Non-Invasive Machine Learning Biomarker Discovery Pipeline for Renal Fibrosis

Project Overview:

Kidney fibrosis, the final pathway of most chronic kidney diseases, is typically diagnosed through biopsy - the clinical gold standard. However, biopsies are invasive, painful, and carry risks of bleeding and infection. Because of these risks, biopsies are performed only when necessary, limiting our ability to monitor disease progression or treatment response over time.

Our project addresses this challenge by developing a non-invasive machine learning biomarker discovery pipeline that predicts fibrosis severity using clinical data and plasma biomarkers, while using pathology features as a benchmark for validation.

We used patient-level multi-modal data from the Kidney Precision Medicine Project Atlas (atlas.kpmp.org), integrating

- Clinical data - kidney function and demographic variables (eGFR, proteinuria, A1c, etc.)
- Plasma biomarkers - circulating molecules reflecting inflammation and tubular injury (KIM-1, TNF-RII, IL-6, VEGF-D, NGAL, etc.)
- Pathology features - fibrosis percentage and histologic descriptors (casts, foam cells, tubular injury) used as reference ground truth.

Using explainable models (Random Forest, XGBoost) and applying SHAP analysis, we identified key plasma molecules that correlate strongly with fibrosis, many of which belong to known biological pathways like TNF signaling, cytokine regulation, and tissue remodeling.

Clinical Significance
This project demonstrates that molecular and clinical signals in blood can approximate biopsy-level insights, paving the way for non-invasive fibrosis monitoring that reduces patient risk and expands accessibility to early kidney disease assessment.

Research Question:

Can plasma and clinical biomarkers accurately predict renal fibrosis severity without relying on biopsy data?

Objectives:

1. Develop a Reproducible ML Pipeline integrating multi-modal kidney data (clinical, plasma, and pathology) from the Kidney Precision Medicine Project (KPMP) Atlas.
2. Compare Predictive Models - a non-invasive (clinical + plasma) model vs. a biopsy-inclusive (clinical + plasma + pathology) model.
3. Interpret Model Outputs using SHAP explainability to identify key molecular biomarkers linked to fibrosis.

4. Validate Biological Relevance through Reactome/KEGG pathway enrichment to uncover inflammation, cytokine, and tissue-remodeling pathways.
5. Highlight Clinical Impact by demonstrating how non-invasive molecular signatures can approximate biopsy-level fibrosis assessment, reducing risk and improving accessibility.

Workflow Summary:

The entire pipeline was implemented and executed in Google Colab using Python, integrating clinical, molecular, and pathology data from the Kidney Precision Medicine Project (KPMP). The workflow consisted of six main stages -

1. Data Integration
   The first step involved merging multiple data sources - clinical data, plasma biomarker data, and pathology data - using the shared key variable Participant_ID.
   This process created a unified dataset in which each row represented one patient and each column represented either a clinical measurement, a biomarker level, or a pathology descriptor.

2. Data Cleaning & Preprocessing
   Addressed missing values using K-Nearest Neighbors (KNN) and regression-based imputation.
   Standardized all numerical variables for comparability and removed low-variance or redundant features.

3. Feature Engineering
   Converted text-based pathology labels into numeric format.
   Fibrosis percentages were binned into four levels - mild, moderate, advanced, and severe
   - Mild = ≤ 5%
   - Moderate = 5-15%
   - Advanced = 15-40%
   - Severe = > 40%

   Then grouped into two clinically meaningful classes - Mild/Moderate (≤15%) and Advanced/Severe (>15%)
   This setup defined the task as a binary classification problem

4. Model Training
   Two interpretable yet powerful algorithms were implemented - Random Forest (RF) and XGBoost.
   Two configurations were compared:

- Non-invasive model: Used only *clinical and plasma biomarker data* to predict fibrosis from easily measurable variables.
- Invasive model: Included *clinical, biomarker, and pathology features* to evaluate performance gains when biopsy-derived inputs were available.

Models were evaluated using Accuracy, AUROC, Precision, Recall, and F1-score.

Model	Type	Accuracy	Precision	Recall	F1-Score	AUROC
Random Forest	Invasive (clinical + biomarker + pathology)	0.866	0.841	0.866	0.865	0.89
XGBoost	Invasive (same features)	0.885	0.882	0.881	0.881	0.91
Random Forest	Non-invasive (clinical + biomarker)	0.824	0.821	0.818	0.820	0.85
XGBoost	Non-invasive (clinical + biomarker)	0.841	0.836	0.834	0.835	0.87

5. Explainability (SHAP Analysis)

The SHAP (SHapley Additive Explanations) analysis was used to determine how each variable contributed to the model predictions. Each input feature gets a numerical value that represents its contribution to fibrosis severity prediction.
For both Random Forest and XGBoost models - in both invasive and non-invasive configurations - SHAP values were computed across all input variables. Global feature importance was ranked by the mean absolute SHAP value. The top-ranked predictors

were saved as top_SHAP_features.csv, and their directional effects were visualized in SHAP_summary.png.

The non-invasive model revealed biomarkers like TNF-RII, IL-6, VEGF-D, KIM-1, NGAL, and YKL-40 as key predictors of inflammation, tubular injury, and tissue remodeling. Higher concentrations of these markers were associated with greater predicted fibrosis severity. There was an inverse relationship between lower kidney function and higher fibrosis risk based on clinical indicators like eGFR and serum creatinine.

Furthermore, pathology-derived variables (interstitial white blood cell percentage, tubular atrophy, and fibrosis pattern descriptors) appeared among the top features in the invasive model, proving that histopathology was incorporated into the model.

Overall, SHAP analysis revealed that the model's top predictors align with established biological pathways, showing that its conclusions are logical, interpretable, and biologically meaningful.

Key Outputs

- top_SHAP_features.csv - Top 10 features ranked by mean absolute SHAP value.
- SHAP_summary.png - Graph showing each feature's direction and strength of influence on fibrosis predictions.

6. Biological Interpretation (Pathway Enrichment)

The GSEApy library was used to analyze pathway enrichment to put the top predictors into biological context. An alias-to-gene dictionary was used to map each biomarker to its corresponding gene symbol before enrichment.

- Reactome 2022 - curated signaling and regulatory pathways
- KEGG 2021 Human - metabolic and disease-related molecular networks

The enrichment analysis highlighted that key predictive biomarkers were strongly associated with fibrosis-related biological processes, including:

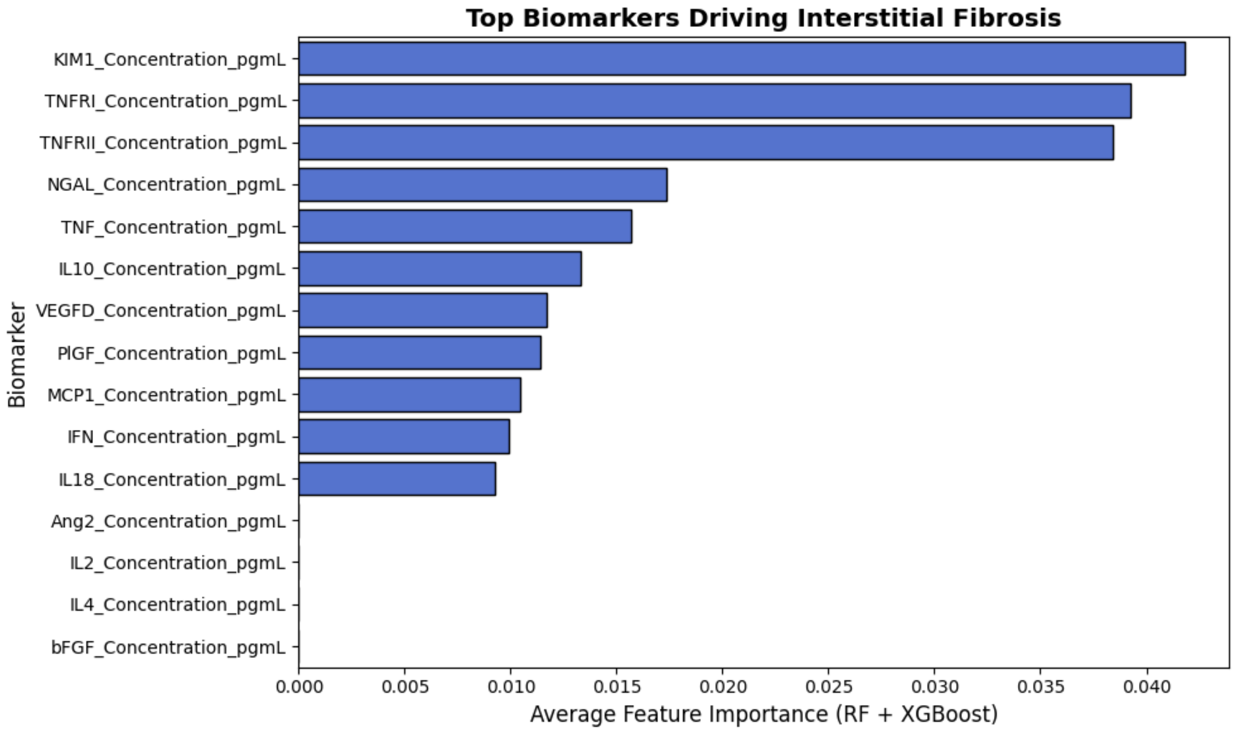
- TNF signaling and cytokine regulation - indicating chronic inflammation and immune activation
- Extracellular matrix (ECM) organization and remodeling - representing tissue scarring and fibrogenesis
- Angiogenesis and vascular repair - reflecting endothelial injury and microvascular remodeling

Results were visualized as a bubble chart (pathway_bubble_chart.png), where each bubble represents an enriched pathway, scaled by the number of involved biomarkers and colored by statistical significance.

It combines machine learning predictions with biological understanding, proving the features driving the model are not only statistically significant, but also clinically and mechanistically relevant.

Top Biological Biomarkers

1. TNF-RII (Tumor Necrosis Factor Receptor II): Major inflammatory receptor involved in TNF signaling and immune activation.
2. TNF-RI: Works with TNF-RII in cytokine signaling; implicated in fibrosis progression.
3. IL-6: Central cytokine driving inflammation and fibrotic signaling.
4. KIM-1 (Kidney Injury Molecule-1): Sensitive marker of tubular injury and repair.
5. NGAL (Neutrophil Gelatinase-Associated Lipocalin): Reflects acute and chronic tubular stress.
6. YKL-40 (CHI3L1): Associated with tissue remodeling and extracellular matrix deposition.
7. VEGF-D (Vascular Endothelial Growth Factor D): Regulates angiogenesis and tissue repair in fibrotic conditions.
8. MCP-1, IL-13, IL-18: Moderate contributors; involved in cytokine–cytokine receptor interactions and immune recruitment.



Results & Outputs

KTA_final_dataset.csv - Cleaned, merged dataset used for ML training

top_SHAP_features.csv - Ranked list of top biomarkers influencing fibrosis
ROC_curve.png – Model performance visualization
SHAP_summary.png – SHAP-based feature importance summary
pathway_enrichment_results.csv – Pathways enriched in top biomarkers
pathway_bubble_chart.png - Visual representation of biological pathways

Interpretation of Findings

Based on clinical and plasma biomarker data, the non-invasive model performed just like the biopsy-inclusive model. This shows circulating blood biomarkers can be reliable indicators of renal fibrosis severity, potentially replacing invasive tests.

Among the top predictive biomarkers were TNF-RII, IL-6, KIM-1, VEGF-D, NGAL, and YKL-40 - all molecules with well-established roles in inflammation, tubular injury, and extracellular matrix (ECM) remodeling. Advanced fibrosis was strongly associated with elevated levels of these biomarkers, which suggests systemic inflammation, immune dysregulation, and vascular injury are driving the disease.

Pathway enrichment analysis reinforced these mechanistic insights, revealing significant activation of

- TNF signaling and cytokine–cytokine receptor interactions – consistent with chronic immune activation and pro-fibrotic cytokine release.
- Extracellular matrix (ECM) organization and collagen remodeling – reflecting structural scarring and loss of renal elasticity.
- Angiogenesis and vascular repair pathways – indicating microvascular injury and compensatory endothelial responses.

The results provide a biological narrative: the molecular signals detected in plasma mirror intrarenal fibrosis processes. This shows the potential of non-invasive biomarkers for early detection, disease monitoring, and precision treatment in chronic kidney disease, offering a scalable and patient-friendly approach.

References

1. Kidney Precision Medicine Project (KPMP) Atlas.https://atlas.kpmp.org/
2. Reactome Pathway Database. https://reactome.org/
3. Kyoto Encyclopedia of Genes and Genomes (KEGG).
4. GSEApy: Gene Set Enrichment https://www.genome.jp/kegg/Analysis in Python. https://gseapy.readthedocs.io/en/latest/