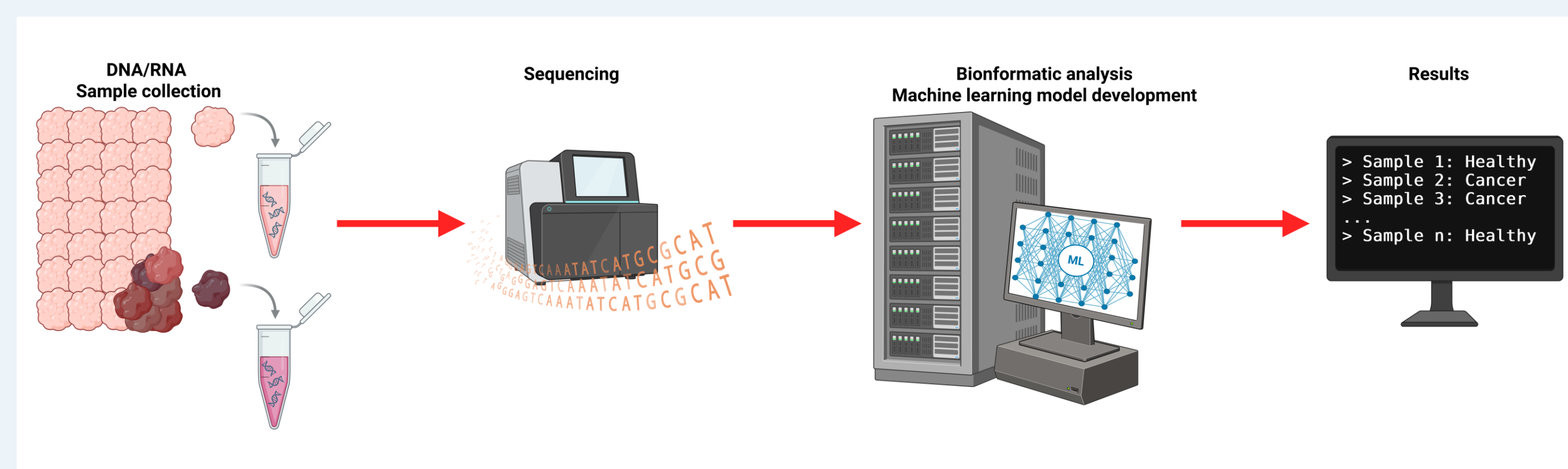


BACKGROUND

Personalised medicine involves detecting patient-specific biomarkers which enable the early detection of the disease thus significantly improving clinical outcomes. In cancer research, epigenomic and transcriptomic data can be used as input in machine learning methodologies in order to identify biomarkers and develop models for cancer detection [1-3]. However, the high dimensionality of omics data and limited sample sizes makes model development especially challenging. Such issues increase the risks of producing suboptimal models with poor generalisability, which can compromise the clinical utility of the biomarkers developed [4]. These risks can only be mitigated through rigorous internal validation which is usually performed by a simple train-test split. Despite its importance, there is limited research into the validity of alternative internal validation methods and therefore a lack of clear consensus on which approaches are most effective.

AIM

The aim of our study was to evaluate the effect of four internal validation methods: train-test split, bootstrapping (including .632 and .632+ corrections), 5-fold cross-validation (CV) and leave-one-group-out cross-validation, on the performance of machine learning classifiers trained on DNA methylation and gene expression data for cancer detection.



INTERNAL VALIDATION COMPARISON

Data collection

Two DNA methylation datasets generated on the Illumina Infinium HumanMethylation 450K BeadChip platform from colorectal cancer patients' samples were downloaded from Gene Expression Omnibus (GEO). Additionally, two gene expression datasets generated with RNA sequencing (RNA-seq), from lung adenocarcinoma patients' samples were downloaded from the same database. All sample sets included matched samples from tumour and adjacent normal tissues.

Cancer type	Sample type	Data type	Data extraction technology	GEO accession number	No of tumour samples	No of healthy samples	Usage
Colorectal cancer	Colon tissue	DNA methylation	Illumina HumanMethylation450k BeadChip	GSE131013	92	92	Training
				GSE193535	53	53	External validation
Lung Adenocarcinoma	Lung tissue	Gene expression	Illumina NovaSeq 6000	GSE229705	123	123	Training
			HiSeq X Ten	GSE14043	49	49	External validation

Data preprocessing

Pre-processing of raw DNA methylation (IDAT files) and RNA-seq data (FASTQ files) was carried out on Apocrita, the High Performance Cluster based at Queen Mary University of London. Differential DNA methylation and differential gene expression analysis were conducted in R version 4.4.1. to detect differential methylated cytosines (DMCs) and differentially expressed genes (DEGs) between healthy and cancer samples.

Machine learning models for cancer detection

To distinguish cancer from normal samples, three supervised ML classifiers were employed using Python's *scikit-learn* library. Support vector machines (SVM), random forest (RF) and multilayer perceptron (MLP) models were selected based on their established effectiveness in cancer detection in multiple studies. The DNA methylation based classifiers were trained using the beta values from the top 20 DMCs as input features while the RNA-seq based classifiers were trained on the VST-transformed read counts of the top 20 DEGs as input features.

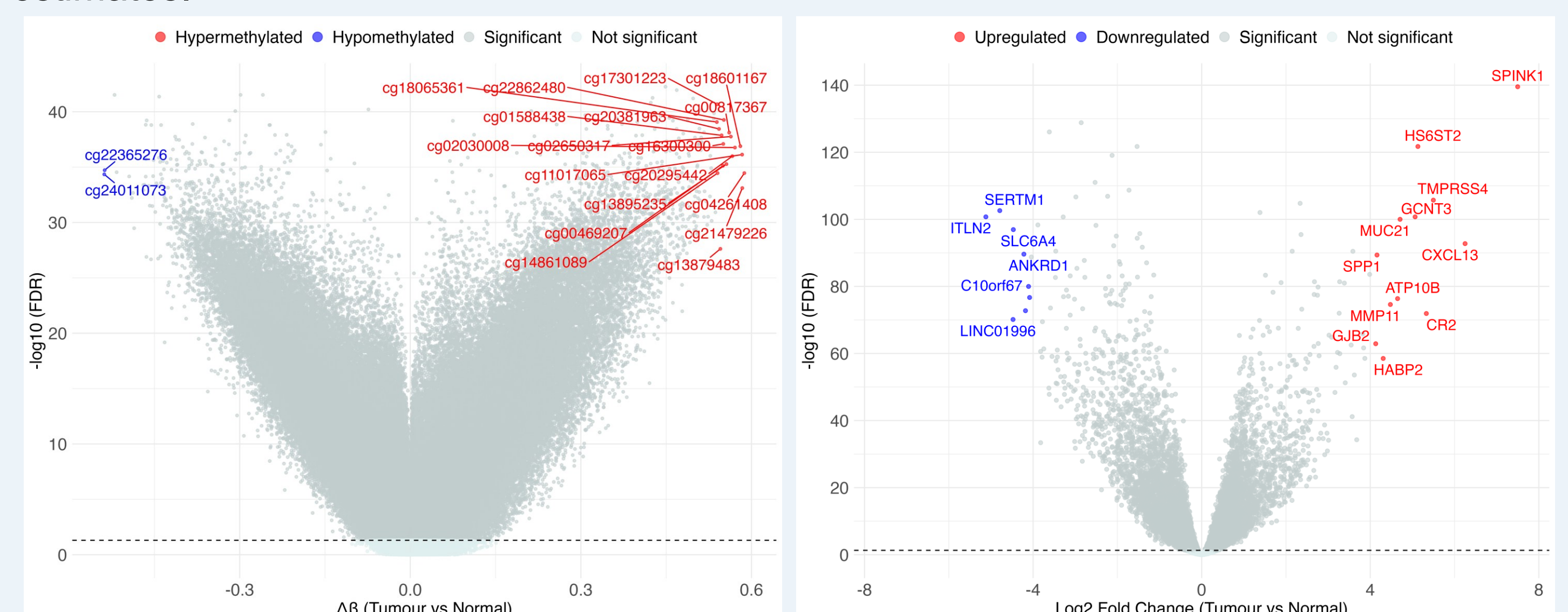
Internal validation methods

ML classifiers for cancer detection were developed and evaluated using four different internal validation methods: train-test split, bootstrapping (including .632 and .632+ corrections), 5-fold CV and LOGOCV. AUC was used as the primary performance metric for model comparison, and sensitivity as a secondary criterion.

RESULTS

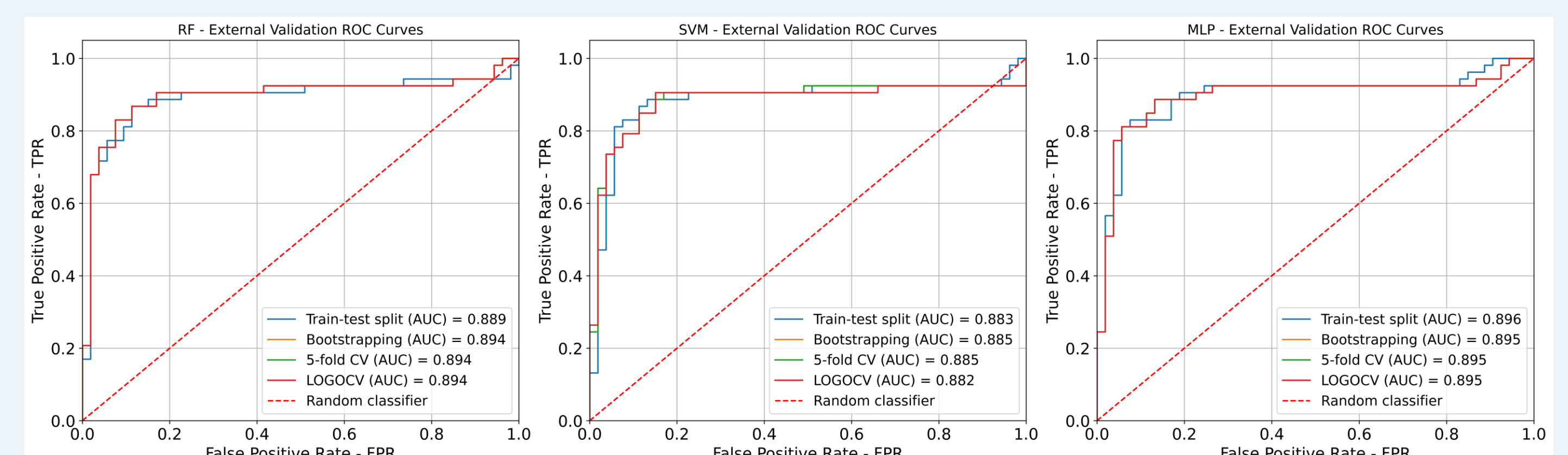
Identification of DMCs and DEGs

Differential methylation analysis using limma on M-values identified 259,062 CpGs sites as significantly differentially methylated (FDR < 0.05). DMCs were ranked on their absolute $\Delta\beta$ values. From the differential gene expression analysis, a total of 12,959 genes were identified as significantly differentially expressed between tumour and normal samples based on an FDR threshold of < 0.05. Log2FC estimates were shrunk to improve effect size estimation and genes were ranked by the absolute value of these estimates.



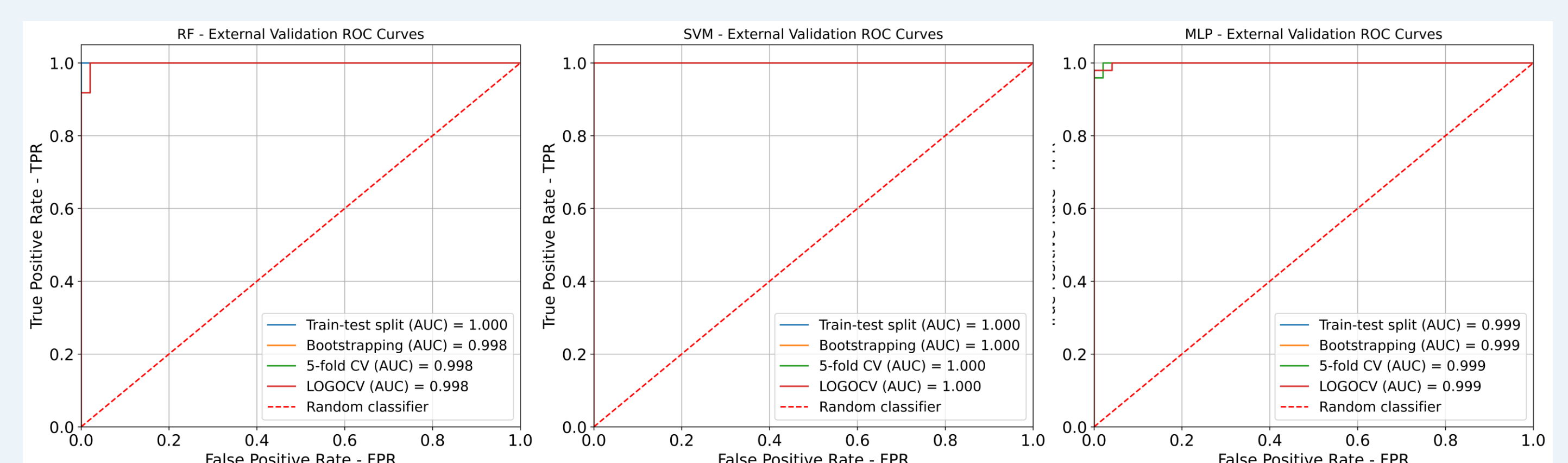
DNA methylation-based classifiers

The models developed using 5-fold CV performed the best in distinguishing between tumour and normal samples for both SVM and RF. For SVM, 5-fold CV yielded an AUC of 0.885 and RF achieved an AUC of 0.894. MLP performed best using the train-test split method, yielding an AUC of 0.896.



Gene expression-based classifiers

Train-test split was the most effective internal validation method for SVM, RF and MLP, yielding a perfect AUC of 1.000 for both SVM and RF, and an AUC of 0.999 for MLP.



CONCLUSIONS

While external validation remains the gold standard for confirming clinical utility, it isn't always feasible. By directly comparing internal validation methods, this study offers practical guidance on selecting approaches that increase the chance of developing models that generalise well to unseen data, strengthening the clinical potential of cancer detection models based on molecular biomarkers. A key finding of our study was that no single internal validation method was optimal across classifiers and data types and the choice of internal validation method depends both on the underlying data and the model used.

REFERENCES

- Modhukur V, Sharma S, Mondal M, Lawarde A, Kask K, Sharma R, et al. Machine learning approaches to classify primary and metastatic cancers using tissue of origin-based dna methylation profiles. *Cancers (Basel)*. 2021 Aug 1;13(15).
- Mahin KF, Robiuddin Md, Islam M, Ashraf S, Yeasmin F, Shatabda S. PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. *Genomics*. 2022 Mar;114(2):110264.
- Gilham C, Nedjai B, Scibior-Bentkowska D, Reuter C, Banwait R, Brentnall AR, et al. Long-term prediction by DNA methylation of high-grade cervical intraepithelial neoplasia: Results of the ARTISTIC cohort. *Int J Cancer*. 2024 Jul 1;155(1):81–92.
- Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016 Jan;69:245–7.