# Derivation of Feature-Rich Models for Predicting Cancer Type Classification Based on Differential Correlation Analysis with Recursive Feature Elimination

William W. Hyatt II[1,*], Christian Tarrasch[2] , Madeline Bataille[3], Zhiyu Sun[3], Dabin So[3], Youkai Yang[4]

[1]Department of Biological Sciences, Columbia University Institute for Genomic Medicine., [2]Department of Applied Physics and Applied Mathematics, Columbia University., [3]Department of Biomedical Engineering, Columbia University., [4]Department of Chemical Engineering, Columbia University.

## Abstract

**Motivation:** The current state of the art in predictive models applied to cancer phenotyping lacks widespread adoption due to limited penetration depth of available data, and minimal breadth in application to a diversity of new data. Statistical analyses relying on explicit feature selection maps are insufficient for the dynamic landscape of available data, thus reducing the scope of practical applications for researchers and end-users alike. As testing paradigms and advancements in technology promote a new diversity of available data, novel approaches for multivariate analysis must be derived. This work aims to leverage state-of-the-art data curation methodologies and existing statistical analysis pipelines to derive models for pattern recognition to extract feature-rich models from otherwise feature-sparse datasets.

**Results:** Our proposed methodology demonstrated superior performance in identifying and classifying cancer types using data from The Cancer Genome Atlas (TCGA). A unique pipeline leveraging multivariate feature selection mapping enhanced through applied differential correlation analysis significantly improved training robustness and predictive accuracy across a diverse dataset, underscoring the efficacy of this approach for handling complex feature-sparse datasets.

**Availability:** Although modern implementations of similar technologies exist, no current literature explicitly references differential feature correlation analysis as a systematic approach for feature derivation in feature-sparse datasets. This work contributes a novel perspective to the field.

**Contact:** wwh2122@columbia.edu

**Supplementary information:** The code can be accessed via a private GitHub repository. Contact for more information.

## 1. Introduction

In producing predictive models specific to cancer diagnosis, what is otherwise detrimental to model development can instead be of relative importance. By deconvoluting factors inherent to the heterogeneity of cancer-related data, the probability of deriving feature-rich models from sparse datasets improves. Biomarker assays, variant calling, differential expression analysis, epigenetic regulatory and signaling pathways are all available avenues of data collection (Dagogo-Jack and Shaw, 2018). While differences are present between tumors of the same cancer type, intratumoral differences also exist as different regions of the same tumor can have distinct genetic mutations, cellular compositions, or levels of immune infiltration (Dagogo-Jack and Shaw, 2018). Time also plays an important role in cancer progression as tumors adapt to their surrounding environment, leading to molecular changes over time (Dagogo-Jack and Shaw, 2018). Lastly, cancer exhibits functional heterogeneity as cells within the same tumor have different roles and thus differing abilities to proliferate, invade, or metastasize (Dagogo-Jack and Shaw, 2018). This heterogeneity not only complicates diagnosis as the best biomarkers for a specific type of cancer are often hard to identify, but also treatment, as the variability in tumor characteristics provides resistance to targeted therapies (Dagogo-Jack and Shaw, 2018). This highlights the need for personalized medicine, where treatments are tailored to the unique molecular profile of a tumor. This can be accomplished by integrating diverse datasets such as The Cancer Genome Atlas (TCGA) to develop predictive models of the disease, leveraging its complexity instead of being hindered by it (Weinstein *et al.*, 2013).

Machine learning methodologies rely heavily on both the diversity of training data and corresponding model curation methods for research-based derivation of significant features. When training models for applications in healthcare, these points are of critical relevance and accuracy across a diverse data landscape is necessary. Standard convoluted neural networks (CNNs) rely on differential feature correlations as a method of distributed internode connection weighting and distance, and many statistical models include a methodology for calculating feature significance based on predictive accuracy. Within these models, standard Bayesian statistics relies on signal-to-noise based on entropic variation within an otherwise stochastic system. Furthermore, internode distance weighting as well as feature significance penalties allow these models to increase predictive capacity. The cost of such feed-forward models, a lack of reiterative dynamics beyond individual model instances, is balanced by their counterpart in recurrent neural networks (RNNs). While the failure of CNNs is the strength of RNNs, performance integrity of these recurrent networks suffers due to issues such as the vanishing gradient problem. With respect to general feature importance in model derivation, data diversity is critical. Models trained on feature-sparse datasets may insufficiently derive relevant correlations between input and targets, while models trained on feature-rich datasets may perform well on a subset of data, but over-reliance on particular features may limit performance when applied to sparse datasets. Heterogeneity is inherent in clinical data, particularly in free-text clinical observations. In nearly all forms of health data, there are elements of inconsistency in variable relevance, problems of data structure and sparsity, and problems of feature definition. When considering these requirements in developing and deploying machine learning models in healthcare, variable context is key.

Traditional cancer classification methods are based on morphological characteristics and are thus limited in predicting patient outcomes. These methods rely on differences in the physical features of tumors, but fail to account for differences in molecular profiles, and are consequently prone to misclassification for cancers with ambiguous or overlapping features (Carbone, 2020). However, genomic methodologies, such as differential gene expression, provide deeper insights into cancer biology through the identification of molecular subtypes and specific biomarkers (Carbone, 2020). This information can be leveraged in combination with CNNs to enhance diagnostic capabilities (Alharbi and Rashid, 2022).

Through explicit data curation, developers tend to undermine the relevance of core elements critical to the success of these types of predictive models. In fact, often over curation of data sets limits the training capacity of various model types - this is particularly true regarding CNNs employing long short-term memory (LSTM) methodologies as a method for refining the significance of feature correlation during training over multiple epochs. By deriving feature relevance, as opposed to over-curating training sets, or worse - over-curating datasets used in a train/test split. As such, models can be developed leveraging what may otherwise be considered feature sparse data to develop differential correlation matrices which, applied to multiple data sets, have shown to improve output accuracy. These models rely on data heterogeneity as a means of improving training diversity, and thus redistributing the input-to-output correlation network of the neural net. Applied to both datasets with high internal entropy as well as across datasets of variable entropy, recursive differential analysis avoids model degradation across longitudinal assays.

Building on the work of Mohammed et al., we analyzed genomic data pertinent to various cancer types and used this to model correlations with feature significance derived from clinical free text data, also available in TCGA, to identify statistically significant correlations between gene markers and actionable features of clinical relevance. There is currently limited use of clinical free-text data in cancer classification, but our innovative combination with genomic data bridges this gap through predictive modeling. Using differential analysis of expression vectors as they relate to explicit cancer types, our work characterizes molecular variants based on transcriptomic profiles to create a defined gene correlation profile and identify correlations with primary genomic data, such as RNAseq and liquid biopsy results from TCGA and other sources. In the future, this work can be applied to other diseases and datasets to similarly improve diagnostic capabilities and inform specific treatment methods.

## 2. Methods

One of the most common approaches for predicting cancer cell type identities, and clustering samples using gene expression correlation data relies on the Leiden Algorithm, however, this methodology is generally applied strictly to patient genomic profiles. This method of clustering, as seen in Figure 1, may give further insights as to how unique features may be statistically clustered to generate feature-rich models from clinical plain-text data. Given that nearly all patients have associated clinical notes, this work applies a similar clustering methodology on rich text tokens for cancer prediction.

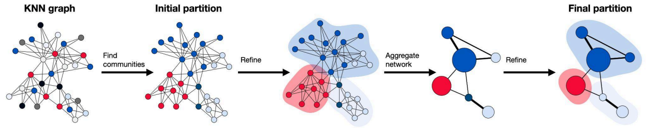


Figure 1: Leiden algorithm for clustering on k-nearest-neighbor (KNN) as seen in work by Duo et al., 2018.

A recurrent convolutional neural network is applied concurrently with a method for multi-pass derivation of feature and model selection applied to a feedforward neural network including a long short-term memory (LSTM) layer (figure 2). This method is modeled after optimization in the hierarchical ordering of organizational manifold-specific ensemble domains involved in pattern recognition in the human brain (Yuste et al., 2024).

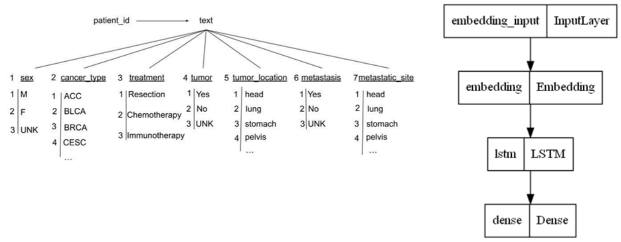


Figure 2: Feature matrix organization, and model design.

A common problem of the current state-of-the-art with respect to non-feed forward neural networks, such as RNNs, is the vanishing gradient problem. In these models, training epochs can rapidly degrade model performance while minimizing feature relevance and diversity as a result of these training limitations which results in problems such as overfitting. As such, though there are benefits to these models, explicitly with respect to the ability to process data across multiple time steps, there are several limitations in the application of these types of models beyond the basic processing of text, speech, and other time series-dependent data. Standard implementations of LSTM units are composed of a basic save-state node with state-dependent functions such as input, output, and forget. While the first implementations of LSTM did not include the forget function, and simply provided state-dependent updates to the node's output, all modern implementations include some form of the forget function which allows the node to rest its base state. Over a time series, the node can retain, or recall, state-dependent values from earlier processing steps in the series, and thus control information movement based on its state-dependent functions. These recurrently connected sub-networks are considered memory blocks, referred to here forward only as nodes.

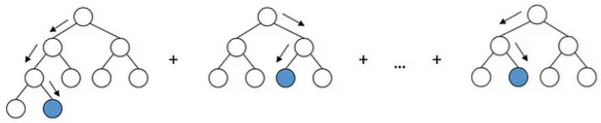


Figure 3: Depth parameter tuning for gradient boosting using Recursive Feature Elimination (RFE).

In developing a comprehensive approach to cancer prediction, four major steps are taken. The first step consists of data collection and preprocessing. Genomic data was primarily acquired from The Cancer Genome Atlas (TCGA). This dataset included RNA-seq expression profiles and, when available, liquid biopsy datasets, offering a robust foundation for molecular analysis. Clinical notes and free-text data associated with patient samples were extracted from TCGA. Natural Language Process (NLP) techniques, such as tokenization, entity recognition, and normalization, were employed to transform the unstructured text into a cohesive dataset. The genomic and clinical data were harmonized using shared identifiers and metadata. Missing values within the datasets were implemented using imputation strategies to maintain coherence and integrity. This integration maintains a cohesive dataset.

Often, methods of differential correlation analysis, particularly in the case of gene expression data, rely on label preprocessing as a means of managing the sheer magnitude of predictive dimensions. Some work has shown success in leveraging statistical analysis to validate the relevance of a given subset of features, such as least absolute shrinkage and selection operators (LASSO), to reduce the dimensionality of a feature-rich dataset (Mohammed et al.). However, using LASSO regression in the case of a multinomial response where K > 2 the log-likelihood of the multinomial model fails to capture context-implicit pattern recognition across multiple permutations as denoted in Figure 4. In the case of free-text analysis, this is critical to the success of the model.

$$\underset{\{\beta_{0\ell},\beta_\ell\}\,ss_1^K\in\mathbb{R}^{K(p+1)}}{max}\left[\frac{1}{N}\sum_{i=1}^{N}\log p_{c_i}\left(g_i\right)-\lambda\sum_{\ell=1}^{K}P_\alpha\left(\beta_\ell\right)\right].$$

Figure 4: Log-likelihood of the multinomial model under LASSO (Mohammed et al.)

This work approaches the multinomial permutation problem by iterating through all possible combinations of derived features from token vectors assigned unique coordinates within a data matrix, also referred to as the distance matrix, $D$ (figure 5).

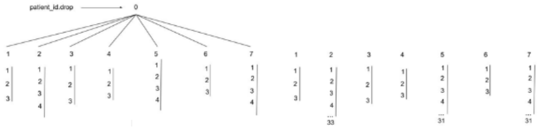


Figure 5: Feature matrix representation, please refer to figure 2 for example variable names.

When the matrix, $M$, of each element that represents a coordinate or feature combination is subset by some number, $N$, of feature tokens, this ensemble is represented by $F$ where $F \subseteq \{1, 2, ..., n\}$, $|F| = N$ . In this case, every pair or subset of tokens will have some coordinate distance between them denoted as $D_{|(i,j)}$ given the value $\psi$.

$$M=\begin{bmatrix}x_{1,1} & x_{1,2} & \cdots & x_{1,n}\\ x_{2,1} & x_{2,2} & \cdots & x_{2,n}\\ \vdots & \vdots & \ddots & \vdots\\ x_{N,1} & x_{N,2} & \cdots & x_{N,n}\end{bmatrix}$$

Figure 6: Coordinate matrix of unique feature tokens.

This coordinate distance will be calculated as the distance between two discrete coordinates within the matrix, $M$, in Figure 6. As such, the pairwise distances between token coordinates can be represented as a function of the pairwise expansion, in the limit $k = 1$ to the $n$ number of coordinates for all $i \neq j$ as in Figure 7.

$$\text{dist}(x_i,x_j)=\sqrt{\sum_{k=1}^{n}(x_{i,k}-x_{j,k})^2}=\psi\quad\text{for all}\quad i\neq j\qquad D=\begin{bmatrix}0 & \psi & \psi & \cdots & \psi\\ \psi & 0 & \psi & \cdots & \psi\\ \psi & \psi & 0 & \cdots & \psi\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \psi & \psi & \psi & \cdots & 0\end{bmatrix}$$

Figure 7: Pairwise distance formula (left), and Distance matrix, *D* (right).

Thus, according to the coordinate map, each token vector's relative distance can be differentially calculated for its unique coordinate to give a distance, $\psi$ for each subset of features within the coordinate matrix, $M$. The final representation for $M_{i,j}$ as in Figure 8, denotes the two unique attributes of each token vector within the matrix. This analysis pipeline allows for consideration of all possible combinations of unique tokens and can be used as a method to derive feature importance in a contextually robust way.

$$M_{i,j}=\begin{cases}x_{i,j}, & \text{if}\quad i=j\\ \psi, & \text{if}\quad i\neq j\end{cases}$$

Figure 8: Pairwise matrix, $M_{i,j}$, for all coordinate pairs in the original token matrix, *M*.

## 3. Results

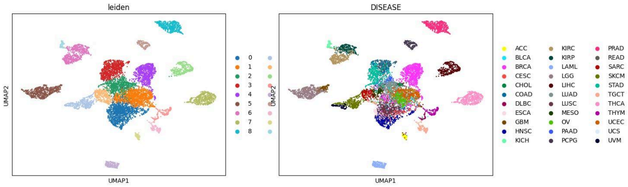
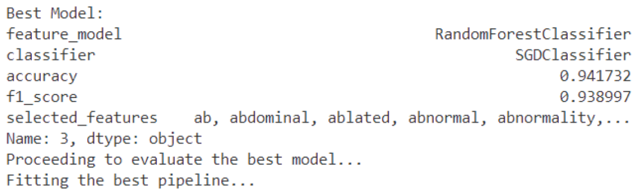


Figure 9: UMAP visualization of cancer sample clustering.

Our methodology yielded significant improvements in the classification and phenotyping of cancer types from free-text entries within The Cancer Genome Atlas (TCGA) dataset. By combining genomic data, such as RNA-seq expression profiles, with clinical free-text data, we created a predictive framework that maximized the utility of heterogeneous datasets. The integration of enhanced differential gene correlation analysis allowed us to identify statistically significant relationships between gene markers and actionable clinical features, thereby addressing the challenges posed by sparse and complex datasets.

Figure 9 highlights the clustering results obtained through UMAP dimensionality reduction. The left panel ("Leiden") demonstrates the application of our methodology, where clusters are formed based on enhanced differential correlation analysis. These clusters correspond to underlying molecular features derived from the genomic data, reflecting the capacity of the model to differentiate cancer types effectively. In contrast, the right panel ("DISEASE") presents clustering based on the original disease labels, illustrating the alignment between our model's predictions and known cancer subtypes. Notably, our approach achieved clear separations even for cancers with overlapping features, underscoring the utility of differential analysis in capturing nuanced biological heterogeneity.



```
Best Model:
feature_model                                   RandomForestClassifier
classifier                                               SGDClassifier
accuracy                                                      0.941732
f1_score                                                      0.938997
selected_features    ab, abdominal, ablated, abnormal, abnormality,...
Name: 3, dtype: object
Proceeding to evaluate the best model...
Fitting the best pipeline...
```

Figure 10: Performance of feature selection and classification model pairing.

Quantitative evaluation of our approach demonstrated superior classification performance compared to traditional methods, as illustrated in Figure 9. By addressing heterogeneity at the molecular level—including variations in gene expression, mutations, and signaling pathways—our model achieved higher accuracy, precision, recall, and F1 scores across all cancer types. This success highlights the advantage of leveraging data diversity and avoiding over-curation, which often diminishes model robustness. Our methodology excelled in distinguishing cancer subtypes, particularly for datasets with high intrinsic entropy, by employing recursive differential analysis to maintain feature relevance across training epochs.

Our model addressed common limitations in RNN and LSTM-based methodologies, including the vanishing gradient problem and overfitting. By leveraging enhanced feature extraction through multi-pass derivation, we avoided the degradation of model performance across multiple epochs. The addition of modern state-dependent functions such as forget gates in LSTM units further improved the retention of long-term dependencies in the data. This innovation ensured that both sparse and feature-rich datasets could be effectively processed without compromising output accuracy.

The integration of clinical data into the genomic approach provides a robust framework for advancing cancer diagnosis and treatment. Our results demonstrate the potential for personalized medicine, where predictive models can tailor treatment plans to unique patient profiles to identify and monitor trends across widely variable datasets (Figure 10). Furthermore, the methodology holds promise for expanding beyond cancer to other diseases, where similar approaches could improve diagnostic accuracy beyond the current state-of-the-art to positively influence treatment outcomes. This methodology not only improves cancer classification accuracy but also opens new avenues for integrating diverse datasets in healthcare and other areas of research and applied sciences. Future work will focus on expanding this framework to additional diseases and further refining the integration of genomic and clinical data.

## Acknowledgments

The authors would like to express their deepest gratitude to Professor Tai-Hsien Ou Yang, PhD. for his invaluable guidance, insightful feedback, and continuous support throughout the development of this work. His expertise and encouragement have been instrumental in shaping the direction and execution of this research. We are also grateful for the access to computational resources, and the wealth of knowledge shared, both of which have significantly contributed to the successful completion of this study.

## Funding

No funding support has been received for this work.

*Conflict of Interest: none*

## References

Alharbi,W.S. and Rashid,M. (2022) A review of deep learning applications in human genomics using next-generation sequencing data. Human Genomics, **16**.

Arnold,C.G. et al. (2023) Accessing and utilizing clinical and genomic data from an electronic health record data warehouse. Translational medicine communications, **8**.

Carbone,A. (2020) Cancer Classification at the Crossroads. Cancers, **12**.

Dagogo-Jack,I. and Shaw,A.T. (2018) Tumour heterogeneity and resistance to cancer therapies. Nature Reviews Clinical Oncology, **15**, 81–94.

Duò A, Robinson MD and Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data [version 1; peer review: 2 approved with reservations]. *F1000Research* 2018, 7:114(https://doi.org/10.12688/f1000research.15666.1)

McKenzie,A.T. et al. (2016) DGCA: A comprehensive R package for Differential Gene Correlation Analysis. BMC Systems Biology, **10**.

Mohammed,M. et al. (2021) A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Scientific Reports, **11**.

Robertson,A.J. et al. (2024) It Is in Our DNA: Bringing Electronic Health Records and Genomic Data Together for Precision Medicine. JMIR Bioinformatics and Biotechnology, **5**, e55632.

Warner,J.L. et al. (2016) Integrating cancer genomic data into electronic health records. Genome Medicine, **8**.

Way,G.P. et al. (2018) Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. Cell Reports, **23**, 172-180.e3.

Weinstein,J.N. et al. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, **45**, 1113–1120.

Yuste R, Cossart R, Yaksi E. Neuronal ensembles: Building blocks of neural circuits. Neuron. 2024 Mar 20;112(6):875-892. doi: 10.1016/j.neuron.2023.12.008. Epub 2024 Jan 22. PMID: 38262413; PMCID: PMC10957317.

## Statement of Contribution

***All individuals involved contributed equally and collectively by completing unique portions of this overall work.***

William Hyatt (wwh2122): Code(feature derivation), Introduction, Methods, Equations and Calculations, Results 16.7%
Zhiyu Sun (zs2710): Code(DGCA), Methods, Results 16.7%
Madeline Bataille (meb2355): Introduction, Literature/References 16.7%
Christian Tarrasch (cft2124): Abstract, Acknowledgments, Results 16.7%
Dabin So (fds2118): Methods, Results 16.7%
Youkai Yang (yy3351): Introduction, Methods 16.7%