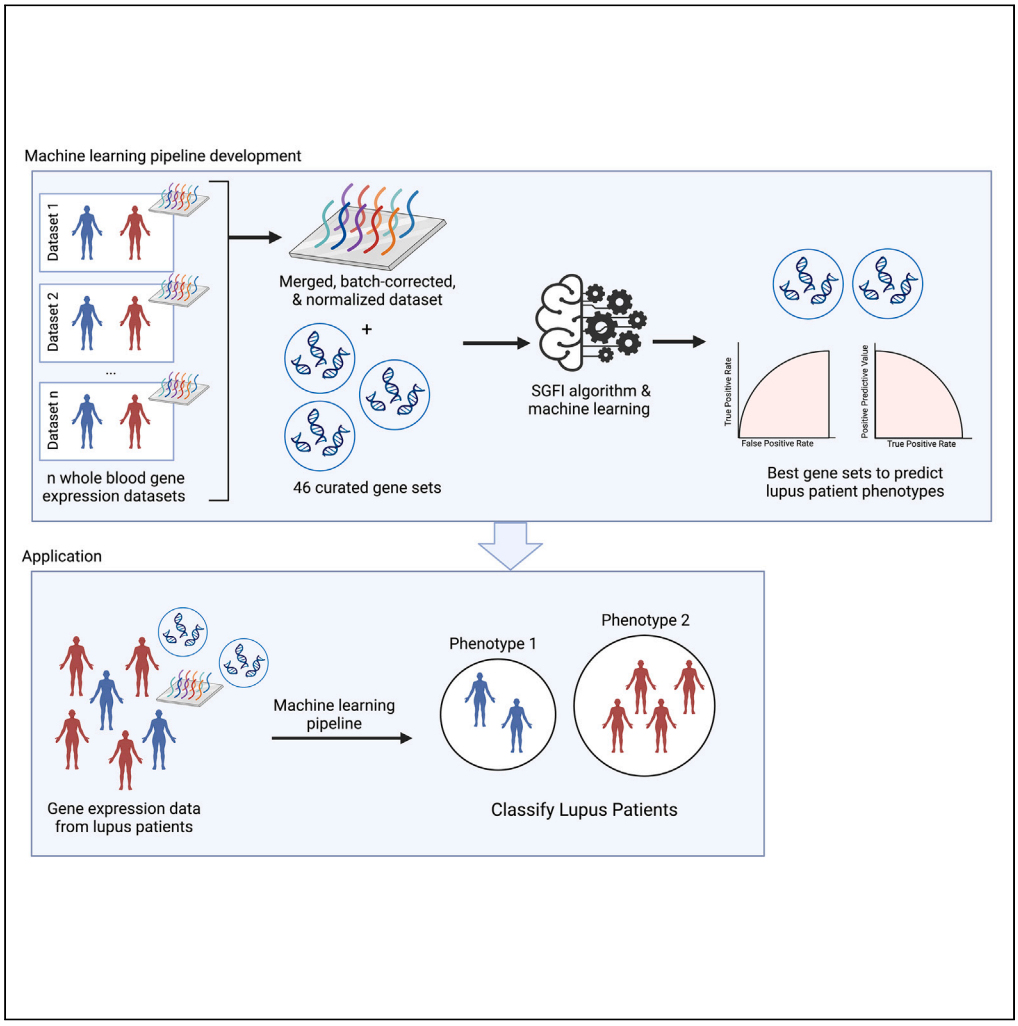
iScience

CellPress
OPEN ACCESS

Article

An interpretable machine learning pipeline based on transcriptomics predicts phenotypes of lupus patients



Emily L. Leventhal, Andrea R. Daamen, Amrie C. Grammer, Peter E. Lipsky

emily.leventhal@ampelbiosolutions.com

Highlights

An interpretable ML pipeline to predict clinical phenotypes in SLE was developed

The SGFI algorithm was employed on blood gene expression data

Disease-relevant gene sets distinguished SLE patient disease status

The SGFI-based ML pipeline outperformed previous efforts to classify patients with SLE

Leventhal et al., iScience 26, 108042
October 20, 2023
https://doi.org/10.1016/j.isci.2023.108042

iScience

CellPress
OPEN ACCESS

Article

An interpretable machine learning pipeline based on transcriptomics predicts phenotypes of lupus patients

Emily L. Leventhal,[1,2,*] Andrea R. Daamen,[1] Amrie C. Grammer,[1] and Peter E. Lipsky[1]

SUMMARY

Machine learning (ML) has the potential to identify subsets of patients with distinct phenotypes from gene expression data. However, phenotype prediction using ML has often relied on identifying important genes without a systems biology context. To address this, we created an interpretable ML approach based on blood transcriptomics to predict phenotype in systemic lupus erythematosus (SLE), a heterogeneous autoimmune disease. We employed a sequential grouped feature importance algorithm to assess the performance of gene sets, including immune and metabolic pathways and cell types, known to be abnormal in SLE in predicting disease activity and organ involvement. Gene sets related to interferon, tumor necrosis factor, the mitoribosome, and T cell activation were the best predictors of phenotype with excellent performance. These results suggest potential relationships between the molecular pathways identified in each model and manifestations of SLE. This ML approach to phenotype prediction can be applied to other diseases and tissues.

INTRODUCTION

Gene expression analysis holds the promise of identifying subsets of patients with specific clinical phenotypes and understanding their distinct pathologies. However, to date, it has not generated consensus information in most circumstances. In some studies, the characterization of patient subsets has relied on differential expression of a small number of genes in a limited number of patients, without considering whether these changes are representative across multiple patient populations.[1,2] Others have attempted to create machine learning (ML) models to predict phenotype by selecting individual genes to use as features, rather than considering how those genes fit into the broader context of disease pathogenesis.[3–5] In this article, we aim to establish an ML pipeline to identify and understand the specific molecular pathways implicated in phenotypic subsets of patients by using a systems biology lens.

Given its multitude of phenotypes and their largely unknown molecular pathologies, we use systemic lupus erythematosus (SLE) as a test case for our pipeline. SLE is a complex autoimmune disease characterized by multi-organ inflammation and a wide range of clinical manifestations.[6,7] Additionally, the course of SLE is characterized by flares of disease activity with variable outcomes interspersed with periods of disease quiescence.[8] Because of the heterogeneity and complexity of the disease, recognition of SLE is oftentimes delayed or inaccurate. Physicians must rely on a combination of criteria, consisting mostly of clinical evaluations and measurements of autoantibodies to arrive at a diagnosis of SLE. In practice, this is a prolonged process that can delay definitive identification.[9–12] Earlier and more accurate detection of SLE and SLE subtypes is important, as prompt recognition is paramount for effective treatment and to prevent irreversible damage.[8] Additionally, delayed treatment is associated with worse prognosis, decreased survival, and worsened quality of life.

As a clinically heterogeneous, multi-system disease, SLE can affect many organs of the body, including the skin, joints, kidneys and the nervous system. However, the detection of organ manifestations of SLE often relies on invasive tests, such as biopsies, to confirm disease after clinical findings suggest organ dysfunction. For example, lupus nephritis (LN) is a severe organ manifestation of SLE that affects up to 50% of patients with SLE.[7,13] The gold standard for diagnosing LN is the kidney biopsy, which is usually carried out after the detection of signs of organ dysfunction in the urine.[14] However, LN can be present in patients without clinical suggestions of kidney disease.[7] Therefore, diagnosis based on clinical presentation is not always adequate. Furthermore, early recognition of LN is particularly important, as LN is a major source of morbidity and mortality in SLE, and late diagnosis is a risk factor for end-stage renal disease.[7]

Blood-based molecular classification of SLE pathogenesis and organ involvement using gene expression data offers a promising alternative to traditional methods of characterizing various presentations of SLE. Gene expression in the blood of patients with SLE has been shown to be quite variable ([15]). However, coupled with ML, this approach has the potential to unravel some of the heterogeneity of the systems involved in lupus.[4,16,17] An interpretable ML model can uncover patterns in gene expression not otherwise observable and further, can provide insight into the biological profiles of patients with SLE.[17] Both supervised and unsupervised ML learning algorithms have been

[1]AMPEL BioSolutions LLC, and the RILITE Research Institute, Charlottesville, VA 22902, USA
[2]Lead contact
*Correspondence: emily.leventhal@ampelbiosolutions.com
https://doi.org/10.1016/j.isci.2023.108042

applied to SLE transcriptomes in previous studies. However, many of the previous studies have not reported all the accuracy metrics necessary for validation or have employed small datasets from a single center and, thus, may be over-fitted.[3,4,18] As a result, no ML algorithms based on blood gene expression have been adapted in the clinic to aid in the diagnosis of SLE or the recognition of lupus phenotypes.

Instead of assessing the performance of individual genes, it can be more biologically reliable and informative to assess the performance of gene sets.[19] Selecting grouped features instead of individual features for ML is especially helpful for interpreting high-dimensional data, such as gene expression data. Uniquely, we have previously created 46 gene sets comprised of immune pathways, metabolic pathways, and inflammatory cell types that are potentially involved in the pathogenesis of SLE.[20,21] To understand the diverse pathways involved in various clinical manifestations of SLE in greater detail, we apply the sequential grouped feature importance (SGFI) algorithm to determine the best combinations of these gene sets to predict SLE and SLE clinical phenotypes.[22] SGFI sequentially adds gene sets as features to the model in terms of leave-one-in-importance until the performance no longer improves. As such, SGFI allows for the selection of the best and sparsest combination of gene sets to classify samples into the correct clinical phenotype.

Here, we describe a systems biology approach utilizing interpretable, supervised ML to predict SLE clinical phenotypes from blood gene expression data. We additionally use this approach to differentiate SLE from another inflammatory autoimmune disease, namely, rheumatoid arthritis (RA). Notably, the models identify gene sets with potential physiologic importance in each disease subtype, providing new insights into the molecular pathways and individual genes most relevant to the classification of SLE disease pathology.[17] Moreover, the excellent performance of these models suggests that this approach could be used to support decision making in clinical care of patients with SLE and those with other diseases.

RESULTS

Predicting systemic lupus erythematosus from healthy control blood from gene expression profiles

Data preparation

To create an ML model to predict SLE from CTL whole blood samples, we developed a pipeline to normalize, combine, and batch correct publicly available gene expression datasets across different microarray platforms (Figure 1; Table S2A).[23–28] Each dataset was pre-processed separately before merging the datasets by genes shared between all arrays to yield a final merged dataset with 13,111 gene features and 943 samples (765 SLE and 178 CTL samples) (Figure S1A).

Visualizing the merged data by both principal component analysis (PCA) and sample density plots (Figure S1B) revealed a batch effect. After adjusting for the batch effects with ComBat, the explained variance of the first principal component decreased dramatically from 35.7% to 5.2%, and the samples across different batches had similar distributions (Figure S2C). We split the merged, batch-corrected dataset into train (70%) and test (30%) sets and standardized the train and test sets.

Feature selection

In selecting the features, or genes, to use in the ML model, we chose to leverage our knowledge of sets of genes with similar functions or expressions. Instead of selecting the best individual features (genes), we selected the best feature groups (gene sets) using SGFI.[22] Utilizing gene sets rather than individual genes allows for a more interpretable model. The inputs were 46 gene sets of curated immune pathways, metabolic pathways, and cell types with relevance to lupus pathology (Table S1).[20,21]

The Sankey diagram in Figure 2A demonstrates the SGFI algorithm workflow. For each of 100 subsample iterations, the algorithm begins from the null model and then sequentially adds in the next best feature group in terms of leave-one-group-in importance (LOGI) if the mean misclassification error (MMCE) improves beyond a threshold. In most iterations, genes upregulated in response to interferon (*IFN*) (52 iterations), *Monocyte* (21 iterations), *Unfolded Protein* (7 iterations), and genes upregulated in response to tumor necrosis factor (*TNF induced genes*) (6 iterations) were chosen as the best single feature group to predict SLE from CTL blood samples in the first round. Combining these gene sets or adding in genes from a different second or third gene set (e.g., mitochondrial ribosome-large subunit (*Mito Large Ribo*), *TCA cycle*) often further improved performance (Figure 2A). We extracted all feature group combinations that were selected by the algorithm in five or more of the 100 iterations as the possible feature sets to use in the final model (Figure 2A; Table S2B). These feature sets advanced to the next step of hyperparameter tuning and cross-validation on the train set.

Hyperparameter tuning and cross-validation

For each feature set, we created a radial support vector machine (SVM) model and carried out hyperparameter tuning. We then ran 10-fold cross-validation (CV) of each tuned model on the train set to predict how the features might perform on new, unseen data, and to determine the combination of features with the best overall performance metrics. The model created with the genes from the combined *IFN*, *TNF induced genes*, and *Mito Large Ribo* gene sets had the highest average F1 score (F1 = 0.88) for predicting SLE from CTL in the train set, so it was chosen as the final model (Table S2B).

Evaluation on the test set

To predict how the ML model will perform on future data, we evaluated the model performance on an unseen test set, that had not been used in the creation of the model. We trained the final model on the train set using the 189 genes from the *IFN*, *TNF induced genes*, and *Mito Large Ribo* gene sets and the hyperparameters determined from hyperparameter tuning. The model achieved excellent classification of SLE vs. CTL

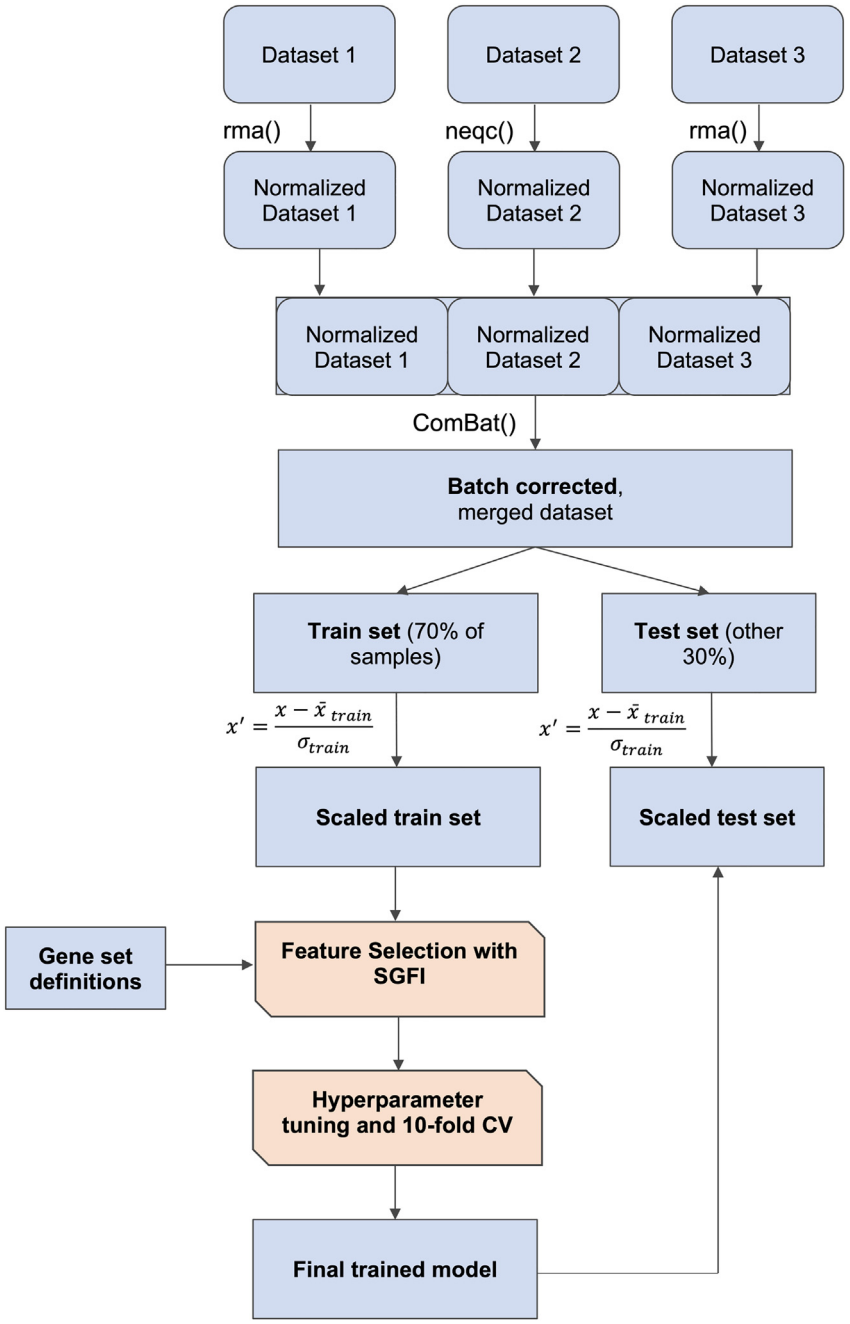


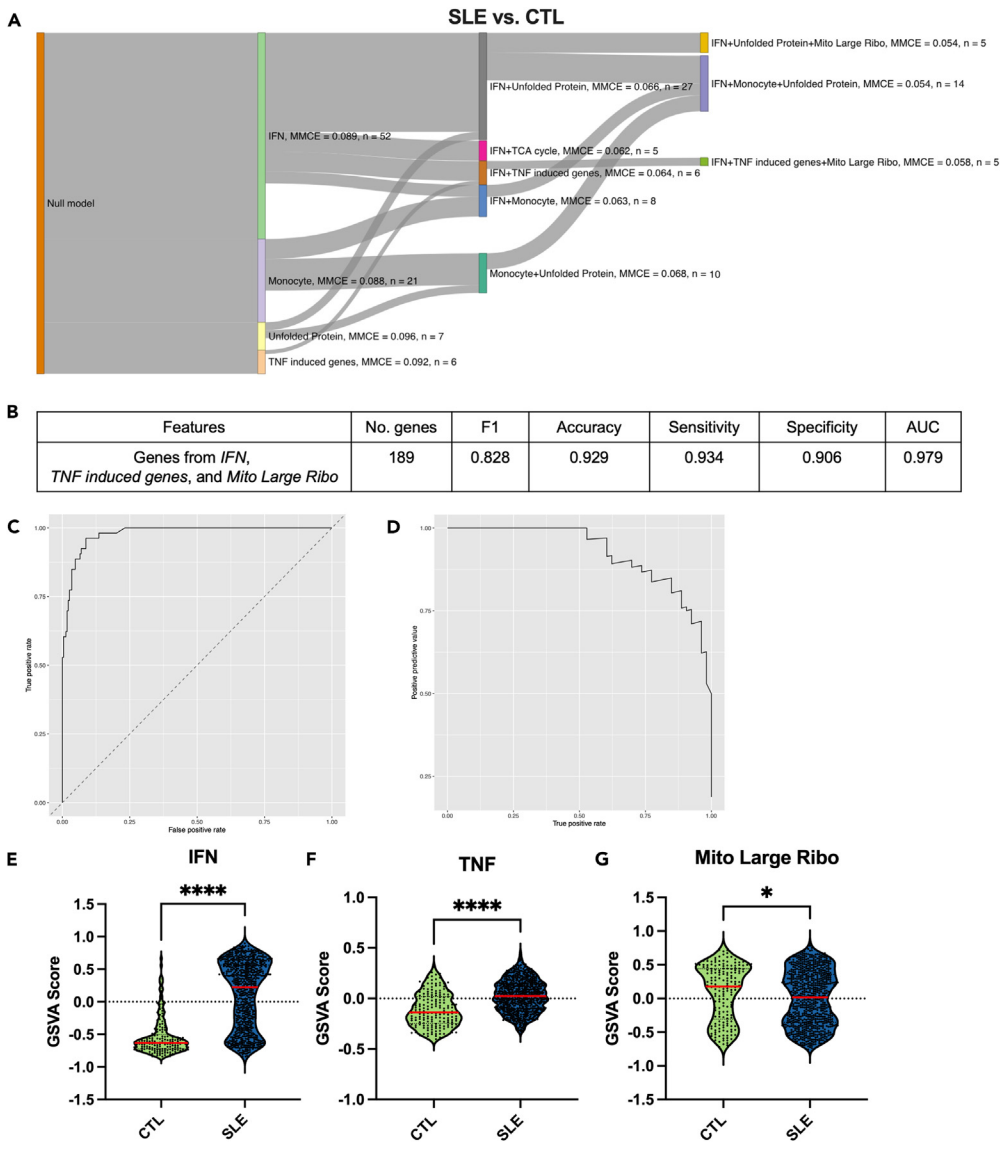
Figure 1. Pipeline to predict disease status using machine learning

The datasets are quantile normalized, merged by common gene symbol, batch-corrected, and split into a train and test set. The scaled train set is inputted into SGFI. For each feature group combination chosen ($n \geq 5$) in SGFI, a tuned radial SVM model is created, and 10-fold cross validation (CV) is performed. The model with the best F1 score is chosen as the final model and is evaluated on the test set.

samples with an area under the curve (AUC) of 0.979 and an accuracy, sensitivity, and specificity above 0.9 (Figure 2B). The receiver operating characteristic (ROC) curve and precision-recall curve are shown in Figures 2C and 2D.

Analysis of genes that classify systemic lupus erythematosus vs. control blood samples

To interpret the feature selection results and determine how the selected gene sets may be differentially enriched in SLE samples compared to CTLs, we ran gene set variation analysis (GSVA) on all the datasets used in the model with the *IFN*, *TNF induced genes*, and *Mito Large Ribo*



Features	No. genes	F1	Accuracy	Sensitivity	Specificity	AUC
Genes from *IFN*, *TNF induced genes*, and *Mito Large Ribo*	189	0.828	0.929	0.934	0.906	0.979

Figure 2. Results from the SLE vs. CTL pipeline

(A) Sankey plot of the results from feature selection. Gene set combinations chosen in five or more subsamples are shown in the plot. Gene sets are separated by the symbol, "+." For each gene set combination, the MMCE, averaged across the subsamples in which that gene set combination was chosen, and the number of subsamples (n) is shown.

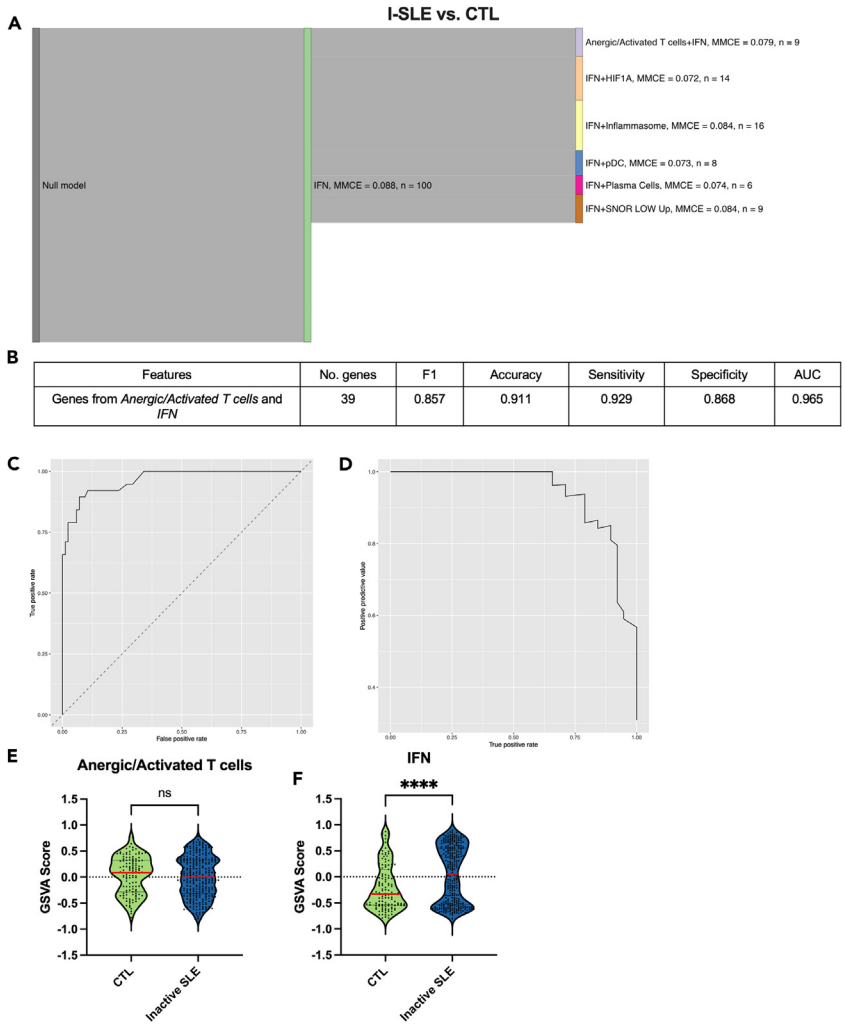
(B) Accuracy metrics from the evaluation of the final model on the test set.

(C–G) (C) ROC and (D) PR curves of the final model on the test set. GSVA results of (E) *IFN*, (F) *Mito Large Ribo*, and (G) *TNF induced genes*. Unpaired t-test, *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

gene sets and merged the GSVA scores (Figures 2E–2G). *IFN* and *TNF induced genes* were significantly higher in SLE compared to CTL blood, whereas *Mito Large Ribo* was significantly lower in SLE compared to CTL. Additionally, we conducted differential expression (DE) analysis of the genes within the top gene sets across all datasets, and all 189 selected genes were differentially expressed in at least one dataset with both SLE and CTL samples (Figure S1D).

Independent validation on a new dataset

To further validate that these genes could distinguish SLE from CTL blood, we tested how the "locked" model, or the model with the same hyperparameters and features, would perform using a completely independent dataset (Table S2C). We carried out the same pre-processing procedure and trained the radial SVM model with the same hyperparameters and features on 70% of the merged validation dataset and



Features	No. genes	F1	Accuracy	Sensitivity	Specificity	AUC
Genes from *Anergic/Activated T cells* and *IFN*	39	0.857	0.911	0.929	0.868	0.965

Figure 3. Results from the inactive SLE vs. CTL pipeline

(A) Sankey plot of the results from feature selection. Gene set combinations chosen in five or more subsamples are shown in the plot. Gene sets are separated by the symbol, "+." For each gene set combination, the MMCE, averaged across the subsamples in which that gene set combination was chosen, and the number of subsamples (n) is shown.

(B) Accuracy metrics from the evaluation of the final model on the test set.

(C–F) (C) ROC and (D) PR curves of the final model on the test set. GSVA results of (E) *Anergic/Activated T cells* and (F) *IFN*. Unpaired t-test, *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

tested it on the remaining 30% of the data. The locked model performed extremely well in classifying the independent validation data with an AUC of 0.992 in 10-fold CV on the train set and 0.985 on the test set (Figures S2A and S2B).

Distinguishing inactive systemic lupus erythematosus from control blood

Because of the excellent performance of the model generated from the SGFI pipeline in classifying patients with SLE from CTLs, we next determined whether this approach could also distinguish I-SLE from CTL blood. We normalized, combined, and batch-corrected three microarray datasets with patients with I-SLE (SLEDAI <6) and CTL (Table S3A; Figures S3A–S3C).[25,29] The final merged dataset consisted of 285 patients with I-SLE, 127 patients with CTL expressing 14,725 genes.

Next, we applied the SGFI algorithm to the train set (70%) of the merged I-SLE vs. CTL dataset. In all 100 iterations, *IFN* was the best single gene set to distinguish I-SLE from patients with CTL (Figure 3A). However, including genes from a second gene set (e.g., *Anergic/Activated T cells, Plasma Cells, Inflammasome)* often improved performance. Gene set combinations chosen in five or more iterations were considered possible feature sets for the final model (Table S3B). Then, we created a radial SVM model for each feature set and performed hyperparameter tuning followed by 10-fold CV on the tuned model (Table S3B). The tuned model created with the *Anergic/Activated T cells* and *IFN* gene sets achieved the highest F1 score at 0.865 and was thus chosen as the final model. Finally, we applied this model using 39 genes from

the *Anergic/Activated T cells* and *IFN* gene sets as features on the train set and evaluated it on the test set. The model achieved excellent classification on the test set with an AUC of 0.965, a sensitivity of 0.929, and a specificity of 0.868 (Figures 3B–3D).

To characterize the differences between I-SLE and CTL in the selected gene sets, we ran GSVA on each dataset and merged the GSVA scores. We found that the *IFN* signature was significantly higher in I-SLE compared to CTL whole blood whereas the *Anergic/Activated T Cells* signature was not significantly different (Figures 3E and 3F). When further analyzing DE of individual genes, 38 of the 39 genes used in the final model were significantly different in I-SLE compared to CTL in at least one of the datasets with both I-SLE and CTL samples (Figure S3D). Notably, although the *Anergic/Activated T Cells* signature was not significantly different by GSVA, the individual genes within the gene set were significantly differentially expressed between I-SLE and CTL.

Predicting active systemic lupus erythematosus from patients with inactive systemic lupus erythematosus

After creating a model to distinguish inactive from control patients, we sought to determine which gene sets best predicted whether a patient had low or high disease activity. We merged available microarray datasets with active (SLEDAI $\geq$6) and inactive (SLEDAI <6) patients to create a final dataset with 14,733 features and 437 patient samples (Table S4A; Figures S4A–S4C).[25,28] Similar to the results from feature selection for I-SLE vs. CTL, *IFN* was chosen in all iterations as the best gene set in the first round of SGFI to differentiate active SLE from I-SLE (Figure 4A). Then, for each gene set combination chosen more than five times in all rounds of SGFI, models were created using the train set, hyperparameter tuned, and assessed with 10-fold CV (Table S4B). The top performing model on the train set was that created with *IFN* genes (F1 = 0.86) and was chosen as the final model. The final SVM classifier was then evaluated on the test set. The model achieved good classification of active from patients with I-SLE with an AUC of 0.842 (Figures 4B–4D). We found the GSVA scores for the *IFN* signature were significantly higher in patients with active SLE compared to I-SLE (Figure 4E).

Distinguishing lupus nephritis from control blood

Next, we created a model to distinguish LN from CTL whole blood samples, using six microarray datasets that were processed to yield a final combined dataset with 14,632 features and 277 samples (Table S5A; Figures S5A–S5C).[29–32] The SGFI algorithm revealed that the *IFN*, *TNF induced genes*, *Monocyte*, *Oxidative Phosphorylation*, and *B cells* gene sets were the best feature groups to predict LN from CTL samples in the first round, and combining these gene sets often improved the model (Figure 5A). Many of the gene sets chosen as the best predictors of LN from CTL samples were also chosen when evaluating the best predictors for SLE from CTL (e.g., *IFN*, *Monocyte*, *TNF induced genes*). The feature group combinations chosen more than five times were considered as possible feature sets for the final model (Table S5B). After hyperparameter tuning and 10-fold CV on the train set, the model created with genes from the *IFN* and *TNF induced genes* gene sets had the highest F1 score (F1 = 0.898) and was thus chosen as the final model (Table S5B).

We trained the final model on the train set with the 143 genes from *IFN* and *TNF induced genes* as features and the hyperparameters determined from hyperparameter tuning, and evaluated the model by its prediction of the test set samples. The model achieved excellent classification of LN vs. CTL on the test set with an AUC of 0.943 (Figures 5B–5D). GSVA with the gene sets used in the final model revealed that both the *IFN* and *TNF induced genes* gene sets were significantly enriched in LN samples compared to CTL samples (Figures 5E and 5F). Through DE analysis, we found that expression of 119 of the 143 genes used in the model were significantly changed in at least one dataset with both LN and CTL samples (Figure S5D). IFN genes were consistently higher in LN samples compared to CTL samples, whereas some TNF induced genes were significantly higher and others were significantly lower in LN compared to CTL.

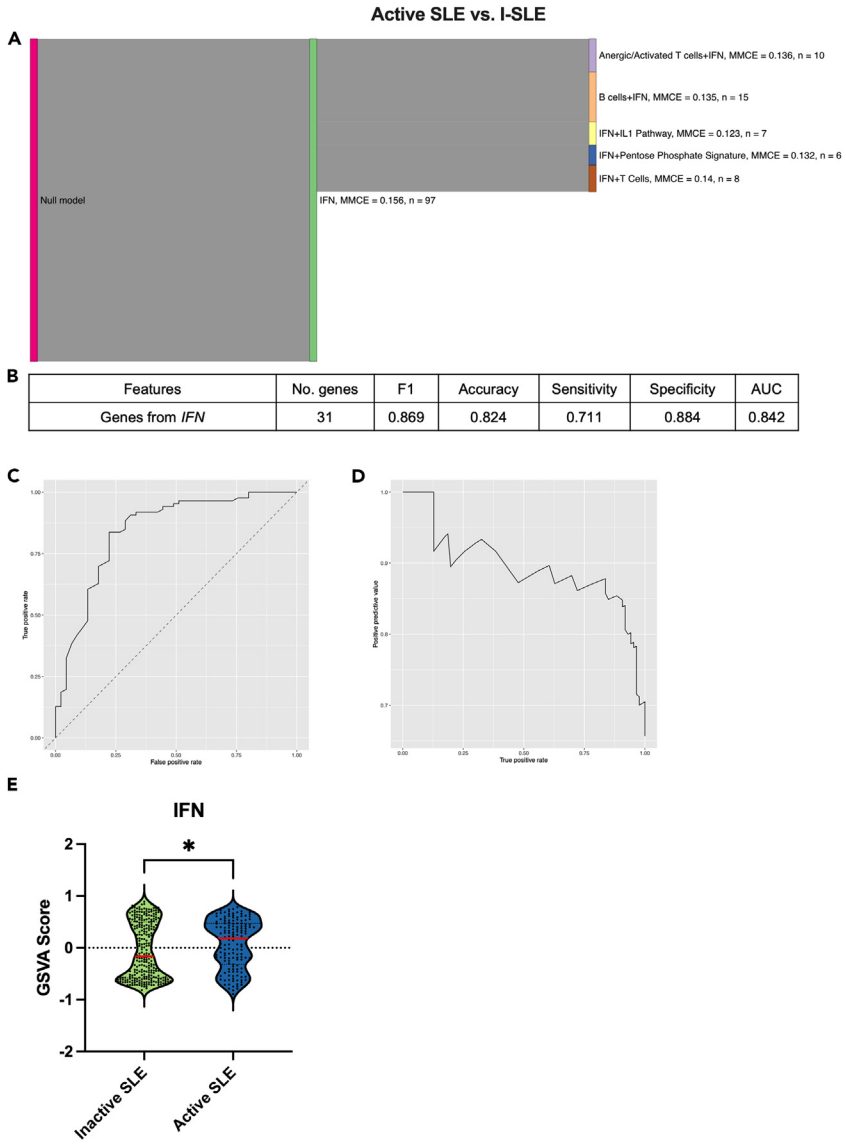
Discriminating lupus nephritis from non-renal lupus blood

We then created a model to predict LN from NRL whole blood, combining six microarray datasets with LN and confirmed NRL blood samples (Table S6A; Figures S6A–S6C).[24,28,30–32] The final batch-corrected, merged dataset had 14,632 features and 415 samples (138 LN, 277 NRL). The results of the SGFI pipeline revealed that the *IFN*, *B cells*, *Mito Large Ribo*, *Mito Small Ribo*, *TCA cycle*, *TNF induced genes*, and *Unfolded Protein* gene sets were most often the best single feature group to predict LN from NRL blood (Figure 6A). Adding a second gene set of immune cell populations or metabolism signatures such as *Inflammatory Cytokines*, *Anergic/Activated T Cells*, *B cells* or *Mito Large Ribo* to *IFN* often increased the performance of the model. We conducted hyperparameter tuning and performed 10-fold CV on the possible gene set combinations (Table S6B). The model created with the 80 genes from the *IFN* and *Mito Large Ribo* gene sets had the highest F1 score at 0.917 to predict LN from NRL, and was chosen as the final model to evaluate on the test set. We found that the model worked well in classifying LN from NRL in the test set with an AUC of 0.894 (Figures 6B–6D).

GSVA revealed that *Mito Large Ribo* was significantly lower in LN compared to NRL samples, whereas *IFN* was not significantly different (Figures 6E and 6F). DE analysis revealed that 73 of the 80 genes were significantly different in at least one of the datasets with both LN and NRL samples (Figure S6D). Although *IFN* was not significantly different when considering the gene sets as a whole, by examining individual gene differences we found that many IFN genes were significantly downregulated in LN compared to NRL in the blood.

Using protein-protein interaction clusters to predict lupus nephritis from non-renal lupus

Next, we wanted to a use a secondary, unsupervised approach to creating gene sets able to distinguish LN from NRL and explore whether it would perform as well as our previously published gene sets. We created a network of protein-protein interactions (PPIs) derived from the STRING database using the top 2,000 variable genes from the merged LN/NRL dataset. We then carried out MCODE clustering on the



Features	No. genes	F1	Accuracy	Sensitivity	Specificity	AUC
Genes from *IFN*	31	0.869	0.824	0.711	0.884	0.842

Figure 4. Results from the active SLE vs. I-SLE pipeline

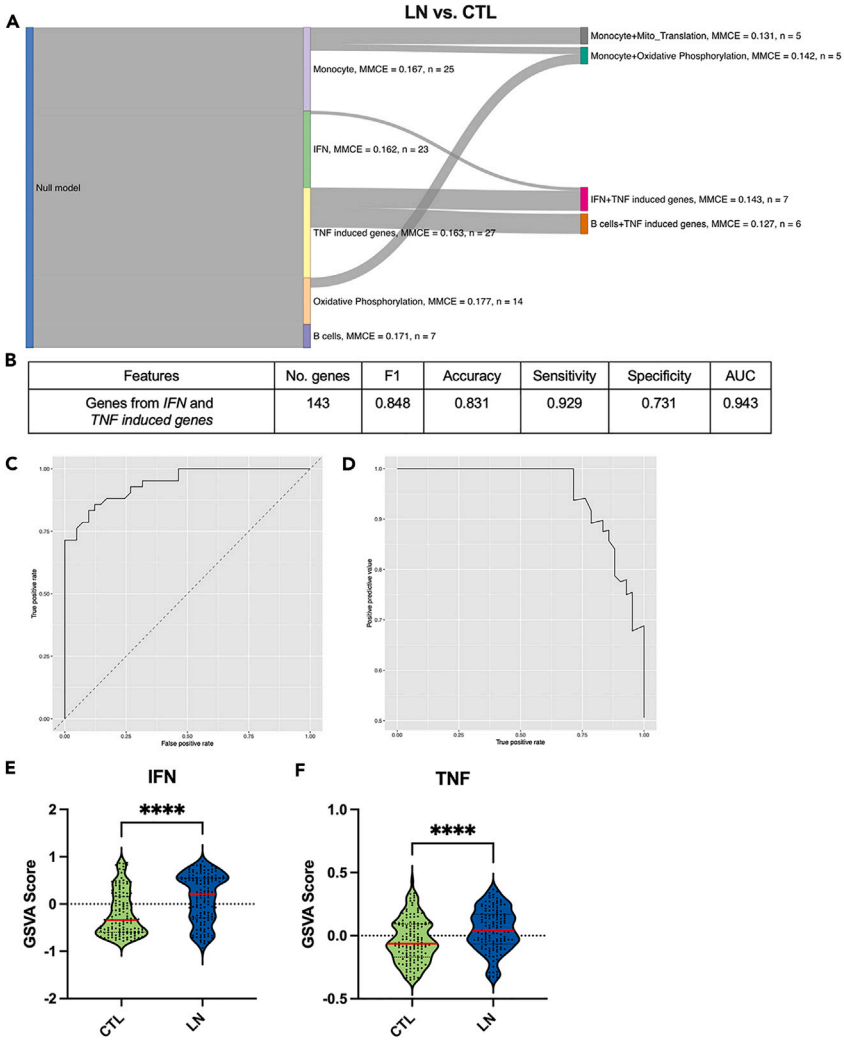
(A) Sankey plot of the results from feature selection. Gene set combinations chosen in five or more subsamples are shown in the plot. Gene sets are separated by the symbol, "+." For each gene set combination, the MMCE, averaged across the subsamples in which that gene set combination was chosen, and the number of subsamples (n) is shown.

(B) Accuracy metrics from the evaluation of the final model on the test set.

(C–E) (C) ROC and (D) PR curves of the final model on the test set. GSVA results of (E) *IFN*. Unpaired t-test, *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

network to create gene clusters to be used as feature group inputs for SGFI. The MCODE clusters were annotated based on overlap with our curated gene sets and gene ontology (GO) terms (Figure 7A). Then, to evaluate whether the PPI-based MCODE gene clusters would be effective feature groups for an ML model to predict disease status, we ran SGFI with the same input datasets and train set previously used to predict LN from NRL (Figure S6A), but with MCODE clusters as feature group inputs.

Clusters *A*, *AA*, *L*, *P*, and *D* best classified LN from NRL in the train set (Figure 7B). In most cases, adding a second MCODE cluster did not improve the model performance, but combining clusters *L* and *P* created a better model in 8 iterations. For each feature set, we created a radial SVM model and tuned the hyperparameters. The results from 10-fold CV indicated that cluster *A* was the best feature group with an F1-Score of 0.922 (Table S6C). The tuned model with cluster *A*, which overlapped with gene sets *IFN, TNF induced genes, Mito Large Ribo, Mito Small Ribo, Mito Translation, GO MITO FISSION, and Cell Cycle*, was chosen as the final model. After evaluation on the test set, the model achieved excellent classification of LN and NRL samples, with an AUC of 0.956 (Figures 7C–7E).



Features	No. genes	F1	Accuracy	Sensitivity	Specificity	AUC
Genes from *IFN* and *TNF induced genes*	143	0.848	0.831	0.929	0.731	0.943

Figure 5. Results from the LN vs. CTL pipeline

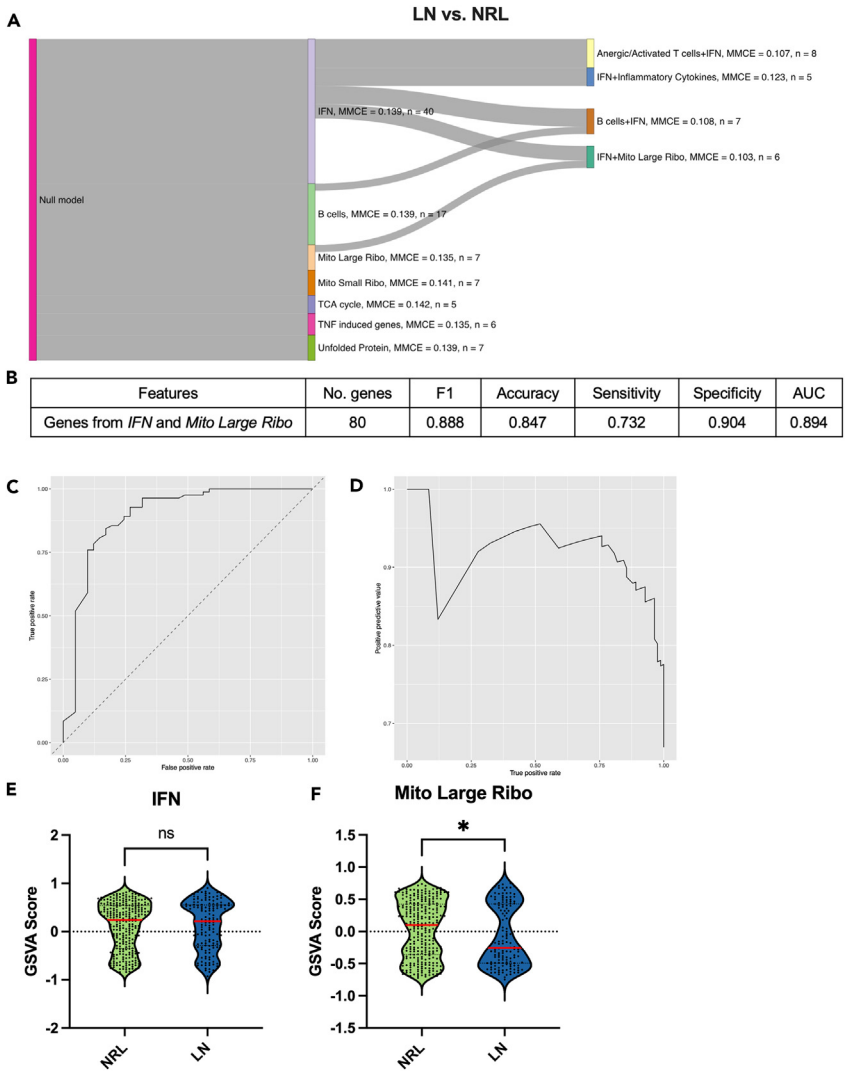
(A) Sankey plot of the results from feature selection. Gene set combinations chosen in five or more subsamples are shown in the plot. Gene sets are separated by the symbol, "+." For each gene set combination, the MMCE, averaged across the subsamples in which that gene set combination was chosen, and the number of subsamples (n) is shown.

(B) Accuracy metrics from the evaluation of the final model on the test set.

(C–F) (C) ROC and (D) PR curves of the final model on the test set. GSVA results of (E) *IFN* and (F) *TNF induced genes*. Unpaired t-test, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

Predicting differential recognition of rheumatoid arthritis and systemic lupus erythematosus

Finally, we used this ML approach to differentiate SLE from a different chronic, inflammatory autoimmune disease, namely, RA. We applied the pipeline to a merged dataset with 809 RA samples, 518 SLE samples, and 16,761 features (Table S7A; Figures S7A–S7C).[24,25,33] In the first round of SGFI, *IFN* best predicted RA and SLE samples in two-thirds of the iterations, while *TNF induced genes* best predicted RA versus SLE in 16 iterations. In 28 iterations, combining *IFN* and *TNF induced genes* improved the performance (Figure 8A). Models were created for each gene set combination selected in more than five iterations, and they were hyperparameter tuned and assessed via 10-fold CV on the train set. The tuned model with *IFN* and *TNF induced genes* had the highest F1 score (F1 = 0.944) and was chosen as the final model. When evaluated on the test set, the model achieved excellent classification with an AUC of 0.989 and all other accuracy metrics above 0.91 (Figures 8B–8D). GSVA revealed that *IFN* was significantly de-enriched in patients with RA as compared to patients with SLE whereas the *TNF induced genes* signature was not significantly different between patients with SLE and RA by GSVA (Figures 8E and 8F). The differential expression results indicated that *IFN* genes were consistently downregulated in RA compared to SLE, whereas genes within the *TNF induced genes* signature were either up- and down-regulated in RA (Figure S7D). The variability in the differential expression of TNF induced genes most likely contributed to the lack of significant GSVA score differences, which highlights the ability of SGFI to consider the expression of individual genes and



Features	No. genes	F1	Accuracy	Sensitivity	Specificity	AUC
Genes from *IFN* and *Mito Large Ribo*	80	0.888	0.847	0.732	0.904	0.894

Figure 6. Results from the LN vs. NRL pipeline

(A) Sankey plot of the results from feature selection. Gene set combinations chosen in five or more subsamples are shown in the plot. Gene sets are separated by the symbol, "+." For each gene set combination, the MMCE, averaged across the subsamples in which that gene set combination was chosen, and the number of subsamples (n) is shown.

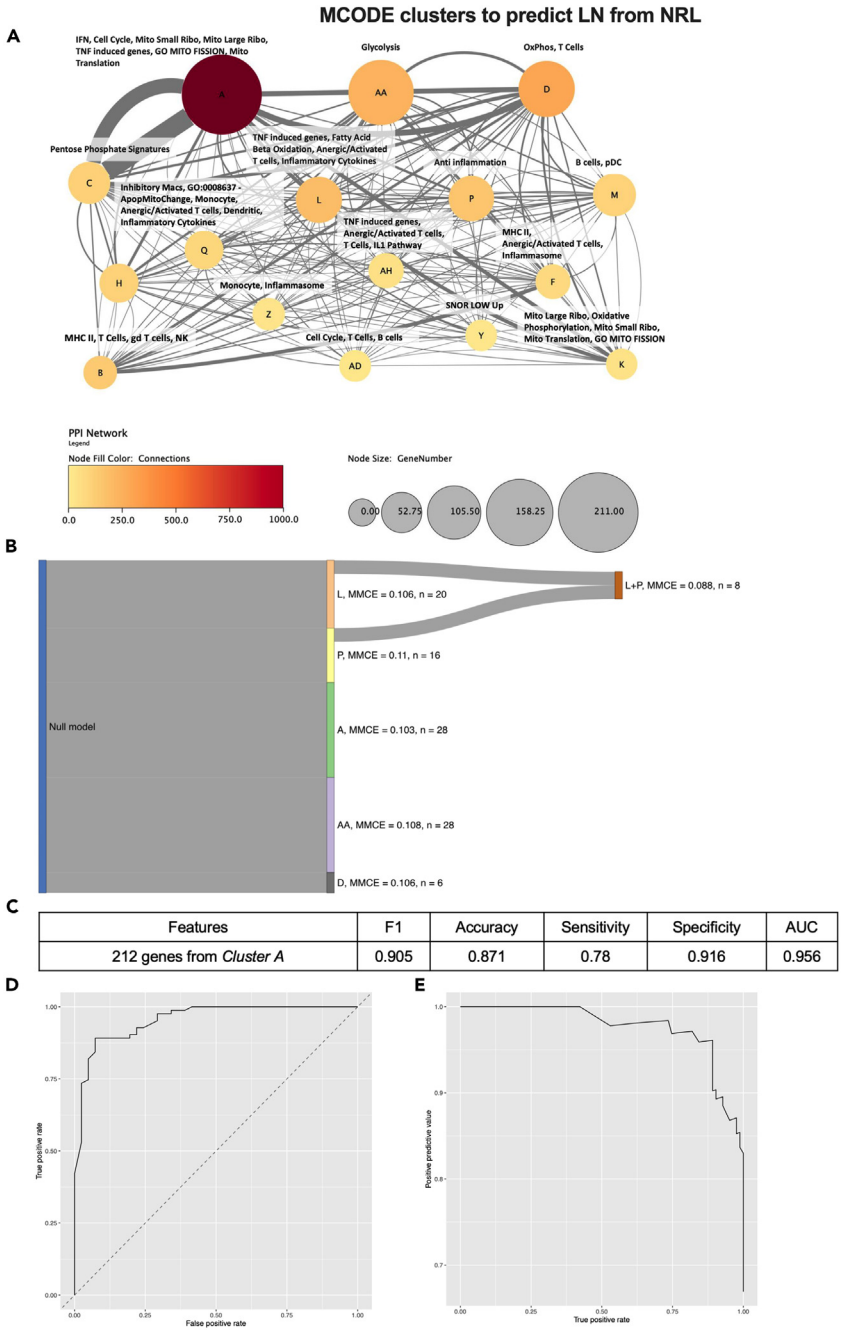
(B) Accuracy metrics from the evaluation of the final model on the test set.

(C–F) (C) ROC and (D) PR curves of the final model on the test set. GSVA results of (E) *IFN* and (F) *Mito Large Ribo*. Unpaired t-test, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$.

preserve their variability while also considering the genes as a meaningful functional group. The ML approach we have described in this article was able to distinguish patients both within and across autoimmune diseases with high accuracy.

DISCUSSION

ML has promising implications for creating accurate predictions about disease status from non-invasive blood samples. However, previous efforts that have used ML for the classification of patients with SLE have been difficult to generalize or have only focused on the contribution of individual genes to predict disease status. In this study, we introduced an interpretable ML approach that leveraged the SGFI algorithm to predict disease status from merged gene expression datasets using a systems biology lens.[22] Though interpretable ML cannot distinguish between causal and noncausal effects, it can be used to suggest potential causal relationships.[17] Therefore, through the process of selecting disease-associated gene sets with the greatest capability of identifying groups of patients, our results additionally highlight potential pathogenic mechanisms with particular relevance to SLE and SLE clinical phenotypes that can be further analyzed through causal inference methods or additional experiments.



Features	F1	Accuracy	Sensitivity	Specificity	AUC
212 genes from *Cluster A*	0.905	0.871	0.78	0.916	0.956

Figure 7. Results from the LN vs. NRL pipeline using MCODE clusters as feature group definitions

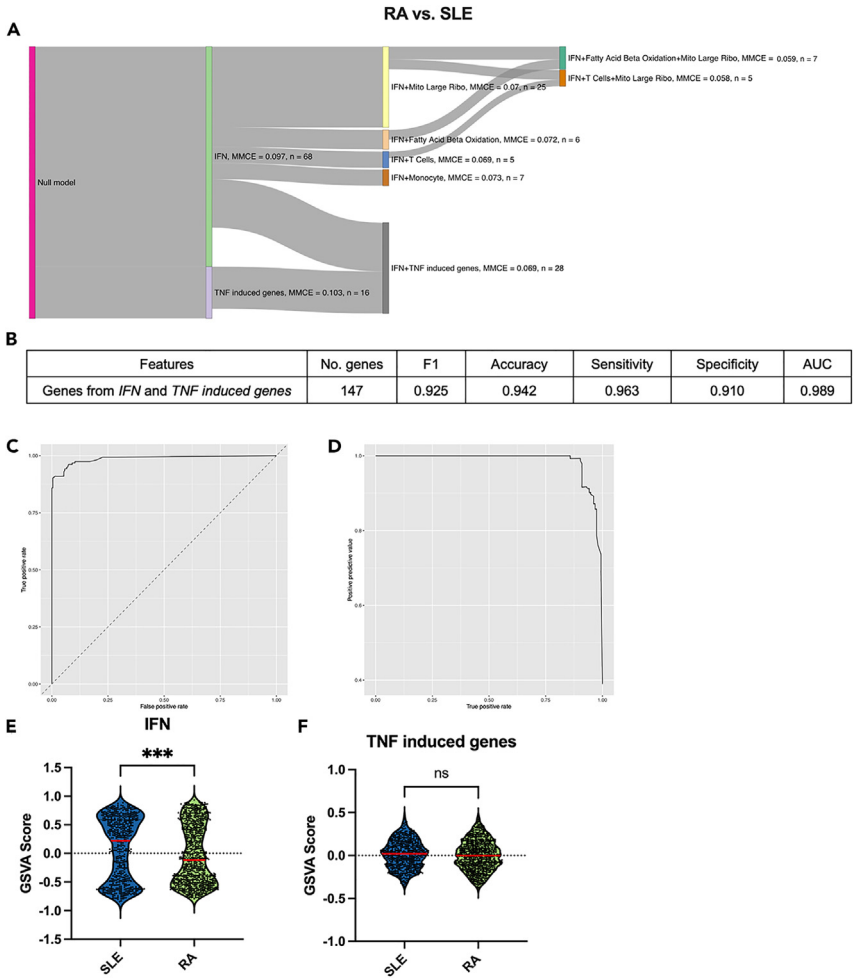
(A) MCODE clusters from protein-protein interactions, annotated with our curated gene sets.

(B) Sankey plot of the results from feature selection. Gene set combinations chosen in five or more subsamples are shown in the plot. Gene sets are separated by the symbol, "+." For each gene set combination, the MMCE, averaged across the subsamples in which that gene set combination was chosen, and the number of subsamples (n) is shown.

(C) Results from the evaluation of the final model on the test set.

(D and E) (D) ROC and (E) PR curves of final model on the test set.

The SGFI algorithm was highly successful at selecting well-performing combinations of gene sets to predict clinical phenotype. In addition to being biologically informative, selecting for gene sets may also be more accurate than using an individual gene approach for a number of reasons. For instance, disease pathogenesis is often associated with changes in expression of sets of genes with similar functionality, so



Features	No. genes	F1	Accuracy	Sensitivity	Specificity	AUC
Genes from *IFN* and *TNF induced genes*	147	0.925	0.942	0.963	0.910	0.989

Figure 8. Results from the RA vs. SLE pipeline

(A) Sankey plot of the results from feature selection. Gene set combinations chosen in five or more subsamples are shown in the plot. Gene sets are separated by the symbol, "+." For each gene set combination, the MMCE, averaged across the subsamples in which that gene set combination was chosen, and the number of subsamples (n) is shown.

(B) Accuracy metrics from the evaluation of the final model on the test set.

(C–F) (C) ROC and (D) PR curves of the final model on the test set. GSVA results of (E) *IFN* and (F) *TNF induced genes*. Unpaired t-test, *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

differences in a group of genes may be more reliable than changes in an individual gene. Additionally, many genes are multi-functional, so their individual selection may lead to ambiguous conclusions.[19] Finally, SGFI provides a way to meaningfully reduce the high dimensionality of gene expression datasets, combatting the "curse of dimensionality" present in biological datasets in which the number of features far exceeds the number of samples. A common approach in computational modeling is to leverage prior knowledge of the system into the model, as it can help find the optimal solution.[34] The SGFI algorithm provides a method to incorporate *a priori* knowledge of biological processes and functions into the data while also selecting pathways for the model in an unbiased manner, explaining how the method can achieve such high performance. Thus, our approach in utilizing curated gene sets indicative of disease-relevant cell types and pathways integrates previous knowledge of processes potentially involved in SLE pathogenesis into a novel pipeline that selects those with the greatest efficacy in patient classification.

Previous studies have attempted to predict clinical phenotypes and disease activity from gene expression data. One study sought to characterize the molecular heterogeneity of SLE in a longitudinal cohort but was limited to pediatric patients. They used mixed models to stratify patients based on immune sets that best correlate with disease activity, and they identified a plasmablast signature as the best marker of disease activity in pediatric patients.[35] Another study postulated that common determinants contribute to both vaccine response and disease activity in SLE. As such, they identified gene signatures related to vaccine response and correlated those with disease activity in SLE.[36] A third study reported signatures related to disease state and disease activity, but the analysis was limited to isolated cell types.[37] The current study is

Table 1. Summary of the accuracy metrics in the seven machine learning models constructed in this article for different classification problems

Classification problem	Gene sets	No. Genes	F1	Accuracy	Sensitivity	Specificity	AUC
SLE vs. CTL	*IFN, TNF induced genes, and Mito Large Ribo*	189	0.828	0.929	0.934	0.906	0.979
I-SLE vs. CTL	*IFN and Anergic/Activated T Cells*	39	0.857	0.911	0.929	0.868	0.965
Active SLE vs. I-SLE	*IFN*	30	0.869	0.824	0.711	0.884	0.842
LN vs. CTL	*IFN and TNF induced genes*	143	0.848	0.831	0.929	0.731	0.943
LN vs. NRL	*IFN and Mito Large Ribo*	80	0.888	0.847	0.732	0.904	0.894
LN vs. NRL	*PPI Cluster A*	212	0.905	0.871	0.78	0.916	0.956
RA vs. SLE	*IFN and TNF induced genes*	147	0.925	0.942	0.963	0.91	0.989

the first effort applying an ML approach to whole blood samples from adult patients to identify gene sets that clearly separate patients with SLE and identify various clinical phenotypes.

Many of the models described here outperform previous attempts at predicting SLE status based on blood gene expression. In a previous study, we classified patients as active SLE or I-SLE by using a random forest classifier using all available gene expression data, which achieved an AUC of 0.89, or by using gene modules generated by Weighted Gene Co-expression Network Analysis (WGCNA), which had a peak AUC of 0.77.[38] Another group distinguished SLE from CTL patients using a six gene signature, achieving an AUC of 0.913 in the validation cohort, but no other accuracy metrics were reported.[5] A different study distinguished LN from CTL blood samples by using a single gene (HERC5, an interferon response gene) selected from traditional feature selection techniques, and obtained an area under the curve (AUC) of 0.880. However, no other accuracy metrics were reported.[3] In contrast, our models developed using expression of genes grouped by disease-relevant cell types or functions achieved higher overall accuracies (with AUCs ranging from 0.842 to 0.989) and presented interpretable results that could be translated to improved understanding of the molecular systems involved in SLE pathology.

Interestingly, subsets of the gene sets *IFN*, *TNF induced genes*, *Mito Large Ribo*, and *Anergic/Activated T cells* were chosen as the best predictors of disease status for all six classification problems (SLE vs. CTL, I-SLE vs. CTL, active SLE vs. I-SLE, LN vs. CTL, LN vs. NRL, and RA vs. SLE; Table 1). The idea that common gene sets were chosen for many classifications implies that these four gene sets may play essential roles in the pathogenesis of SLE and SLE clinical phenotypes.[17] However, other feature sets chosen by SGFI also performed well in classifying the train set samples in some cases and, therefore, may also be viable predictors of disease status. To choose the best gene set combinations overall from the SGFI results, we carried out CV on all combinations selected by SGFI and chose the model with the highest F1 score. Additional investigation of the other potential gene set combinations highlighted by SGFI might lead to interesting biological interpretations about the interplay of different gene sets. We expect that as more data becomes available, the confidence in these results will increase.

The use of PPI-based MCODE clusters as feature group inputs instead of our curated gene sets also created an excellent model for predicting LN from NRL in the blood (AUC = 0.956), implying that this pipeline can be applied to gene expression datasets by using unsupervised clustering methods to create gene sets. Additionally, using MCODE clusters led to the selection of genes that significantly overlapped with our curated *IFN* and *Mito Large Ribo* signatures. Therefore, using both our curated gene sets and unsupervised MCODE clusters led to the selection of genes with similar functions, validating our initial results and suggesting that these gene sets are meaningful in the development of LN.

It is not surprising that *IFN* was chosen as an important gene set in all classification problems, as the overproduction of IFN is a major hallmark of SLE.[39,40] When further interpreting the results from feature selection, we found that *IFN* signature genes were significantly enriched in SLE compared to CTL, I-SLE compared to CTL, active compared to I-SLE, LN compared to CTL, and SLE compared to RA, which corresponds to previous work describing the upregulation of Type I IFNs in SLE and SLE subtype patients.[39] The *IFN* gene set was chosen across all subsample iterations as the best single gene set to predict both I-SLE from CTL and active from I-SLE, pointing to its stability in predicting disease phenotype across different data architectures. Though SLE is extremely heterogeneous and patients with I-SLE do not show many clinical indications of disease, it appears that the *IFN* signature is stably enriched across most patients with I-SLE. This conclusion is consistent with previous studies that have detected the IFN gene signature in the blood of inactive lupus patients.[41] Additionally, the finding that *IFN* response genes can accurately predict disease activity corresponds to studies that have determined that *IFN* levels are positively correlated with SLEDAI.[42] The ability of the *IFN* gene signature to distinguish SLE from patients with RA was also in line with past studies that found SLE has significantly higher IFN activity compared to RA, and that IFN activity in patients with RA is no different than healthy controls.[42]

The pro-inflammatory cytokine TNF has also been well-documented to be linked to SLE, but the direction of this relationship is unclear. Whereas some studies have found that TNF expression is related to SLE susceptibility, others have found that TNF expression is protective against SLE.[43] Another recent study found that TNF was the best discriminator of SLE from CTL patients out of 26 biomarkers, and it positively correlated to SLEDAI.[44] Likewise, we found that the module of *TNF induced genes* was one of the best discriminators of both SLE from CTL and LN from CTL blood. In this case, using both traditional laboratory methods and bioinformatics results yielded the same results. In the datasets used in this study, in general the *TNF induced genes* signature was significantly enriched in patients with SLE compared to CTL and in patients with LN compared to CTL via GSVA. However, some TNF induced genes, such as *CD38, HP,* and *LGALS3BP*, were significantly upregulated, whereas others, such as *FCER2*, *INSIG1*, and *MAP3K4*, were generally significantly downregulated across datasets. Additionally,

TNF is the main therapeutic target in RA, so it is not surprising that it was selected as a main differentiator of RA and SLE.[45] Similarly to SLE vs. CTL and LN vs. CTL, some TNF induced genes in RA as compared to SLE were upregulated, such as *NR3C1*, *PTGS2*, and *EREG*, whereas others were downregulated, such as *MMP19*, *MRPS15*, and *MSC*. Despite this variability, it is clear that the expression levels of TNF induced genes are important predictors of disease status in SLE.

The role of the mitochondria in SLE and LN has been extensively studied. However, the role of the large subunit of the mitoribosome, specifically, has not been elucidated.[46] Defective mitophagy, or mitochondrial death, has been reported in SLE, and disease severity is thought to be driven by excess reactive oxygen species that result from defective mitophagy.[47,48] Additionally, extracellular oxidized mitochondrial DNA can induce Type I IFN signaling in lupus mouse models.[47] *Mito Large Ribo* was found to be de-enriched from SLE to CTL and from LN to NRL. No other mitochondria-related signatures were chosen across SGFI subsamples in SLE vs. CTL, whereas the small subunit of the mitoribosome (*Mito Small Ribo*) was also chosen in LN vs. NRL in some iterations. Therefore, there may be a specific role of the mitoribosome in the pathogenesis of SLE or it may be a specifically potent marker of mitochondrial dysfunction.

T cells have also been reported to contribute to SLE, as they amplify inflammation and help B cells generate autoantibodies.[49] The *Anergic/Activated T cells* gene signature consists of genes that have either inhibitory or activating effects on the immune system, depending on the context, and it was assessed because patients with SLE have dysfunctions in T cell anergy and activation.[50] The *Anergic/Activated T cells* was used in the final model classifying I-SLE and CTL samples. Although the GSVA signature was not significantly different between groups, half of the genes (*CD160, CD244, CTLA4,* and *KLRG1*) from *Anergic/Activated T cells* were downregulated, whereas the others (*HAVCR2, ICOS, LAG3,* and *PDCD1*) were upregulated in I-SLE. This might indicate a compensatory response to autoimmunity, in which the inactive disease state is maintained, but patients are still poised to return to an inflammatory state if the balance shifts to favor inflammatory signals.

Our approach exemplifies the power of ML in detecting subtle differences in blood gene expression between disease status groups that may point out new areas of research investigation. Some differences between gene sets were undetectable by GSVA, yet the radial SVM machine learning algorithm could uncover patterns by projecting the data into a higher feature space. Additionally, although in some cases GSVA scores were not different from class to class, oftentimes the individual genes within that gene set were differentially expressed. This approach of leveraging knowledge of gene sets with similar expression or function, while retaining the individual expression differences of each gene, ensures that we create an interpretable model without removing true biological variation in the genes within those sets.

Overall, we constructed five ML models classifying clinical phenotypes of SLE using expression of genes grouped by disease-relevant cell types or function that all performed with extremely high accuracy (AUCs >0.842). Additionally, we demonstrated that this ML pipeline can be applied to diseases with shared pathways and common symptoms, such as RA and SLE.[51] These models lay the framework for a potential approach to creating non-invasive diagnostic tests for SLE that may support clinical decisions and improve patient care, if more reproducible gene expression data becomes available. As demonstrated here, because the pipeline involves selecting molecular pathways that best predict disease status, the ML models are interpretable and point to future areas of research for SLE and SLE clinical phenotypes. We have focused on SLE, but the approach described in this study can be applied to other diseases or tissues to create an interpretable model that predicts disease status based on gene expression. Namely, it may be applied to other rheumatic autoimmune/inflammatory diseases with similarities to lupus, such as Sjögren's syndrome or dermatomyositis.

Limitations of the study

This study outlines a validated approach to predict disease status based on gene expression data and identifies potential pathways involved in SLE manifestations. However, there are limitations to this study. First, this method has not been externally validated, in that a completely independent, separately run gene expression dataset was not used as the test set. In order for our model to perform well on an unseen test set, both the train and test set needed to be comprised of a random sample of patients from all datasets, rather than using n-1 datasets in the train set and using the n^{th} dataset as the test set. One potential reason for this requirement is that ComBat does not completely remove the batch effect, and in the case of multiple datasets in the train and test set, the classifier can learn patterns specific to each dataset. Therefore, better normalization methods or more reproducible gene expression data would be important to translate this methodology to a diagnostic test. A leave-one-dataset-out approach may achieve higher accuracy using RNA-seq rather than microarray data, as RNA-seq has less technical variation across datasets and is typically more reproducible.[52] However, we did not have enough RNA-seq samples available for all SLE subtypes to test this hypothesis. Another potential limitation of the study relates to the known effect of standard of care medication on gene expression profiles.[20] To obviate this potential issue, we avoided the use of gene modules that were most impacted by medications and focused on those that were detected in most patients and more affected by patient intrinsic processes than by medication. Despite these caveats, our study simulates the potential performance of an interpretable diagnostic test based on gene expression data that is standardized and reproducible. Additionally, because we had additional data available with SLE and CTL samples, we were able to externally validate the "locked" model, which used the same genes and hyperparameters determined from the pipeline but was trained (70%) and tested (30%) on new data, implying that these genes were not just discriminatory in the original data but rather perform well in predicting SLE from CTL samples in other datasets as well.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108042.

ACKNOWLEDGMENTS

The work presented in this article was funded by a grant awarded to A.C.G. and P.E.L. of the RILITE Research Institute by the John and Marcia Goldman Foundation (jmgoldmanfoundation.org). We would like to thank the authors of previous studies for making their data publicly available and facilitating this study. We would also like to thank the team at AMPEL BioSolutions for providing their support and insights throughout the development of this work. The graphical abstract was created with biorender.com.

AUTHOR CONTRIBUTIONS

Conceptualization: E.L.L., A.R.D., and P.E.L. Methodology: E.L.L. Software: E.L.L. Investigation: E.L.L. Resources: A.R.D. and P.E.L. Data curation: A.R.D. and E.L.L. Writing-original draft: E.L.L. and A.R.D. Writing-review and editing: E.L.L., A.R.D., and P.E.L. Visualization: E.L.L. and A.R.D. Supervision: A.C.G. and P.E.L. Funding acquisition: A.C.G. and P.E.L.

DECLARATION OF INTERESTS

The authors declare that they have no conflict of interest.

Received: April 21, 2023
Revised: July 3, 2023
Accepted: September 21, 2023
Published: September 25, 2023

REFERENCES

1. Cui, M., Li, T., Yan, X., Wang, C., Shen, Q., Ren, H., Li, L., and Zhang, R. (2021). Blood Genomics Identifies Three Subtypes of Systemic Lupus Erythematosus: "iFN-High," "nE-High," and "mixed. Mediators Inflamm. *2021*, 6660164. https://doi.org/10.1155/2021/6660164.
2. Bradley, S.J., Suarez-Fueyo, A., Moss, D.R., Kyttaris, V.C., and Tsokos, G.C. (2015). T cell transcriptomes describe patient subtypes in systemic lupus erythematosus. PLoS One *10*, e0141171. https://doi.org/10.1371/journal.pone.0141171.
3. Wang, L., Yang, Z., Yu, H., Lin, W., Wu, R., Yang, H., and Yang, K. (2022). Predicting diagnostic gene expression profiles associated with immune infiltration in patients with lupus nephritis. Front. Immunol. *13*, 839197. https://doi.org/10.3389/fimmu.2022.839197.
4. Yones, S.A., Alva Annett, S., Diamanti, K., Holmfeldt, L., Fredrik Barrenäs, C., Jennifer Meadows, S., and Komorowski, J. (2021). Interpretable Machine Learning Identiies Paediatric Systemic Lupus Erythematosus Subtypes Based On Gene Expression Data. https://doi.org/10.21203/rs.3.rs-588542/v2.
5. Zhong, Y., Zhang, W., Hong, X., Zeng, Z., Chen, Y., Liao, S., Cai, W., Xu, Y., Wang, G., Liu, D., et al. (2022). Screening Biomarkers for Systemic Lupus Erythematosus Based on Machine Learning and Exploring Their Expression Correlations With the Ratios of Various Immune Cells. Front. Immunol. *13*, 873787. https://doi.org/10.3389/fimmu.2022.873787.
6. Cojocaru, M., Cojocaru, I.M., Silosi, I., and Doina Vrabie, C. (2011). Manifestations of Systemic Lupus Erythematosus. Maedica *6*.
7. Fava, A., and Petri, M. (2019). Systemic lupus erythematosus: Diagnosis and clinical management. J. Autoimmun. *96*, 1–13. https://doi.org/10.1016/j.jaut.2018.11.001.
8. Sebastiani, G.D., Prevete, I., Iuliano, A., and Minisola, G. (2016). The Importance of an Early Diagnosis in Systemic lupus Erythematosus. Isr. Med. Assoc. J. *18*, 212–215.
9. Fanouriakis, A., Tziolos, N., Bertsias, G., and Boumpas, D.T. (2021). Update in the diagnosis and management of systemic lupus erythematosus. Ann. Rheum. Dis. *80*, 14–25. https://doi.org/10.1136/annrheumdis-2020-218272.
10. Aringer, M., Costenbader, K., Daikh, D., Brinks, R., Mosca, M., Ramsey-Goldman, R., Smolen, J.S., Wofsy, D., Boumpas, D.T., Kamen, D.L., et al. (2019). 2019 European League Against Rheumatism/American College of Rheumatology Classification Criteria for Systemic Lupus Erythematosus. Arthritis Rheumatol. *71*, 1400–1412. https://doi.org/10.1002/art.40930.
11. Petri, M., Orbai, A.M., Alarcón, G.S., Gordon, C., Merrill, J.T., Fortin, P.R., Bruce, I.N., Isenberg, D., Wallace, D.J., Nived, O., et al. (2012). Derivation and validation of the systemic lupus international collaborating clinics classification criteria for systemic lupus

erythematosus. Arthritis Rheum. *64*, 2677–2686. https://doi.org/10.1002/art.34473.
12. Md Yusof, M.Y., and Vital, E.M. (2022). Early intervention in systemic lupus erythematosus: Time for action to improve outcomes and health-care utilization. Rheumatol. Adv. Pract. *6*, rkab106. https://doi.org/10.1093/rap/rkab106.
13. Haładyj, E., and Cervera, R. (2016). Do we still need renal biopsy in lupus nephritis? Reumatologia *54*, 61–66. https://doi.org/10.5114/reum.2016.60214.
14. Rovin, B.H., Parikh, S.V., and Alvarado, A. (2014). The kidney biopsy in lupus nephritis: Is it still relevant? Rheum. Dis. Clin. North Am. *40*, 537–552. ix. https://doi.org/10.1016/j.rdc.2014.04.004.
15. Rai, R., Chauhan, S.K., Singh, V.V., Rai, M., and Rai, G. (2016). RNA-seq analysis reveals unique transcriptome signatures in systemic lupus erythematosus patients with distinct autoantibody specificities. PLoS One *11*, e0166312. https://doi.org/10.1371/journal.pone.0166312.
16. Ceccarelli, F., Natalucci, F., Picciariello, L., Ciancarella, C., Dolcini, G., Gattamelata, A., Alessandri, C., and Conti, F. (2023). Application of Machine Learning Models in Systemic Lupus Erythematosus. Int. J. Mol. Sci. *24*, 4514. https://doi.org/10.3390/ijms24054514.
17. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. USA *116*, 22071–22080. https://doi.org/10.1073/pnas.1900654116.
18. Figgett, W.A., Monaghan, K., Ng, M., Alhamdoosh, M., Maraskovsky, E., Wilson, N.J., Hoi, A.Y., Morand, E.F., and Mackay, F. (2019). Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus. Clin. Transl. Immunol. *8*, e01093. https://doi.org/10.1002/cti2.1093.
19. Maleki, F., Ovens, K., Hogan, D.J., and Kusalik, A.J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. Front. Genet. *11*, 654. https://doi.org/10.3389/fgene.2020.00654.
20. Catalina, M.D., Bachali, P., Yeo, A.E., Geraci, N.S., Petri, M.A., Grammer, A.C., and Lipsky, P.E. (2020). Patient ancestry significantly contributes to molecular heterogeneity of systemic lupus erythematosus. JCI Insight *5*, e140380. https://doi.org/10.1172/jci.insight.140380.
21. Kingsmore, K.M., Bachali, P., Catalina, M.D., Daamen, A.R., Heuer, S.E., Robl, R.D., Grammer, A.C., and Lipsky, P.E. (2021). Altered expression of genes controlling metabolism characterizes the tissue response to immune injury in lupus. Sci. Rep. *11*, 14789. https://doi.org/10.1038/s41598-021-93034-w.
22. Au, Q., Herbinger, J., Stachl, C., Bischl, B., and Casalicchio, G. (2022). Grouped feature importance and combined features effect plot. Data Min. Knowl. Discov. *36*, 1401–1450. https://doi.org/10.1007/s10618-022-00840-5.
23. Lauwerys, B.R., Hachulla, E., Spertini, F., Lazaro, E., Jorgensen, C., Mariette, X., Haelterman, E., Grouard-Vogel, G., Fanget, B., Dhellin, O., et al. (2013). Down-regulation of interferon signature in systemic lupus erythematosus patients by active immunization with interferon α-kinoid. Arthritis Rheum. *65*, 447–456. https://doi.org/10.1002/art.37785.
24. Hu, Y., Carman, J.A., Holloway, D., Kansal, S., Fan, L., Goldstine, C., Lee, D., Somerville, J.E., Latek, R., Townsend, R., et al. (2018). Development of a Molecular Signature to Monitor Pharmacodynamic Responses Mediated by In Vivo Administration of Glucocorticoids. Arthritis Rheumatol. *70*, 1331–1342. https://doi.org/10.1002/art.40476.
25. Bienkowska, J., Allaire, N., Thai, A., Goyal, J., Plavina, T., Nirula, A., Weaver, M., Newman, C., Petri, M., Beckman, E., and Browning, J.L. (2014). Lymphotoxin-LIGHT pathway regulates the interferon signature in rheumatoid arthritis. PLoS One *9*, e112545. https://doi.org/10.1371/journal.pone.0112545.
26. Berry, M.P.R., Graham, C.M., McNab, F.W., Xu, Z., Bloch, S.A.A., Oni, T., Wilkinson, K.A., Banchereau, R., Skinner, J., Wilkinson, R.J., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. Nature *466*, 973–977. https://doi.org/10.1038/nature09247.
27. Banchereau, R., Jordan-Villegas, A., Ardura, M., Mejias, A., Baldwin, N., Xu, H., Saye, E., Rossello-Urgell, J., Nguyen, P., Blankenship, D., et al. (2012). Host immune transcriptional profiles reflect the variability in clinical disease manifestations in patients with staphylococcus aureus infections. PLoS One *7*, e34390. https://doi.org/10.1371/journal.pone.0034390.
28. Houssiau, F.A., Thanou, A., Mazur, M., Ramiterre, E., Gomez Mora, D.A., Misterska-Skora, M., Perich-Campos, R.A., Smakotina, S.A., Cerpa Cruz, S., Louzir, B., et al. (2020). IFN-α kinoid in systemic lupus erythematosus: results from a phase IIb, randomised, placebo-controlled study. Ann. Rheum. Dis. *79*, 347–355. https://doi.org/10.1136/annrheumdis-2019-216379.
29. Bigler, J., Boedigheimer, M., Schofield, J.P.R., Skipp, P.J., Corfield, J., Rowe, A., Sousa, A.R., Timour, M., Twehues, L., Hu, X., et al. (2017). A severe asthma disease signature from gene expression profiling of peripheral blood from U-BIOPRED cohorts. Am. J. Respir. Crit. Care Med. *195*, 1311–1320. https://doi.org/10.1164/rccm.201604-0866OC.
30. Wither, J.E., Prokopec, S.D., Noamani, B., Chang, N.H., Bonilla, D., Touma, Z., Avila-Casado, C., Reich, H.N., Scholey, J., Fortin, P.R., et al. (2018). Identification of a neutrophil-related gene expression signature that is enriched in adult systemic lupus erythematosus patients with active nephritis: Clinical/pathologic associations and etiologic mechanisms. PLoS One *13*, e0196117. https://doi.org/10.1371/journal.pone.0196117.
31. Ducreux, J., Houssiau, F.A., Vandepapelière, P., Jorgensen, C., Lazaro, E., Spertini, F., Colaone, F., Roucairol, C., Laborie, M., Croughs, T., et al. (2016). Interferon α kinoid induces neutralizing anti-interferon α antibodies that decrease the expression of interferon-induced and B cell activation associated transcripts: Analysis of extended follow-up data from the interferon α kinoid phase I/II study. Rheumatology *55*, 1901–1905. https://doi.org/10.1093/rheumatology/kew262.
32. Hou, G. (2022). Expression data of whole blood samples from SLE patients and controls. BioStudies.
33. Tasaki, S., Suzuki, K., Kassai, Y., Takeshita, M., Murota, A., Kondo, Y., Ando, T., Nakayama, Y., Okuzono, Y., Takiguchi, M., et al. (2018). Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. Nat. Commun. *9*, 2755. https://doi.org/10.1038/s41467-018-05044-4.
34. Feldner-Busztin, D., Firbas Nisantzis, P., Edmunds, S.J., Boza, G., Racimo, F., Gopalakrishnan, S., Limborg, M.T., Lahti, L., and De Polavieja, G.G. (2023). Dealing with dimensionality: the application of machine learning to multi-omics data. Data Text Min. *39*, 1–8. https://doi.org/10.5281/zenodo.7361807.
35. Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A.M., Acs, P., Turner, J., Anguiano, E., et al. (2016). Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. Cell *165*, 551–565. https://doi.org/10.1016/j.cell.2016.03.008.
36. Kotliarov, Y., Sparks, R., Martins, A.J., Mulè, M.P., Lu, Y., Goswami, M., Kardava, L., Banchereau, R., Pascual, V., Biancotto, A., et al. (2020). Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. Nat. Med. *26*, 618–629. https://doi.org/10.1038/s41591-020-0769-8.
37. Nakano, M., Ota, M., Takeshima, Y., Iwasaki, Y., Hatano, H., Nagafuchi, Y., Itamiya, T., Maeda, J., Yoshida, R., Yamada, S., et al. (2022). Distinct transcriptome architectures underlying lupus establishment and exacerbation. Cell *185*, 3375–3389.e21. https://doi.org/10.1016/j.cell.2022.07.021.
38. Kegerreis, B., Catalina, M.D., Bachali, P., Geraci, N.S., Labonte, A.C., Zeng, C., Stearrett, N., Crandall, K.A., Lipsky, P.E., and Grammer, A.C. (2019). Machine learning approaches to predict lupus disease activity from gene expression data. Sci. Rep. *9*, 9617. https://doi.org/10.1038/s41598-019-45989-0.
39. Rönnblom, L., and Leonard, D. (2019). Interferon pathway in SLE: One key to unlocking the mystery of the disease. Lupus Sci. Med. *6*, e000270. https://doi.org/10.1136/lupus-2018-000270.
40. Elkon, K.B., and Stone, V.V. (2011). Type I interferon and systemic lupus erythematosus. J. Interferon Cytokine Res. *31*, 803–812. https://doi.org/10.1089/jir.2011.0045.
41. Catalina, M.D., Bachali, P., Geraci, N.S., Grammer, A.C., and Lipsky, P.E. (2019). Gene expression analysis delineates the potential roles of multiple interferons in systemic lupus erythematosus. Commun. Biol. *2*, 140. https://doi.org/10.1038/s42003-019-0382-x.
42. Miyachi, K., Iwamoto, T., Kojima, S., Ida, T., Suzuki, J., Yamamoto, T., Mimura, N., Sugiyama, T., Tanaka, S., Furuta, S., et al. (2023). Relationship of systemic type I interferon activity with clinical phenotypes, disease activity, and damage accrual in systemic lupus erythematosus in treatment-naive patients: a retrospective longitudinal analysis. Arthritis Res. Ther. *25*, 26. https://doi.org/10.1186/s13075-023-03010-0.
43. Ghorbaninezhad, F., Leone, P., Alemohammad, H., Najafzadeh, B., Nourbakhsh, N.S., Prete, M., Malerba, E., Saeedi, H., Tabrizi, N.J., Racanelli, V., and Baradaran, B. (2022). Tumor necrosis factor-α in systemic lupus erythematosus: Structure, function and therapeutic implications (Review). Int. J. Mol. Med. *49*, 43. https://doi.org/10.3892/ijmm.2022.5098.
44. Idborg, H., Eketjäll, S., Pettersson, S., Gustafsson, J.T., Zickert, A., Kvarnström, M.,

Oke, V., Jakobsson, P.J., Gunnarsson, I., and Svenungsson, E. (2018). TNF-α and plasma albumin as biomarkers of disease activity in systemic lupus erythematosus. Lupus Sci. Med. *5*, e000260. https://doi.org/10.1136/lupus-2018-000260.

45. Farrugia, M., and Baron, B. (2016). The role of TNF-α in rheumatoid arthritis: a focus on regulatory T cells. J. Clin. Transl. Res. *2*, 84–90. https://doi.org/10.18053/jctres.02.201603.005.
46. Galvan, D.L., Green, N.H., and Danesh, F.R. (2017). The hallmarks of mitochondrial dysfunction in chronic kidney disease. Kidney Int. *92*, 1051–1057. https://doi.org/10.1016/j.kint.2017.05.034.
47. Zhang, C.X., Wang, H.Y., Yin, L., Mao, Y.Y., and Zhou, W. (2020). Immunometabolism in the pathogenesis of systemic lupus erythematosus. J. Transl. Autoimmun. *3*, 100046. https://doi.org/10.1016/j.jtauto.2020.100046.
48. Quintero-González, D.C., Muñoz-Urbano, M., and Vásquez, G. (2022). Mitochondria as a key player in systemic lupus erythematosus. Autoimmunity *55*, 497–505. https://doi.org/10.1080/08916934.2022.2112181.
49. Suárez-Fueyo, A., Bradley, S.J., and Tsokos, G.C. (2016). T cells in Systemic Lupus Erythematosus. Curr. Opin. Immunol. *43*, 32–38. https://doi.org/10.1016/j.coi.2016.09.001.
50. Banica, L.M., Besliu, A.N., Pistol, G.C., Stavaru, C., Vlad, V., Predeteanu, D., Ionescu, R., Stefanescu, M., and Matache, C. (2016). Dysregulation of anergy-related factors involved in regulatory T cells defects in Systemic Lupus Erythematosus patients: Rapamycin and Vitamin D efficacy in restoring regulatory T cells. Int. J. Rheum. Dis. *19*, 1294–1303.
51. Toro-Domínguez, D., Carmona-Sáez, P., and Alarcón-Riquelme, M.E. (2014). Shared signatures between rheumatoid arthritis, systemic lupus erythematosus and Sjögren's syndrome uncovered through gene expression meta-analysis. Arthritis Res. Ther. *16*. https://doi.org/10.1186/s13075-014-0489-x.
52. Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One *9*, e78644. https://doi.org/10.1371/journal.pone.0078644.
53. Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307–315. https://doi.org/10.1093/bioinformatics/btg405.
54. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.
55. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883. https://doi.org/10.1093/bioinformatics/bts034.
56. Huang, S., Cai, N., Pacheco, P.P., Narrandes, S., Wang, Y., and Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics *15*, 41–51. https://doi.org/10.21873/cgp.20063.
57. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z.M. (2016). mlr: Machine Learning in R. J. Mach. Learn. Res. *17*, 1–5.
58. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinf. *14*, 7.
59. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452. https://doi.org/10.1093/nar/gku1003.
60. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Microarray data	Gene Expression Omnibus	GEO: GSE22098
Microarray data	Gene Expression Omnibus	GEO: GSE29536
Microarray data	Gene Expression Omnibus	GEO: GSE61635
Microarray data	Gene Expression Omnibus	GEO: GSE39088
Microarray data	Gene Expression Omnibus	GEO: GSE69683
Microarray data	Gene Expression Omnibus	GEO: GSE185047
Microarray data	Gene Expression Omnibus	GEO: GSE72326
Microarray data	BioStudies	BioStudies: E-MTAB-11191
Microarray data	Gene Expression Omnibus	GEO: GSE72747
Microarray data	Gene Expression Omnibus	GEO: GSE99967
Microarray data	Gene Expression Omnibus	GEO: GSE45291
Microarray data	Gene Expression Omnibus	GEO: GSE93272
Microarray data	Gene Expression Omnibus	GEO: GSE110169
Microarray data	Gene Expression Omnibus	GEO: GSE110174

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the lead contact, Emily Leventhal (emily.leventhal@ampelbiosolutions.com).

Materials availability

This study did not generate any new materials.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data pre-processing

The publicly available microarray gene expression datasets used for each model are listed in Table S2A (SLE vs. CTL blood), Table S3A (I-SLE vs. CTL blood), Table S4A (active SLE vs. I-SLE), Table S5A (LN vs. CTL blood), Table S6A (LN vs. non-renal lupus [NRL] blood), and Table S7A (RA vs. SLE). We used all microarray datasets found in GEO that were collected using either the Affymetrix or Illumina platforms and had samples from either cohort for each classification problem. I-SLE was defined as Systemic Lupus Erythematosus Disease Activity (SLEDAI) < 6 and active SLE was defined as SLEDAI ≥ 6. When we had information on nephritis, we excluded nephritis patients from the SLE vs. CTL model. For the LN vs. NRL model, we chose to only include GEO: GSE185047 and GEO: GSE110174 as they were the only datasets with SLE patients confirmed as non-nephritic. For the RA vs. SLE model, we split GEO: GSE45291 into two batches given the apparent batch effect between SLE and RA samples implying the samples were run at different times. We additionally included GEO: GSE110174 in the RA vs. SLE model to balance the number of SLE samples.

For each model, we combined microarray datasets across different platforms. Figure 1 shows the method for normalizing, merging, and batch-correcting microarray data. For each model, every dataset was normalized independently with a quantile normalization method. For Affymetrix chip microarray datasets, we applied the multichip average (RMA) pre-processing method from the rma package to carry out background correction, quantile normalization, and probe summarization to each raw microarray dataset (CEL files).[53] For Illumina BeadChip microarray datasets, we applied the neqc() function from the limma package, which performs background correction using negative control

probes and quantile normalization using negative and positive control probes.[54] In the case of GEO: GSE72326, the quantile normalized data was already available on GEO.

Genes with an interquartile range (IQR) of 0 were filtered out, which includes genes not expressed over the detection level of the microarray. In the case of duplicate gene symbols, we kept the probe with a higher IQR across samples.

We merged the datasets by gene symbol. To adjust for batch effects, we used the ComBat method from the sva package on the merged dataset.[55] We then split the data into train (70%) and test (30%) sets with class stratification. After splitting the data, we standardized the train and test set separately according to the following:

$$x'_{test} = \frac{x_{test} - \overline{x}_{train}}{\sigma_{train}}$$

$$x'_{train} = \frac{x_{train} - \overline{x}_{train}}{\sigma_{train}}$$

We standardized the test set based on the mean and the standard deviation of the train set because the train and test set come from the same population, and this would simulate testing an individual sample in which the mean and standard deviation of the test set are not known.

Support vector machine (SVM)

To classify samples into disease status, we implemented a support vector machine (SVM) model because of its previous success in predicting disease status from gene expression data and its effectiveness in classifying high-dimensional data in which the number of features is greater than the number of samples.[56] We employed the Radial Basis Function (RBF) kernel as a non-linear transformation to convert the features to a higher dimensional space.[56]

Feature selection

To select the genes to use in the SVM model, we employed the SGFI algorithm implemented in R on the train set.[22] SGFI aims to find well-performing combinations of feature groups. SGFI starts from the null or 'featureless' model, and then sequentially adds the next best feature group in terms of leave-one-group-in importance (LOGI) until no further improvement in mean misclassification error (MMCE) over an improvement threshold, δ, is achieved. We set δ to 0.0001. Here, the genes are the features, and the feature groups are pre-defined gene sets. The gene sets we used as inputs were 46 curated groups of genes previously reported to identify immune pathways, metabolic pathways, and immune/inflammatory cell types.[20,21] To account for the potential effects of glucocorticoid treatment on immune cell gene expression signatures, gene sets for low-density granulocytes (LDGs) and neutrophils were excluded from the analysis.

Because LOGI is a refitting-based method requiring retraining the model to determine the expected gain from adding in a feature group, the results of SGFI depend on the learner used. We used the default hyperparameters (degree = 3, cost = 1) of the SVM classifier implemented in the mlr package, 'classif.svm,' with a radial (RBF) kernel during feature selection.[57] To account for variability introduced by the model, we used repeated subsampling (67%) with 100 repetitions with an inner resampling approach of 10-fold cross validation.

We extracted all combinations that were chosen in 5 or more of the 100 subsampling repetitions as possible feature sets. This generated a list of possible feature group combinations to use in our final model (Figure 1).

Hyperparameter tuning and 10-fold cross validation on the train set

For each combination of feature groups, we tuned the hyperparameters, Cost (C) and gamma (γ), using tuneParams() of the mlr package. We used the following parameter sets:

- C: $\{C \in \mathbf{R} \mid 2^{-5} < C < 2^{15}\}$
- γ: $\{\gamma \in \mathbf{R} \mid 0 < \gamma < 0.01\}$

After tuning C and γ, we then tuned the threshold, or the probability above which samples are classified as the positive (disease) class to adjust for the class imbalance. To determine the best threshold, we did a search of the set {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9} and identified the threshold with the best geometric mean. We repeated this with 50 different train (70%) and validation (30%) sets, each being a subset of the original train set, and took the average best threshold of all 50 subsamples as the final threshold used.

$$Geometric\ mean = \sqrt{sensitivity \times specificity}$$

10-fold cross validation was then run on each tuned model with the optimal hyperparameters. In 10-fold cross validation, the train set is divided into 10 parts, and the model is trained using nine of the folds and tested on one of the folds, and this process is repeated 10 times. We recorded the F1 score, F1 standard deviation across the folds, accuracy, sensitivity, specificity, and area under the curve (AUC) for each

model. The best model was chosen from these tuned models by selecting the model with the highest F1 score (rounded to 3 decimal places) and lowest F1 standard deviation (if there was a tie).

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Evaluation of final model on the test set

The best model by 10-fold cross validation on the train set was then evaluated on the test set. The F1 score, F1 standard deviation, accuracy, sensitivity, specificity, and AUC on the test set were recorded.

QUANTIFICATION AND STATISTICAL ANALYSIS

Gene set variance analysis (GSVA)

We used the GSVA package from R/Bioconductor (v1.44.5) as a non-parametric, unsupervised method to estimate the enrichment of gene sets in microarray data.[58] The inputs for GSVA were log2 expression values from each dataset and curated gene sets used in the final model (Table S1). After running GSVA on each dataset, the GSVA values from each dataset were combined. P values and visualizations of GSVA were generated via GraphPad Prism 9.4.0 software. Comparisons between GSVA scores between classes were calculated using an unpaired t test. The number of samples used for each classification problem can be found in the supplemental tables.

Differential expression

Differential expression analysis of the genes used in the final model was carried out in R using the limma package.[54] Significance was considered as an adjusted p-value of less than 0.2.

Network analysis and visualization

To create protein-protein interaction (PPI) clusters for the LN vs. NRL datasets, the top 2000 genes with the highest variance across samples were extracted from the merged dataset and inputted into Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (V11.5).[59] The STRING output was inputted into Cytoscape (V3.8.2). with the clusterMaker plugin, and clusters were generated with Molecular Complex Detection (MCODE) clustering algorithm within clusterMaker.[60] The clusters were annotated with the top significant curated gene sets by a Fisher's exact test, $p < 0.05$.