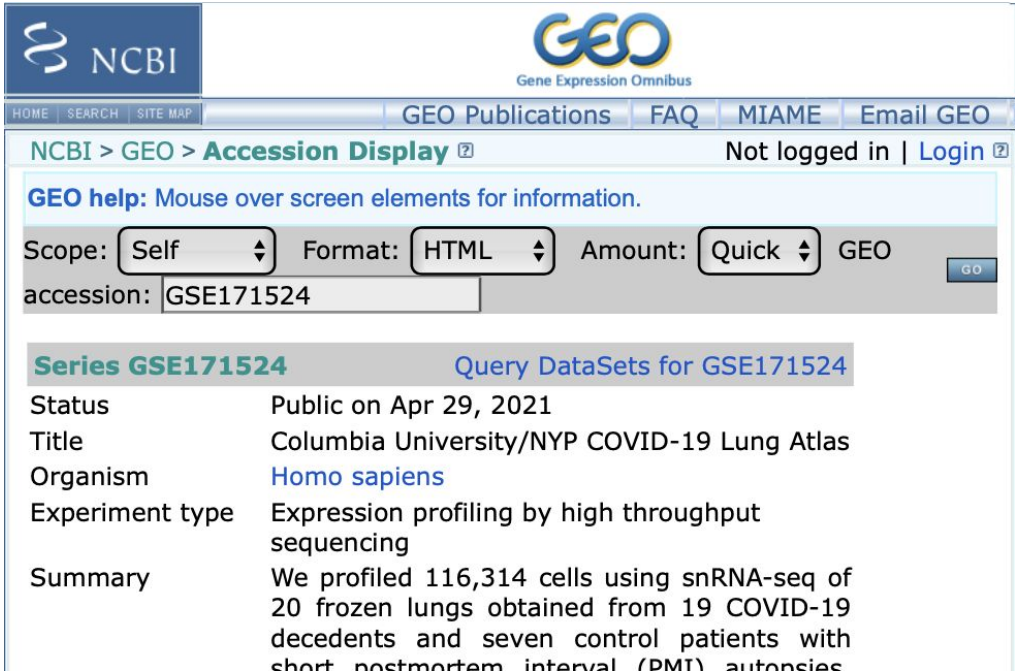
Using Supervised Learning for Single-Cell Analysis in Disease Identification

Data Gathering and Preprocessing

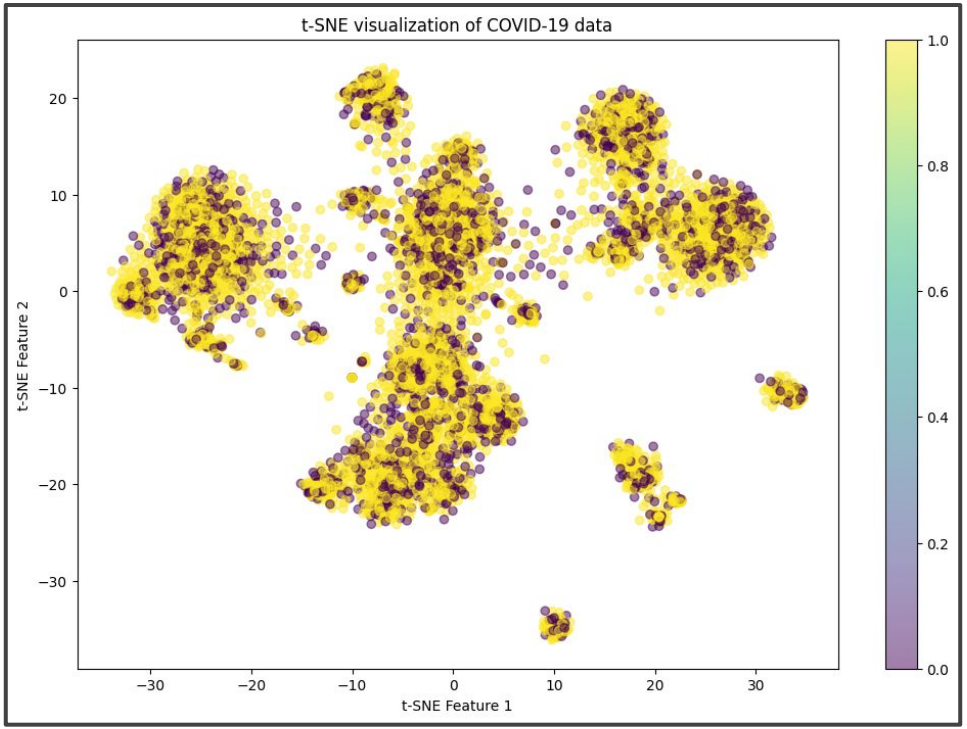
- Gathered Lung data on various cell types
- Converted the data into numeric format
- Clean and normalized the data
- Handled Missing values



```r
# Initialize a new data table for storing numeric converted data
exp_data_numeric <- data.table(exp_data)
rownames(exp_data_numeric) <- rownames(exp_data)
# Convert all columns of exp_data_numeric to numeric, coercing non-numeric entries to NA
for (col_name in names(exp_data_numeric)) {
  # Attempt to convert each column to numeric; assign NA where conversion fails
  exp_data_numeric[, (col_name) := as.numeric(get(col_name))]
}

# Check for any columns that have become entirely NA (indicating conversion failures)
na_columns <- sapply(exp_data_numeric, function(x) all(is.na(x)))
if (any(na_columns)) {
  cat("The following columns couldn't be converted to numeric and contain only NAs:\n")
  print(names(exp_data_numeric)[na_columns])
}
```

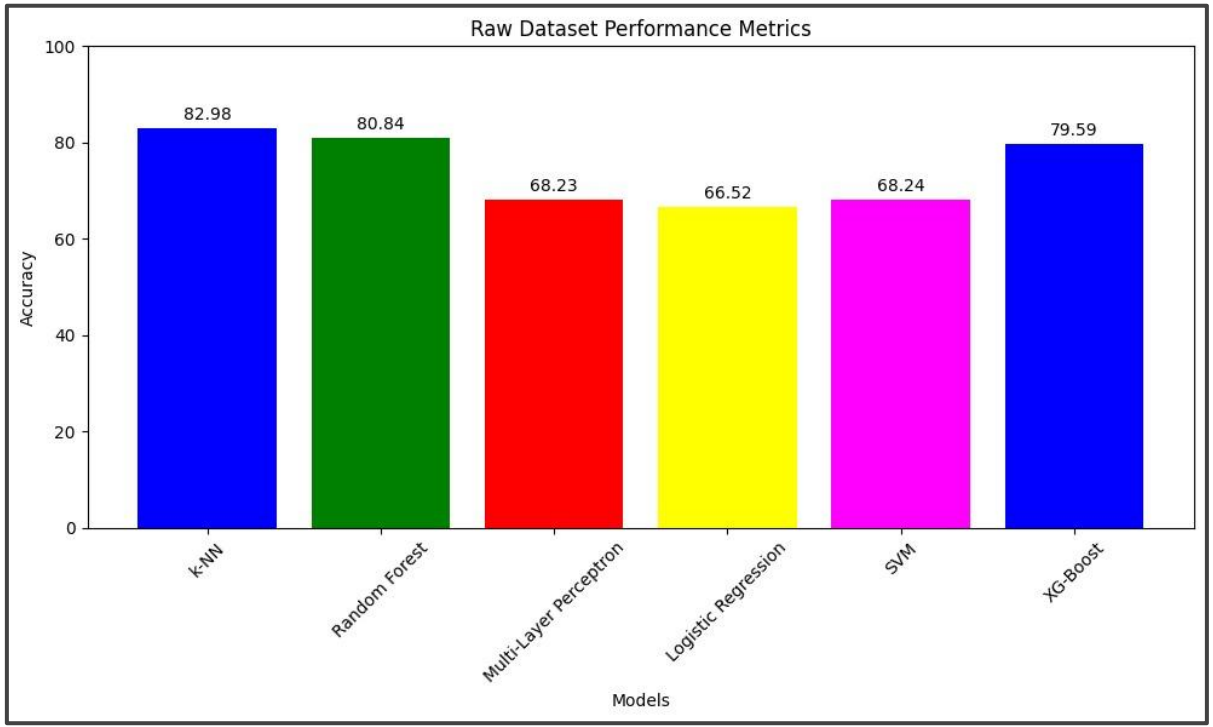
t-SNE Visualisation of Data



Initial ML Model On Raw Data

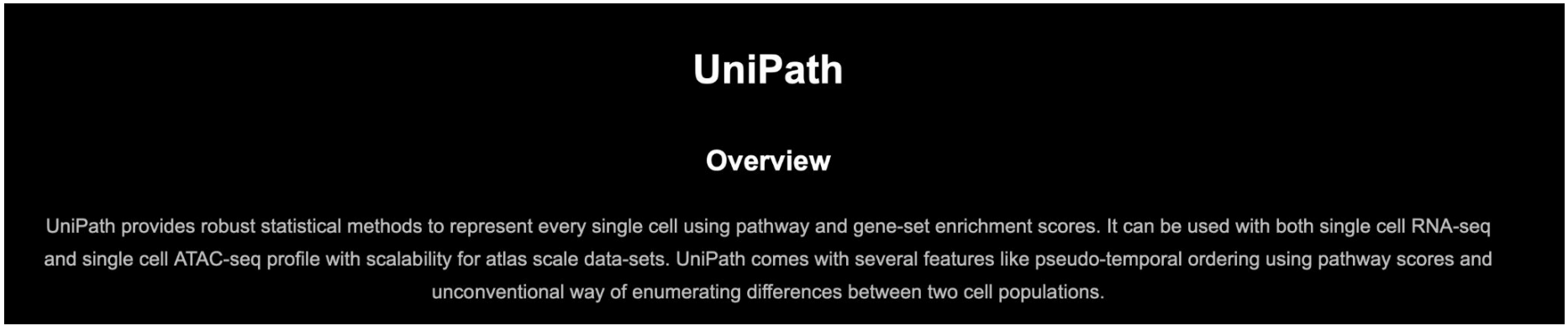
1. Chosen a machine learning Model
2. Trained the model on the raw data
3. Assessed the model performance using metrics like accuracy, precision and recall.

Raw Dataset Performance Metrics



Utilizing UniPath for Scoring

- Leverage UniPath to computer pathway activity scores
- Applied to lung data set to Obtain Pathway Scores for each Sample
- Helped in providing biological insignt into the data

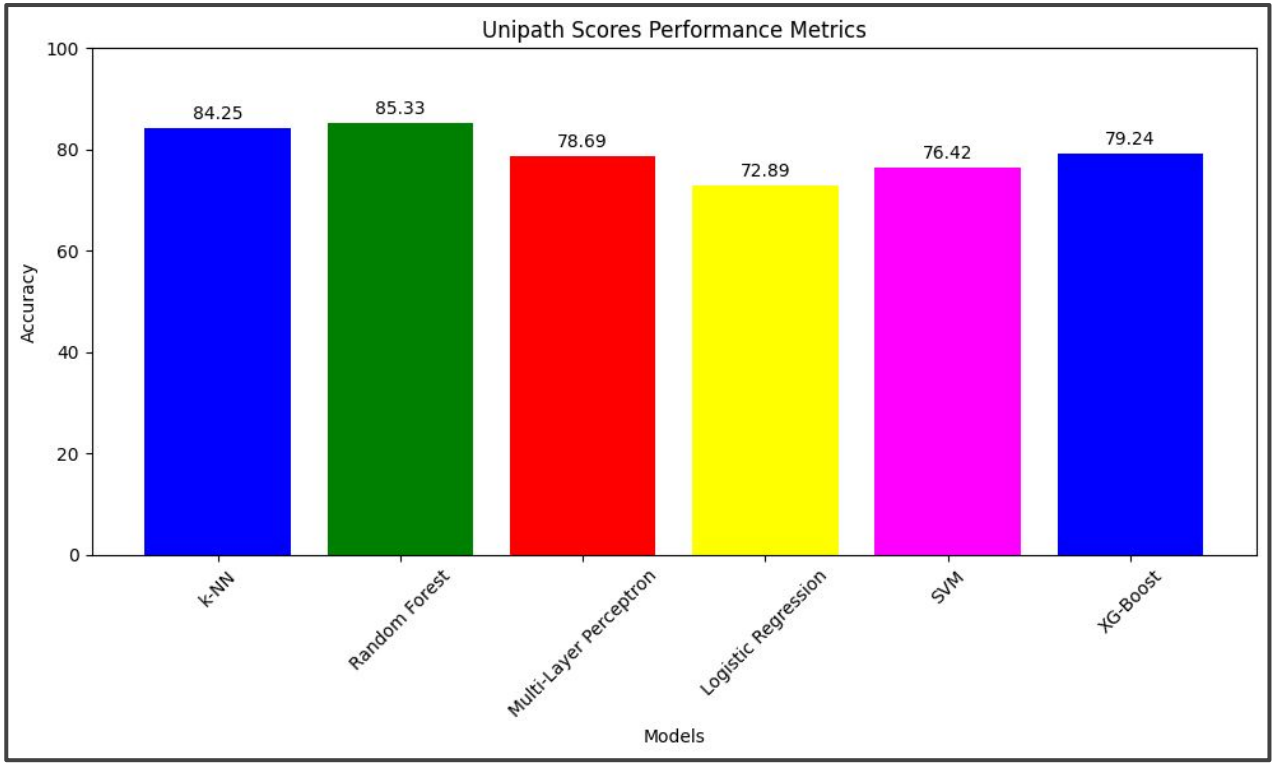


Ref:https://reggenlab.github.io/UniPathWeb/index.html

Running ML Model On UniPath Scores

- Used Pathway Activity scores obtained from uniPath as input features
- Ensured that dataset is properly split into training and testing sets
- Assessed the model performance using metrics like accuracy, precision, recall, and F1-score

Unipath Scores Performance Metrics



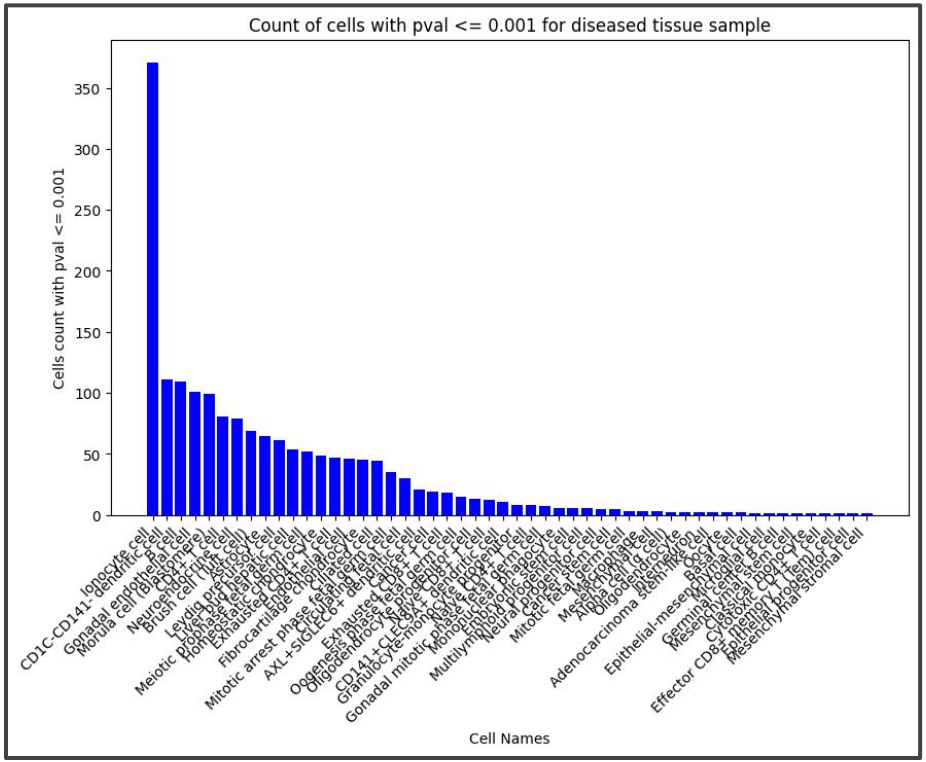
Identifying Significant P-Values

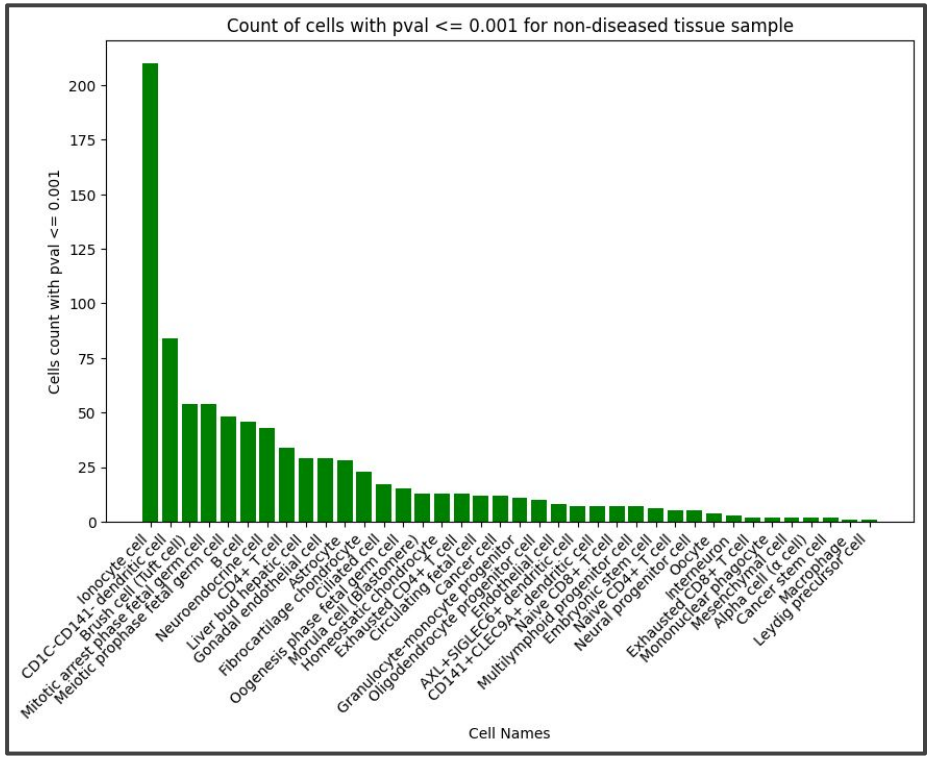
- Analysed Unipath scores for various cell types
- Discovered significantly low p-values for Ionocyte Cells in the lung Dataset
- Indicated Potential Key biological pathways and cell types

Filter | Cols: « < 1 – 50 > »

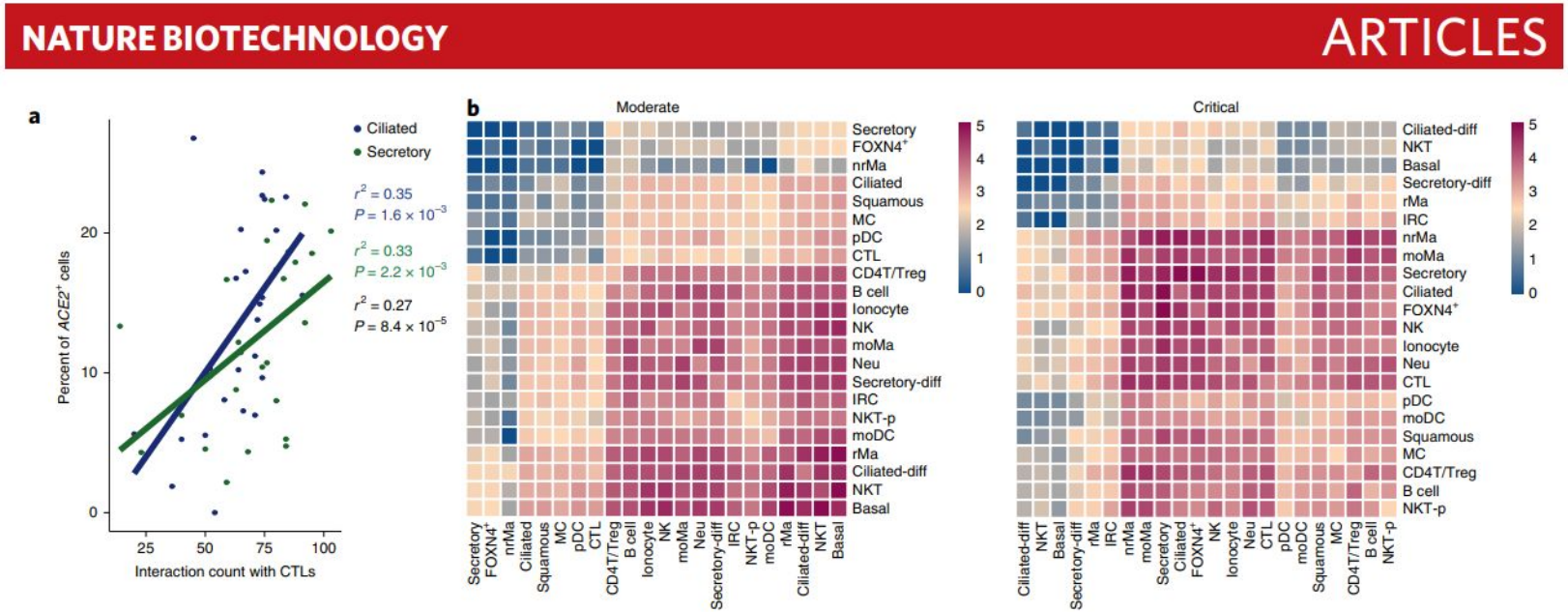
	V1	V2	V3	V4	V5	V6	V7	V8
IGKC	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
SFTPC	0.12536932	1.00000000	0.21196544	0.20704864	1.00000000	0.34068752	1.00000000	0.18998841
SFTPA2	0.36278869	1.00000000	1.00000000	0.29768174	1.00000000	0.83022922	1.00000000	0.97306320
IGLC2	1.00000000	1.00000000	0.98993564	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
SFTPA1	0.61026415	1.00000000	1.00000000	0.42360708	1.00000000	1.00000000	1.00000000	1.00000000
IGHG3	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
CEMIP	1.00000000	1.00000000	1.00000000	1.00000000	0.30228551	0.12569270	1.00000000	0.98970821
COL1A1	1.00000000	0.35631170	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
IGLC1	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
SERPINE1	1.00000000	0.26162590	0.40429166	1.00000000	0.40859603	1.00000000	0.55010212	1.00000000
HS3ST2	1.00000000	0.57265402	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
IGHM	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	0.54586589	1.00000000	1.00000000
BPIFB1	1.00000000	1.00000000	0.59649885	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
FN1	1.00000000	1.00000000	1.00000000	1.00000000	0.47338225	1.00000000	1.00000000	1.00000000

Significantly low p-values for Ionocyte Cells found in Unipath scores of the dataset





Literature backing/survey



Complex Interactions of epithelial and immune cells types in COVID-19 cells
Source: https://www.nature.com/articles/s41587-020-0602-4

Cell states associated with COVID-19 severity

Next, we examined how each cell type responds according to different peak disease severity scores. We performed pairwise differential expression (DE) tests between control WHO 0, COVID-19 WHO 1–5, and COVID-19 WHO 6–8 groups (Tables S2, S3, and S4). Among all coarse cell types, the largest transcriptional changes (measured by the number of DE genes with FDR < 0.001, and log fold change > 0.25) are observed within the epithelial compartment, including ciliated cells, developing ciliated cells, secretory cells, goblet cells, and ionocytes (Figure S4A). Among detailed cell types, we observed the largest transcriptional changes among *AZGP1*^{high} goblet cells, early-response *FOXJ1*^{high} ciliated cells, *FOXJ1*^{high} ciliated cells, *MUC5AC*^{high} goblet cells, *SERPINB11*^{high} secretory cells, early-response secretory cells, and IFN-responsive ciliated cells (Figure 3A). When we directly compared mild or moderate to severe COVID-19, we found that multiple cell types show robust transcriptional changes, most drastically among ciliated cell subtypes (IFN-responsive ciliated cells, *FOXJ1*^{high} ciliated cells, early-response *FOXJ1*^{high} ciliated cells, developing ciliated cells), ionocytes, *SERPINB11*^{high} secretory cells, early-response secretory cells, and *AZGP1*^{high} goblet cells.

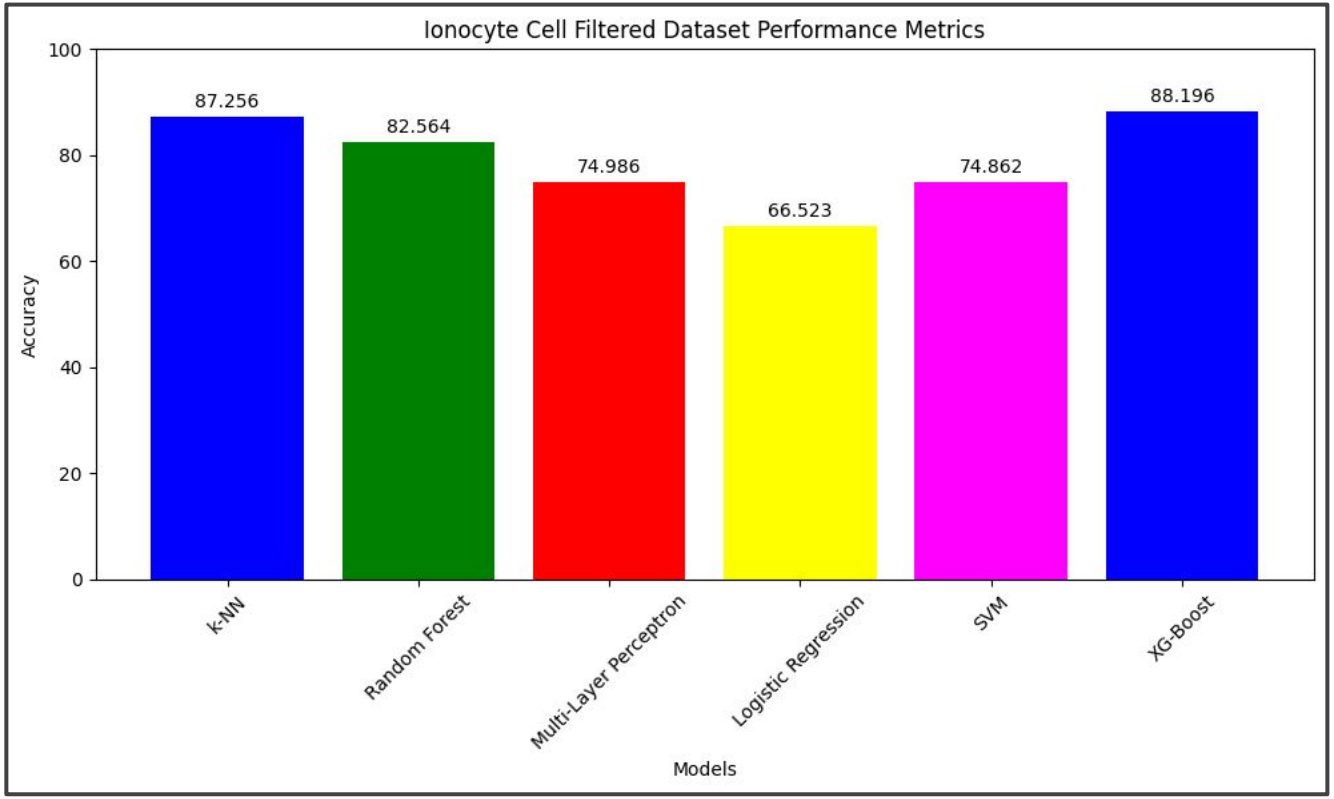
Source: Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19
Paper Link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8299217/

Next, we filtered the tissue samples of Ionocyte cells that had p-value scores of 0.001 or lower.

Ran ML Model for tissue data for only Ionocyte Cells

- Concentrated on raw data of filtered Ionocyte cells
- Trained the ML model on this refined subset of raw data
- Assessed the models specifically for Ionocyte cells
- Gained targeted insights for those

Ionocyte Cell Dataset Performance Metrics



Gathered UniPath Scores for Ionocyte Cell Tissue Samples

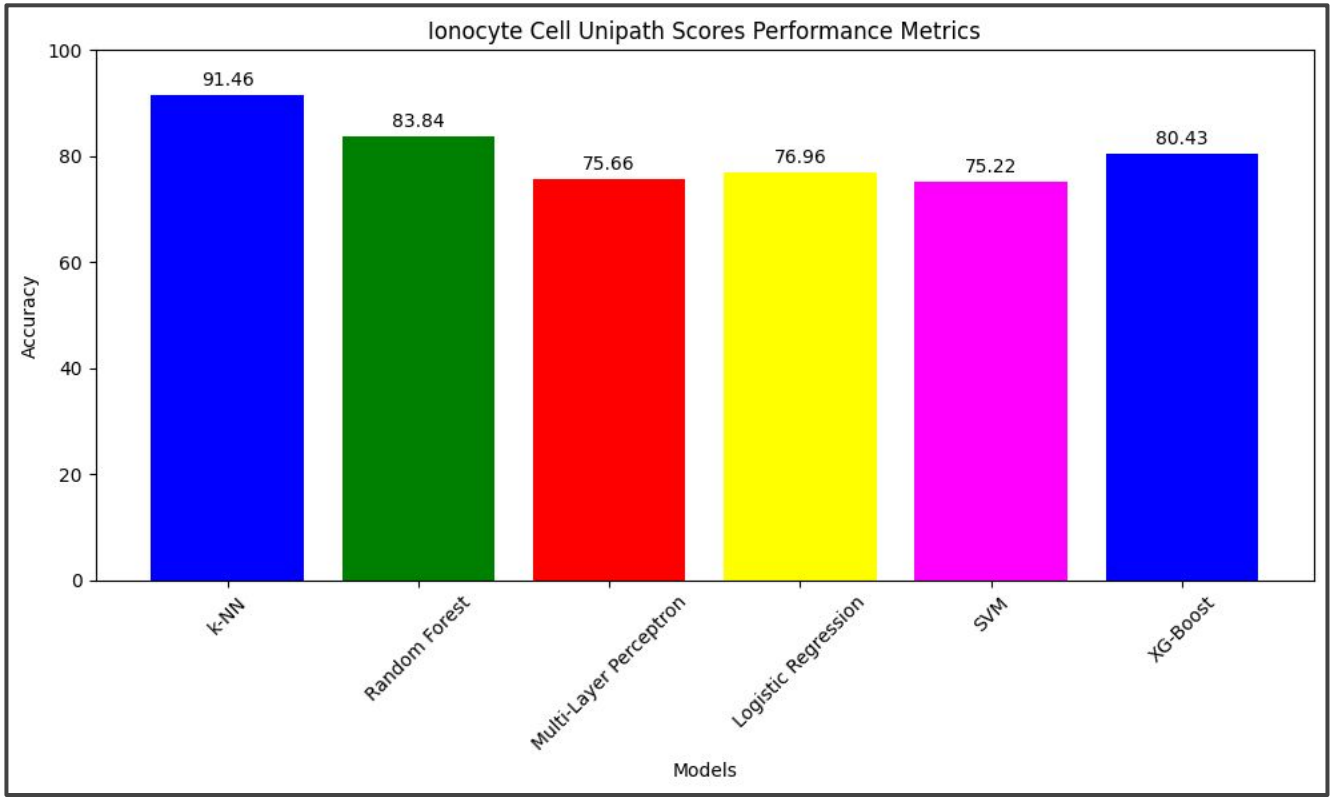
- Used Unipath to generate pathway activity scores for the filtered Ionocyte tissue samples
- Reviewed and interpreted the Pathway scores
- Got findings with biological context

scores	list [3]	List of length 3
adjpva	double [4436 x 581]	0.000 0.714 0.000 0.002 0.000 0.760 0.000 0.197 0....
adjpvaraw	double [4436 x 581]	1.000 0.286 1.000 0.998 1.000 0.240 1.000 0.803 1....
adjpvalog	double [4436 x 581]	-0.000144 1.805409 -0.000144 0.002744 -0.000144...

Final ML Model on UniPath Scores for Ionocyte Cells

- Trained the ML Model using Unipath Scores of Ionocyte tissue samples
- Analysed the Model performance and validated findings
- Highlight the potential insights and validated findings

Ionocyte Cell Unipath Scores Metrics



Results

k-NN Strong Performance: The k-NN model consistently outperforms other models across multiple datasets, achieving the highest accuracy in the Ionocyte Cell Unipath Scores with 91.46%.

Random Forest Reliability: Random Forest exhibits robust performance, especially notable in the Raw Dataset and Ionocyte Cell Filtered Dataset with accuracies close to 88.84% and 82.564%.

XGBoost High Scores: XGBoost shows strong results, particularly in the Ionocyte Cell Filtered Dataset, marking an accuracy of 88.196%.

We summarize the results on the basis that when filtered the dataset with tissue samples having p-value <= 0.001, every model performs outperforms its previous scores.

This tells that the tissue samples having Ionocyte pval <= 0.001 shows much relevancy in predicting the diseased and non-diseased cells.