6(40.0%)	
age,years			1
median	44	51	
range	41-69	28-81	
EBV copy number at baseline (10 ³ IU/ml)			1
median	12.000	8.790	
range	1.540-3050.000	0.767-162.000	
LMP1-1 sequencing depth			0.12
median	23116	4018	
range	5238-144796	675-306018	

Table S2. EBV LMP1 fragment PCR amplification primers

Name		Sequence (5'-3')
EBV-NEST	Forward	CCATTGTCATCAGTGTTGTCAGG
	Reverse	GCAGACCCCGCAAATCCC
LMP1	Forward	CCAAGTAAGCACCCGAAGATG
	Reverse	CTTGCTCCTGACACACTG

Table S3. Model parameter settings

Model	Parameter	Mutation	Entropy	Mutation + Entropy
KNN	n_neighbors	4	11	9
LR	C	0.01	100	10000
GBDT	min_samples_split	9	6	8
RF	n_estimators	60	70	20
	C	1	100	100
SVM	degree	3		
	kernel	poly	rbf	linear
	gamma		1	

Table S4. Comparison of classification performance of six models on the test dataset

Feature	Model	Accuracy	Precision	Recall	F1 score
Mutation	NB	0.808	0.869	0.883	0.876
	LR	0.859	0.877	0.950	0.912
	RF	0.910	0.921	0.967	0.943
	GBDT	0.782	0.852	0.867	0.860
	kNN	0.756	0.860	0.817	0.838
	SVM	0.872	0.879	0.967	0.921
Entropy	NB	0.910	0.921	0.967	0.943
	LR	0.897	0.919	0.950	0.934
	RF	0.885	0.905	0.950	0.927
	GBDT	0.872	0.891	0.950	0.919
	kNN	0.846	0.887	0.917	0.902
	SVM	0.897	0.933	0.933	0.933
Mutation+Entropy	NB	0.808	0.857	0.900	0.878
	LR	0.872	0.903	0.933	0.918
	RF	0.872	0.879	0.967	0.921
	GBDT	0.910	0.921	0.967	0.943
	kNN	0.821	0.838	0.950	0.891
	SVM	0.923	0.935	0.967	0.951