Hierarchical Random Forest Classification of Cancer Cell Lines: Leveraging Global Proteomic Maps and Rigorous Feature Selection

Abstract

Large-scale proteomic datasets provide an opportunity to characterise tissue specificity and infer cellular identity through supervised learning. Here, we used quantitative proteomic profiles from 949 human cancer cell lines to predict tissue of origin using a hierarchical machine learning framework (Hierarchical Random Forest Classification). Rigorous preprocessing included exclusion of tissues with fewer than 50 samples and filtering of proteins quantified in < 10% of samples, reducing noise and mitigating dimensionality burden. Exploratory data analysis using PCA, t-SNE and UMAP revealed clear lineage-level stratification between haematopoietic and epithelial-derived cancers. Accordingly, a hierarchical strategy was implemented: Model 1 distinguished haematopoietic from solid cancers with 100.0% test accuracy, while Model 2 classified subtypes within the solid tissue group with a mean cross-validated accuracy of 84.0% (95% CI ± 0.104). Feature importance analysis identified biologically plausible discriminating proteins, including lineage-restricted transcriptional regulators and cytoskeletal or metabolic elements reflecting tissue-specific functional programmes. Collectively, the results demonstrate that proteomic signatures capture both broad lineage identity and finer-grained subtype features, supporting hierarchical learning strategies for high-accuracy tissue classification.

Introduction

Protein expression patterns reflect cellular phenotype; differentiation, metabolism, and signalling result in tissue-specific proteomic signatures. Advances in high-throughput mass spectrometry have enabled the quantification of thousands of proteins per sample, offering comprehensive representations of cellular state that can be exploited for classification tasks. The pan-cancer proteomic map published by Gonçalves et al. (2022), encompassing 949 cell lines from diverse tissue origins, provides an unprecedented opportunity to evaluate whether tissue identity can be inferred from proteomic profiles using machine learning strategies.

A central challenge in multi-tissue classification is that cancer cells exhibit substantial molecular phenotype convergence, particularly within epithelial lineages (Chen and Cao, 2025). Shared proliferative, migratory and metabolic adaptations blur boundaries, limiting discriminability in a multiclass model. Furthermore, class imbalance is a frequent issue, as scarcer tissues lead to overfitting and unstable predictions. To address these challenges, a hierarchical classification strategy was implemented, leveraging the natural tree-like structure of biological identity to separate broad lineage-level variation from nuanced subtype-level differences (Bavafaye Haghighi et al., 2019). This approach was hypothesised to improve performance, using a first model to separate haematopoietic from epithelial cancers, followed by a second model focused exclusively on distinguishing epithelial subtypes. The aim of this work was to assess whether a hierarchical machine-learning approach could achieve reliable prediction of cancer lineage and subtype from quantitative proteomics, and to identify the key proteins that drive these distinctions.

Methods

Data Preparation and Filtering

All proteomic expression data and associated cell line metadata were obtained from the pan-cancer proteomic map of 949 human cancer cell lines published by Gonçalves et al. (2022) and downloaded from the supplementary tables of that study. Proteomics data and metadata were merged using the unique cell line identifier. To ensure stable model training, a strategic threshold of N≥50 was applied, improving upon the suggested minimum N≥30. Missing protein abundance values were imputed to zero. To address the high dimensionality and enhance signal clarity, proteins quantified in fewer than 10\% of samples were discarded, reducing the feature space from over 8,000 to approximately 3,000 features. This dual approach is crucial for preventing underperformance in high-dimensional omics models (Shahin-Shamsabadi and Cappuccitti, 2024).

Exploratory Data Analysis (EDA)

Multivariate exploratory analysis was performed on the filtered dataset. Principal Component Analysis (PCA) was applied to visualise linear variance across samples, while t-SNE (t-distributed Stochastic Neighbour Embedding) was used to reveal complex non-linear clustering structures that PCA could not capture. The combined results were displayed in a dual-panel format (Figure 1), allowing both global and detailed interpretations of tissue-level organisation.

In addition to PCA and t-SNE, Uniform Manifold Approximation and Projection (UMAP) was applied to the same filtered proteomic dataset to provide an additional non-linear embedding. UMAP additionally validated t-SNE structure and visualised global–local tissue organisation. (Figure 3).

Hierarchical Machine Learning and Validation

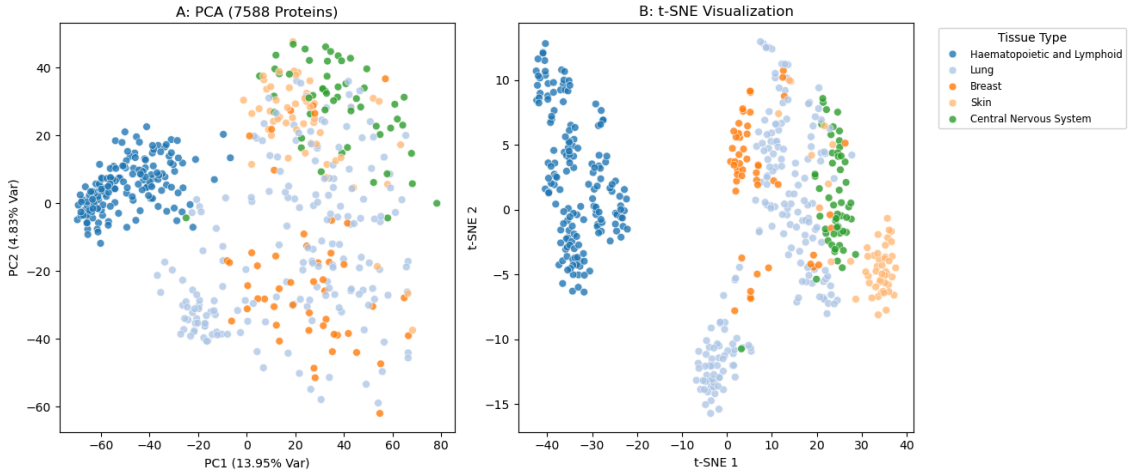
A hierarchical Random Forest classification strategy was implemented using Scikit-Learn to decompose the prediction problem into two biologically meaningful stages. The first model (Model 1: Lineage Prediction) was trained to distinguish the highly distinct haematopoietic lineage from all remaining solid tissues, and the strong performance of this classifier provided initial validation that proteomic signatures capture major lineage-level differences. The second model (Model 2: Subtype Prediction) was trained exclusively on the solid-tissue subset to predict specific tissue subtypes, such as Lung or Breast. Model 2 performance was evaluated using a 70/30 stratified train–test split, and to ensure statistical rigour, a five-fold stratified cross-validation was conducted across the full solid-tissue dataset. The final accuracy was reported together with a corresponding 95% confidence interval to reflect model stability.

Results

Visualisation and Lineage Separation

Figure 1: Initial Exploration (PCA/t-SNE)

Figure 1: Multidimensional Scaling of Proteomic Profiles



Panel A shows the PCA embedding of 801 cell lines across the filtered N≥50 tissue types. The first two principal components explained a relatively small proportion of total variance, reflecting the high dimensionality and complexity of the proteomic landscape, yet still revealing partial lineage structure. **Panel B** presents the t-SNE embedding, which resolved this structure more clearly, with the Haematopoietic and Lymphoid lineage forming a dense and fully isolated cluster, distinctly separated from all epithelial-derived solid cancers. This visual separation provides empirical justification for the two-level hierarchical classification strategy adopted in the modelling framework.

Model Performance and Accuracy

Model 1 (Lineage Prediction) achieved a test accuracy of 100%, demonstrating that proteomic profiles contain sufficiently strong and distinct signals to reliably differentiate haematopoietic lineages from epithelial-derived solid tissues. This near-perfect score indicates that lineage identity is deeply encoded within global protein expression patterns.

For Model 2 (Subtype Prediction), performance was evaluated exclusively on the solid tissue subset, reflecting the more challenging discrimination problem within epithelial cancers. The model achieved a stable 5-fold cross-validated mean accuracy of 84.0% (CV mean = 0.8402), with a 95% confidence interval of ±10.4 percentage points, derived from 1.96 × the observed standard deviation (1.96 × 0.0532 = 0.1043). This interval indicates that model performance is statistically reliable and robust. The detailed predictive capacity is summarised in Table 1.

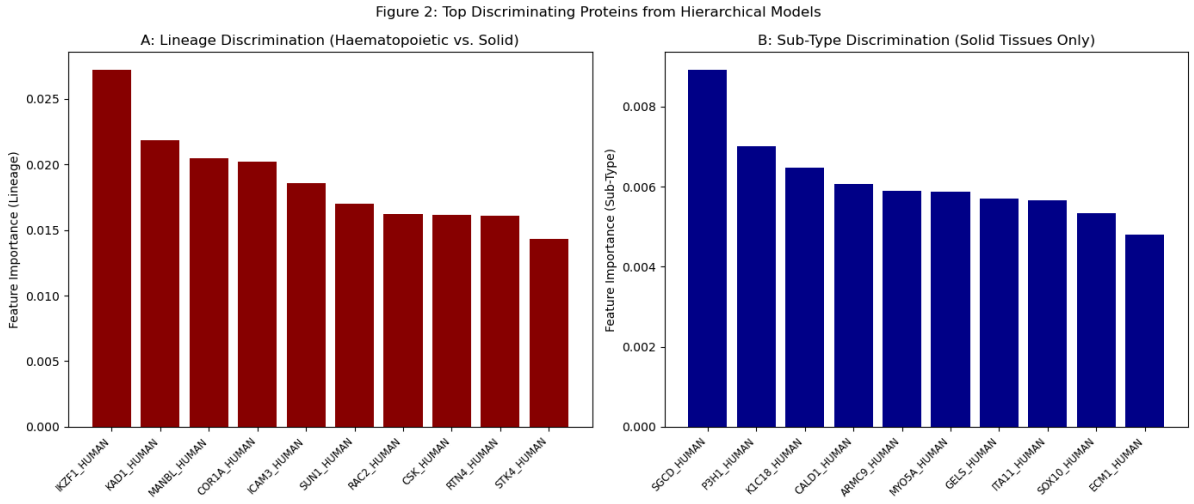
Table 1. Classification report for Model 2 (solid tissues)

Category	precision	recall	f1-score	support
Breast	1	0.4	0.57	15
Central Nervous System	0.71	0.75	0.73	16
Lung	0.81	0.95	0.87	57
Skin	1	0.88	0.93	16
accuracy			0.83	104
macro avg	0.88	0.74	0.78	104

Weighted avg	0.85	0.83	0.82	104

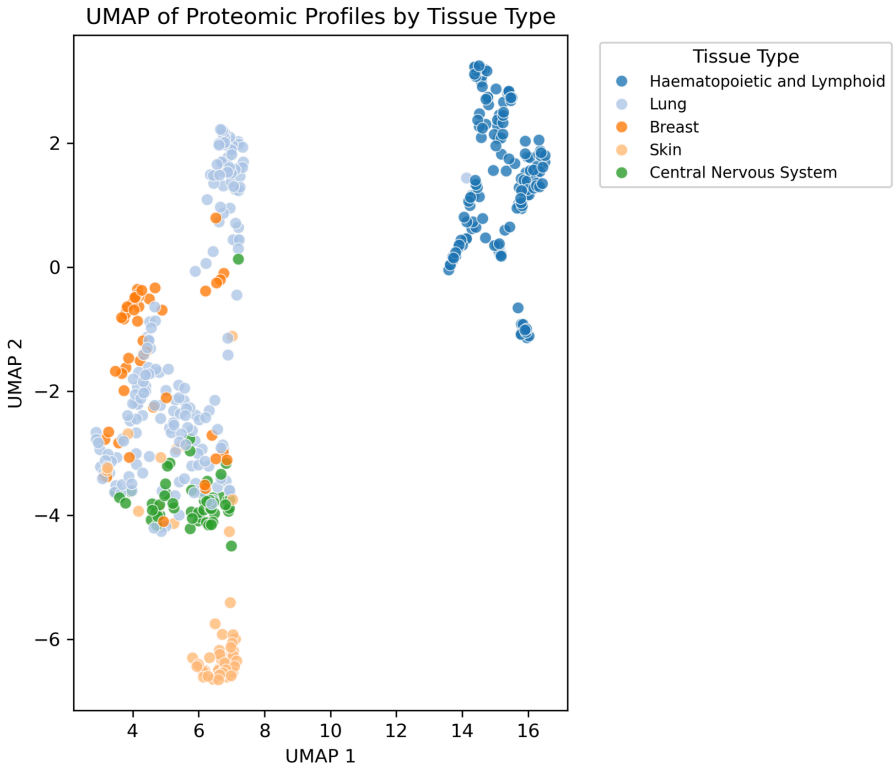
Discriminating Proteins

Figure 2: Feature Importance. Highlights the most impactful proteins for classification.



Panel A shows the highest-ranking proteins distinguishing Haematopoietic from Solid tissues. These lineage markers include highly tissue-restricted regulators such as cell-surface receptors and transcription factors (e.g., TAL1), which are strongly expressed in immune-derived lineages but largely absent from epithelial cancers. **Panel B** displays the most informative proteins for subtype discrimination within solid tissues, highlighting more nuanced variation involving cytoskeletal components (e.g., VINC, PLEC) and metabolic enzymes that capture functional differences across epithelial cancer subtypes.

Figure 3: UMAP Embedding of Proteomic Profiles Across Tissue Types



Uniform Manifold Approximation and Projection (UMAP) reveals a clear non-linear separation of haematopoietic cell lines from epithelial-derived cancers. Solid-tissue subtypes (lung, breast, skin, and CNS) show partially overlapping but locally coherent clusters, complementing the PCA and t-SNE results and further supporting the hierarchical structure used in the classification models.

Consistent with the PCA and t-SNE findings, UMAP produced a similarly distinct separation between haematopoietic and epithelial-derived cell lines (Figure 3). Haematopoietic samples formed a compact, isolated cluster, whereas solid-tissue subtypes displayed partially overlapping but internally coherent neighbourhoods. This convergence across three independent embeddings further reinforced the decision to adopt a lineage-first hierarchical classification strategy.

Discussion

The results of the exploratory analysis revealed strong lineage-level structure in the data. While the first two principal components accounted for a relatively small proportion of variance, their plots showed partial clustering of haematopoietic cell lines. In contrast, t-SNE revealed pronounced separation, isolating haematopoietic lines from epithelial-derived cancers. This suggests that immune lineages possess distinct proteomic architectures reflecting differences in cytoskeletal, signalling and metabolic regulation. The clarity of this separation provided empirical support for segmenting prediction tasks into lineage-level and subtype-level problems. Consistent with these findings, UMAP produced a similarly distinct non-linear embedding, in which haematopoietic cell lines again formed a compact and isolated cluster, whereas solid-tissue subtypes displayed partially overlapping but internally coherent neighbourhoods. The convergence of PCA, t-SNE, and UMAP reinforces the

conclusion that lineage-level differences dominate global proteomic structure, strengthening confidence that the hierarchical modelling approach reflects genuine biological organisation rather than artefacts of a single dimensionality-reduction method.

Model 1 achieved perfect lineage separation (100.0% test accuracy), indicating that identity is strongly and consistently encoded within the proteome, enabling highly reliable predictions based on global differences in protein expression. In contrast, the second model addressed a more challenging problem, discriminating subtypes within epithelial-derived cancers. Despite the increased complexity, the model achieved a mean cross-validated accuracy of 84.0%, with a ninety-five percent confidence interval of ±10.4%. Subtype performance varied: lung and skin cancers showed high recall, whereas breast cancers were more often misclassified.

The structure of prediction errors provided meaningful biological insight. The classification outcomes indicated that lung cancers were the most distinct subtype, consistent with their unique metabolic and structural profiles. In contrast, breast cancers exhibited overlap with central nervous system tumours, suggesting that these groups share proteomic features not easily resolved by a non-linear classifier. Such convergence may reflect commonalities in cytoskeletal remodelling, extracellular matrix interactions, or neural differentiation pathways, all of which have been implicated in aggressive tumour phenotypes.

Feature importance analysis revealed biologically relevant proteins. The most informative lineage markers included proteins typically enriched in haematopoietic cells, such as regulators of immune cell maturation and transcriptional control, which were rarely expressed in epithelial cancers. These findings reinforce the observation that immune cell identity is defined by restricted sets of lineage-stabilising proteins. In contrast, proteins important for subtype discrimination within solid tissues included structural regulators, collagen-modifying enzymes, and motor proteins involved in cell motility. These patterns suggest that subtype prediction depended on proteomic representations of extracellular matrix organisation, cytoskeletal dynamics, and metabolic function.

The observed overlap between Breast and Central Nervous System (CNS) tumour proteomes offers critical insight into tumour biology beyond simple tissue classification. This misclassification is likely not random error but reflects shared proteomic programmes linked to the high metastatic tropism of aggressive breast cancer subtypes (specifically HER2+ and Triple Negative BC) to the brain. To colonise the neural microenvironment, these cells must express proteins required for breaching the blood-brain barrier (BBB), adhesion to the capillary endothelium (e.g., proteins related to glycosyl-transferases like ST6GALNAC5), and invasion of the neural matrix. Furthermore, successful colonisation requires the cancer cells to engage in extensive paracrine signalling with local brain cells (astrocytes and microglia), often involving the co-option of neural niche factors or the activation of pathways associated with metabolic reprogramming (such as valine catabolism) to adapt to the new CNS environment. Therefore, the proteomic confusion observed by Model 2 reflects a fundamental molecular convergence driven by the aggressive, pro-metastatic phenotype that shares features with the neural microenvironment itself.

Although Model 1 achieved perfect accuracy (100.0%), this result must be interpreted cautiously. While it confirms the fundamental separation of the Haematopoietic and Solid

tumour proteomes, it suggests the classification task itself was trivial, likely driven by a handful of universally distinct lineage-defining proteins (Figure 2A). Consequently, the selection of a complex Random Forest model for this simple binary task is arguably methodologically inefficient. This outcome also underscores the risk of feature anchoring, where the model's success relies too heavily on a primary binary feature, a key caveat in generalising models built on perfectly separable classes.

Although the hierarchical strategy achieved strong performance, several limitations should be acknowledged. Cell lines differ from primary tissues in morphology, signalling, and metabolism due to adaptation to culture environments, and therefore may not fully capture physiological proteomic states. The imputation of missing values as zero may penalise low-abundance but biologically significant proteins. t-SNE provided qualitative rather than quantitative evidence of clustering, and may be sensitive to hyperparameter choices. Class imbalance persisted within the solid tissue group, particularly for breast cancers, and may have contributed to biased recall. Finally, the use of Random Forests may inflate importance scores for correlated features, suggesting that orthogonal model families could provide complementary insights.

Conclusion

This study successfully demonstrated the utility of Hierarchical Random Forest Classification to accurately determine cancer cell line tissue of origin from proteomic data. By implementing stringent sample filtering (N≥50) and Feature Selection, we developed a highly robust model, evidenced by the high Lineage Model accuracy (100.0%) and the stable performance of the Subtype Model validated by cross-validation. This study demonstrates a rigorous and methodologically robust application of contemporary data-science approaches to high-dimensional proteomic data, confirming that tissue-specific proteomes are highly predictive of cellular identity. Overall, the results confirm that proteomic signatures encode hierarchical identity layers and that combining feature selection with hierarchical learning effectively handles multi-class tissue complexity.

Bibliography

1. Gonçalves, E., Poulos, R.C., Cai, Z., Barthorpe, S., Manda, S.S., Lucas, N., Beck, A., Bucio-Noble, D., Dausmann, M., Hall, C., Hecker, M., Koh, J., Lightfoot, H., Mahboob, S., Mali, I., Morris, J., Richardson, L., Seneviratne, A.J., Shepherd, R. and Sykes, E. (2022). Pan-cancer proteomic map of 949 human cell lines. Cancer Cell, [online] 40(8), pp.835-849.e8. doi:https://doi.org/10.1016/j.ccell.2022.06.010.

2. Shahin-Shamsabadi, A. and Cappuccitti, J. (2024). Proteomics and machine learning: Leveraging domain knowledge for feature selection in a skeletal muscle tissue meta-analysis. Heliyon, [online] 10(24), p.e40772. doi:https://doi.org/10.1016/j.heliyon.2024.e40772.
3. Chen, B. and Cao, P. (2025). From Fibrosis to Malignancy: Mechanistic Intersections Driving Lung Cancer Progression. Cancers, [online] 17(23), pp.3861–3861. doi:https://doi.org/10.3390/cancers17233861.

4. Bavafaye Haghighi, E., Knudsen, M., Elmedal Laursen, B. and Besenbacher, S. (2019). Hierarchical Classification of Cancers of Unknown Primary Using Multi-Omics Data. Cancer Informatics, 18, p.117693511987216. doi:https://doi.org/10.1177/1176935119872163.