Machine Learning Models for Automated Annotation of Liver Cell Clusters in Single-Cell RNA Sequencing Data

I. Abstract

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity in complex tissues. However, the annotation of cell clusters remains a significant bottleneck to the processing of scRNA-seq data, often requiring extensive manual curation and expert knowledge of cell type marker genes. We present a novel machine learning approach for automated classification of liver cells and annotation of liver cell clusters in scRNA-seq data. We developed and compared Random Forest Classifier (RFC) and Neural Network (NN) models for both human and mouse liver datasets. Our models achieved high accuracy in cell type classification, with the human models reaching 88% classification accuracy on independent datasets and mouse models achieving 98% accuracy; the models also demonstrated robust and generalizable annotation capabilities: even when annotated cell types were incorrect, they were often physically and functionally close to the real cell type.The NN models demonstrated superior performance on sparse datasets. The models, however, still do show signs of class imbalance, where cell types with large populations in the training data have a higher rate of false positives. The models were published as a Python library that works out of the box, without any fine tuning, on scRNA-seq data of liver samples from mouse and human data. These tools offer a simple, rapid, objective method for annotating liver cell populations, potentially accelerating liver research.

II. Introduction

Single-cell RNA sequencing has emerged as a powerful tool for dissecting cellular heterogeneity within tissues(Saliba et al., 2014). This technology has been particularly valuable in studying complex organs like the liver, where diverse cell types interact during various biological processes(MacParland et al., 2018). However, a critical challenge in scRNA-seq analysis is the accurate annotation of cell clusters, which is often performed manually and requires extensive domain expertise(Abdelaal et al., 2019).

Current solutions for cluster annotation have significant limitations. Manual annotation is time-consuming, subjective, and requires in-depth knowledge of tissue and cell-type specific marker genes(Abdelaal et al., 2019). Existing computational tools often require fine-tuning on

pre-annotated reference datasets or lists of cell-type specific marker genes, which can be scarce or hard to find(Abdelaal et al., 2019). Furthermore, fine tuning is often a computationally and time expensive process.

To address these challenges, we sought to develop machine learning models specifically trained on liver scRNA-seq data for automated cell classification and cluster annotation. Our approach aimed to create models that could work "out of the box" for liver samples, eliminating the need for extensive fine-tuning or pre-existing expert-annotated data. By training and fine tuning our models on tissue-specific data, we hypothesized that our models would offer more convenient, accurate and consistent annotations compared to existing general-purpose tools.

The models serve 2 purposes: annotation of clusters and classification of cells. Here, classification is used to refer to the labeling of individual cells with their predicted cell type. This can occur before clustering. These classifications can then be used to group the cells into clusters, and offer a way to do so that aligns directly with the mapping onto the model training datasets. Annotation, however, is used to refer to a more common use case. It refers to the labeling of clusters with the inferred cell type that the cells within that cluster correspond to.

In this study, we present the development and evaluation of Random Forest Classifier and Neural Network models for both human and mouse liver cell annotation. We demonstrate the high accuracy of these models on independent datasets and compare their performance to existing annotation tools. Our work provides a valuable resource for the liver research community and establishes a framework for developing similar tissue-specific annotation tools for other organs.

The models were released as part of a Python library for classification of cells and annotation of clusters. The library was designed to work seamlessly with AnnData and Scanpy, which are used extensively in the analysis of scRNA-seq data in Python(Virshup et al., 2023). Furthermore, it contributes to an efficient development pipeline, as it does not require fine tuning or re-training, as other models do(Abdelaal et al., 2019).

III. Methods

A. Data Collection and Preprocessing

For the human liver models, we utilized a healthy liver atlas with around 8,000 cells as our primary training data(MacParland et al., 2018). The data included pre-clustered and pre-annotated cell populations. For benchmarking the model, we used an independent hepatic carcinoma dataset containing about 7,000 cells, as well as a dataset of integrated scRNA-seq data from 4 healthy human livers(Andrews et al., 2022; Massalha et al., 2020).

The mouse model was trained on a healthy subset of an scRNA-seq dataset of healthy and MASH mice livers, which included approximately 13,000 healthy, pre-clustered and pre-annotated cells(Wang et al., 2023). For benchmarking the model, we used a healthy subset of an atlas of healthy and injured mice livers(Carlessi, 2023).

B. Model Development

We implemented the Random Forest Classifier using scikit-learn in Python(*Scikit-Learn: Machine Learning in Python — Scikit-Learn 1.5.1 Documentation*, n.d.). Initial models were trained with 100 trees and default hyperparameters. These models displayed suboptimal performance, and suffered from class imbalance (Fig 1, initial classification reports of models). We then optimized the hyperparameters using RandomizedSearchCV, with the search space including parameters such as max_depth, max_features, min_samples_leaf, min_samples_split, and n_estimators.

Neural Network models were implemented using PyTorch, with the architecture and training procedures optimized for each dataset(*PyTorch*, n.d.).

Further specifics regarding the model development pipeline can be found in the source code on GitHub (https://github.com/mThkTrn/liver_annotation_workbooks)

C. Classification and Annotation Functionalities

The models were initially trained as classifiers, as there were an abundance of annotated cell types to train them on, while there were not many clusters to train said models on. Furthermore, cells proved to be clearly defined data points, while gauging the expression profile of a whole cluster would have been a more subjective task, as there are varying methods and algorithms to do so.

Still, the intent of the models was to also serve as annotators. Thus, algorithms were created to infer the cell type identity of a cluster. One algorithm takes the mode cell type identity of the cells inside a cluster, and annotates the cluster with that cell type. An alternative technique sums the probabilities that the cells inside a cluster are of any given cell type, and annotates the cluster with the cell type that has the highest sum of probabilities. Both methods are implemented in the Python library.

D. Model Evaluation Metrics

Models were evaluated on test data from the same datasets the models were trained on using accuracy scores, precision, recall, and F1-scores. We also generated confusion matrices to assess performance across different cell types.

E. Benchmarking Against Other Tools and Datasets

We compared our models' performance against existing tools, such as scmap and SingleR(Kiselev et al., 2018; *SingleR*, n.d.). We also evaluated the models on independent datasets to assess generalizability.

When computing accuracy of classification on another model, all cell type annotations, both the manual annotations of the original dataset and the computer generated classifications, were mapped to broad liver cell types (Hepatocytes, Cholangiocytes, Hepatic Stellate Cells, etc.). If more specific cell types had designated cluster(s) in both the machine classifications and manual annotations, mappings were generated to those cell types, as well. Cells belonging to a cluster that had no corresponding cluster in the machine annotations were ignored. Thus, the accuracy scores largely only reflect the accurate classification of cell types, and not sub-cell type populations represented by separate clusters.

Machine annotations of clusters were manually compared and inspected for accuracy.

IV. Results

A. Human Liver Model

The optimized human RFC model achieved an accuracy of 88.83% on an independent hepatic carcinoma dataset. The model demonstrated robustness in annotating both known and novel cell types. The human NN model slightly outperformed the RFC, achieving an accuracy of 90.85% on the same benchmarking dataset. Both models showed similar annotation patterns, with the NN exhibiting marginally better performance in classifying rare cell populations. The confusion matrices for the human NN and RFC models can be found in **Figure 2.**

When tested on a dataset of healthy human livers, which had limited shared genes with the training data, the RFC model's performance dropped significantly (23.10% accuracy), while the NN model maintained higher accuracy (82.78%). This stark difference in performance highlights

the NN model's superior ability to handle sparse datasets. Specifics regarding performance and annotation can be seen in **Table 1.**

Table 1. Post-training classification and annotation performance of human NN and RFC models across training and test datasets.

	Neural Network	Random Forest Classifier
Accuracy (test data)	96%	94%
Accuracy (cancer dataset)	88%	88%
Annotations (cancer dataset)	B cells → Plasma Cells CAFs → Hepatic Stellate Cells Carcinoma 1 → Cholangiocytes Carcinoma 2 → Cholangiocytes Hepatocytes → Hepatocyte 4 Kupffer cells → Non inflammatory Macrophage LSEC → Central venous LSECs LVEC → Periportal LSECs LVECt → Periportal LSECs Pericytes → Hepatic Stellate Cells Proliferation → Inflammatory Macrophage SAMs → Inflammatory Macrophage Stellate cells → Hepatic Stellate Cells T cells → Alpha beta T Cells TM1 → Inflammatory Macrophage cDC1 → Inflammatory Macrophage cDC2 → Inflammatory Macrophage vSMC → Hepatic Stellate Cells	B cells → Plasma Cells CAFs → Hepatic Stellate Cells Carcinoma 1 → Cholangiocytes Carcinoma 2 → Cholangiocytes Hepatocytes → Hepatocyte 4 Kupffer cells → Non inflammatory Macrophage LSEC → Central venous LSECs LVEC → Periportal LSECs LVECt → Periportal LSECs Pericytes → Hepatic Stellate Cells Proliferation → Inflammatory Macrophage SAMs → Inflammatory Macrophage Stellate cells → Hepatic Stellate Cells T cells → Alpha beta T Cells TM1 → Inflammatory Macrophage cDC1 → Inflammatory Macrophage cDC2 → Inflammatory Macrophage vSMC → Hepatic Stellate Cells
Accuracy (healthy dataset)	82%	23%
Annotations (healthy dataset)	Bcells --> Alpha beta T Cells Cholangiocyte --> Cholangiocytes Erythroid --> Hepatocyte 6 Hepatocyte --> Hepatocyte 4 LSECs --> Central venous LSECs Macrophage --> Inflammatory Macrophage NKTcell --> Alpha beta T Cells Stellate --> Hepatic Stellate Cells	B Cells → Plasma Cells Cholangiocyte → Cholangiocytes Erythroid → Plasma Cells Hepatocyte → Cholangiocytes LSECs → Central venous LSECs Macrophage → Inflammatory Macrophage NKTcell → Plasma Cells Stellate → Hepatic Stellate Cells

B. Mouse Liver Model

The mouse RFC model demonstrated exceptional performance, achieving 98.32% accuracy on an independent healthy liver dataset. The mouse NN model performed similarly well, with 98.13% accuracy on the same benchmarking dataset. Both mouse models showed nearly identical performance, with very high accuracy in both annotation and classification tasks. The confusion matrices for the mouse NN and RFC models can be found in **Figure 3.**

Both models maintained their high performance when tested on the independent healthy liver dataset, demonstrating robust generalization. Specifics regarding performance and annotation can be seen in **Table 2.**

Table 2. Post-training classification and annotation performance of mouse NN and RFC models across training and test datasets.

	Neural Network	RFC
Accuracy (training dataset)	72%	71%
Accuracy (test dataset)	98%	98%
Annotations (test dataset)	B Cells → B Cell BECs → Cholangiocytes Endo → Endo1 Hep → Hep1 Mesenchymal → HSC1 Meso → Mesothelial Myeloid → Macrophages T_NK Cells → T Cell pDCs → Unknown Immune	B Cells → B Cell BECs → Cholangiocytes Endo → Endo1 Hep → Hep1 Mesenchymal → HSC1 Meso → Mesothelial Myeloid → Macrophages T_NK Cells → T Cell pDCs → Unknown Immune

C. Comparison of models with other libraries

Our human NN model outperformed existing tools like scmap, which achieved only 44.50% accuracy when used on the same benchmarking and reference datasets (Andrews et al., 2022; MacParland et al., 2018). Furthermore, it was more efficient than scmap. SingleR could not perform the fine tuning in any comparable timeframe, so its comparison with our models was unfeasible.

Table 3. Comparison of performance of human NN model against other automated annotation libraries.

Library	Accuracy (%)	Time (s)
liver_annotation		
scMap	44.5	48
OnClass		
scBalance		
scNym		

D. Comparison of Human and Mouse Model Performance

Although the human models performed better than the mouse models when the models were evaluated on test data from the same dataset as their training data, on datasets that were unrelated to the original dataset, the mouse model outperformed the human models.

V. Discussion

Our machine learning models for liver cell annotation demonstrated high accuracy and robustness across different datasets, showcasing a high potential for generalisability. The strong performance of both RFC and NN architectures suggests that the models have successfully captured the underlying patterns of gene expression that define liver cell types. Furthermore, they maintained relatively accurate predictions across

The superior performance of the mouse models when generalized to datasets unrelated to the training data was expected, as the human test datasets were quite different from the training dataset: one represented a tumorous liver, and the other had very few shared genes with the training dataset(Andrews et al., 2022; MacParland et al., 2018; Massalha et al., 2020). Furthermore, discrepancies may be due to the higher genetic and environmental homogeneity in mouse samples compared to human samples, as the samples for most mouse datasets come from mouse models conducted in a controlled, laboratory environment. This may have led to more consistent performance on data from different experiments.

A notable finding was the NN model's superior performance on sparse datasets, particularly evident in the human liver benchmarking. When tested on a dataset with limited shared genes,

the NN model maintained high accuracy (82.78%) while the RFC model's performance plummeted (23.10%). This suggests that the NN architecture is more adept at extracting meaningful patterns from limited data. This capability could be particularly valuable when working with datasets that represent unhealthy livers, as well as datasets with few shared genes with the training dataset.

One limitation of our models is the issue of class imbalance. As evidenced by the confusion matrices, larger classes tend to have a higher rate of false positives, while smaller classes suffer from a higher rate of false negatives. For example, in both human and mouse models, the Hepatocyte_1/Hep 1 cluster, which has the most cells, also has a very low precision, as seen in **Fig 1** and **Fig 2**. This imbalance reflects the natural distribution of cell types in the liver but can lead to biased predictions. This problem persists even after fine tuning the RFC models to counteract class imbalance, and even as RFC models are known to be resistant to class imbalance.

Even when annotations were wrong, they were often close to the real function of the cell type. This can be seen in **Figure 2** and **Figure 3**, where the bulk of the hepatocytes' false positives are with other hepatocytes, or in **Table 1**, where the human NN model annotated a mesenchymal cluster in the cancer dataset to be a Hepatic Stellate Cell.

The improved performance of our models with comparison is likely due to the fine training of the model architecture involved in our model development pipeline. The improved efficiency is likely due to the dataset fine tuning process required by other libraries, which require providing a reference dataset.

VI. Conclusion

We have developed high-performing machine learning models for automated annotation of liver cell clusters in scRNA-seq data. These models demonstrate superior accuracy compared to existing tools and offer a valuable resource for the liver research community. By accelerating the annotation process and providing consistent results, our tools have the potential to significantly advance our understanding of liver biology and disease.

However, this study has several limitations that should be addressed in future research. The class imbalance issue, while partially mitigated through hyperparameter optimization, remains a challenge. Future work could explore more sophisticated techniques to balance class representations without losing biological relevance.

This research contributes to the field by providing a robust, tissue-specific solution for cell type annotation in liver scRNA-seq data. It demonstrates the value of tailoring machine learning models to specific biological contexts and highlights the potential advantages of neural network architectures in handling sparse biological datasets.

Future work could explore the transferability of these models across species, the integration of spatial information to improve annotation accuracy, and the development of interpretable models that can provide insights into the key genes driving cell type classifications. Building upon this work, similar tissue-specific models for other organs could be developed, creating a comprehensive toolkit for scRNA-seq analysis across multiple tissues.

VII. Data and Code Accession.

The models are accessible as a Python library, the code for which can be found at https://github.com/mThkTrn/liver_annotation. The scripts used to preprocess the data and train the models can be found at https://github.com/mThkTrn/liver_annotation_workbooks/.

The datasets used to train and evaluate the models can be found in NCBI GEO. The human training dataset, “Dissecting the human liver cellular landscape by single cell RNA-seq reveals novel intrahepatic monocyte/ macrophage populations”, can be found under GSE115469. The healthy human test dataset, “Single-Cell, Single-Nucleus and Spatial RNA Sequencing of the Human Liver Identifies Hepatic Stellate Cell and Cholangiocyte Heterogeneity”, can be found under GSE185477. The human cancer test dataset, “A single cell atlas of the human liver tumor microenvironment”, can be found under GSE146409. The mouse training data, “ An autocrine signaling circuit in hepatic stellate cells underlies advanced fibrosis in non-alcoholic steatohepatitis”, can be found under GSE212837. The mouse test data, “Single Nucleus RNA Sequencing of Pre-Malignant Liver Reveals Disease-Associated Hepatocyte State with HCC Prognostic Potential”, can be found under GSE200366.

VIII. References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J. T., & Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, *20*(1), 194. https://doi.org/10.1186/s13059-019-1795-z

Andrews, T. S., Atif, J., Liu, J. C., Perciani, C. T., Ma, X.-Z., Thoeni, C., Slyper, M., Eraslan, G., Segerstolpe, A., Manuel, J., Chung, S., Winter, E., Cirlan, I., Khuu, N., Fischer, S., Rozenblatt-Rosen, O., Regev, A., McGilvray, I. D., Bader, G. D., & MacParland, S. A. (2022). Single-Cell, Single-Nucleus, and Spatial RNA Sequencing of the Human Liver Identifies Cholangiocyte and Mesenchymal Heterogeneity. *Hepatology Communications*, *6*(4), 821–840. https://doi.org/10.1002/hep4.1854

Carlessi, R. (2023). *Single Nucleus RNA Sequencing of Pre-Malignant Liver Reveals Disease-Associated Hepatocyte State with HCC Prognostic Potential*. *2*. https://doi.org/10.17632/w7yh4yjvbw.2

Kiselev, V. Y., Yiu, A., & Hemberg, M. (2018). scmap: Projection of single-cell RNA-seq data across data sets. *Nature Methods*, *15*(5), 359–362. https://doi.org/10.1038/nmeth.4644

MacParland, S. A., Liu, J. C., Ma, X.-Z., Innes, B. T., Bartczak, A. M., Gage, B. K., Manuel, J., Khuu, N., Echeverri, J., Linares, I., Gupta, R., Cheng, M. L., Liu, L. Y., Camat, D., Chung, S. W., Seliga, R. K., Shao, Z., Lee, E., Ogawa, S., … McGilvray, I. D. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature Communications*, *9*(1), 4383. https://doi.org/10.1038/s41467-018-06318-7

Massalha, H., Bahar Halpern, K., Abu-Gazala, S., Jana, T., Massasa, E. E., Moor, A. E., Buchauer, L., Rozenberg, M., Pikarsky, E., Amit, I., Zamir, G., & Itzkovitz, S. (2020). A single cell atlas of the human liver tumor microenvironment. *Molecular Systems Biology*, *16*(12), e9682. https://doi.org/10.15252/msb.20209682

PyTorch. (n.d.). PyTorch. Retrieved August 26, 2024, from https://pytorch.org/

Saliba, A.-E., Westermann, A. J., Gorski, S. A., & Vogel, J. (2014). Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Research*, *42*(14), 8845–8860. https://doi.org/10.1093/nar/gku555

Scikit-learn: Machine learning in Python—Scikit-learn 1.5.1 documentation. (n.d.). Retrieved August 26, 2024, from https://scikit-learn.org/stable/

SingleR. (n.d.). Bioconductor. Retrieved August 26, 2024, from http://bioconductor.org/packages/SingleR/

Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Berger, B., Pe'er, D., Regev, A., Teichmann, S. A., Finotello, F., Wolf, F. A., Yosef, N., Stegle, O., & Theis, F. J. (2023). The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology*, *41*(5), 604–606. https://doi.org/10.1038/s41587-023-01733-8

Wang, S., Li, K., Pickholz, E., Dobie, R., Matchett, K. P., Henderson, N. C., Carrico, C., Driver, I., Borch Jensen, M., Chen, L., Petitjean, M., Bhattacharya, D., Fiel, M. I., Liu, X., Kisseleva, T., Alon, U., Adler, M., Medzhitov, R., & Friedman, S. L. (2023). An autocrine signaling circuit in hepatic stellate cells underlies advanced fibrosis in nonalcoholic steatohepatitis. *Science Translational Medicine*, *15*(677), eadd3949. https://doi.org/10.1126/scitranslmed.add3949

IX. Figures

Figure 1. Pre hyperparameter adjustment performance of human and mouse RFC models.

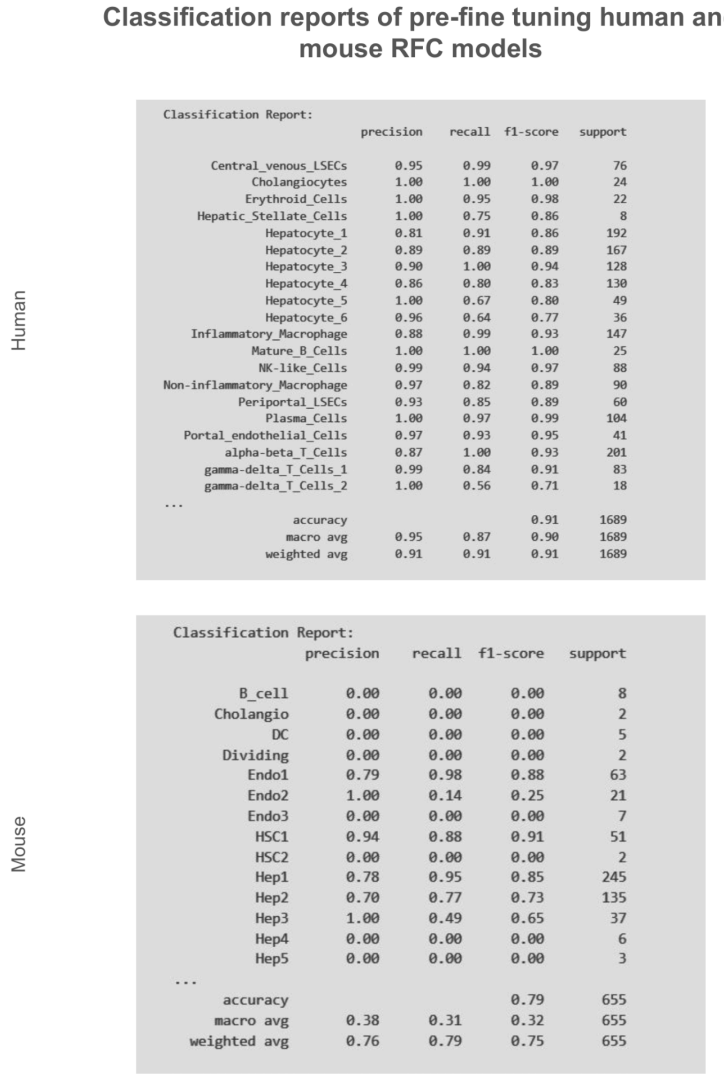


Figure 2. Confusion matrices of human NN and RFC models on test data reserved from the human training dataset. The numbers in the cells represent how many predictions of a given cell type the model made for cells of a given cell type.

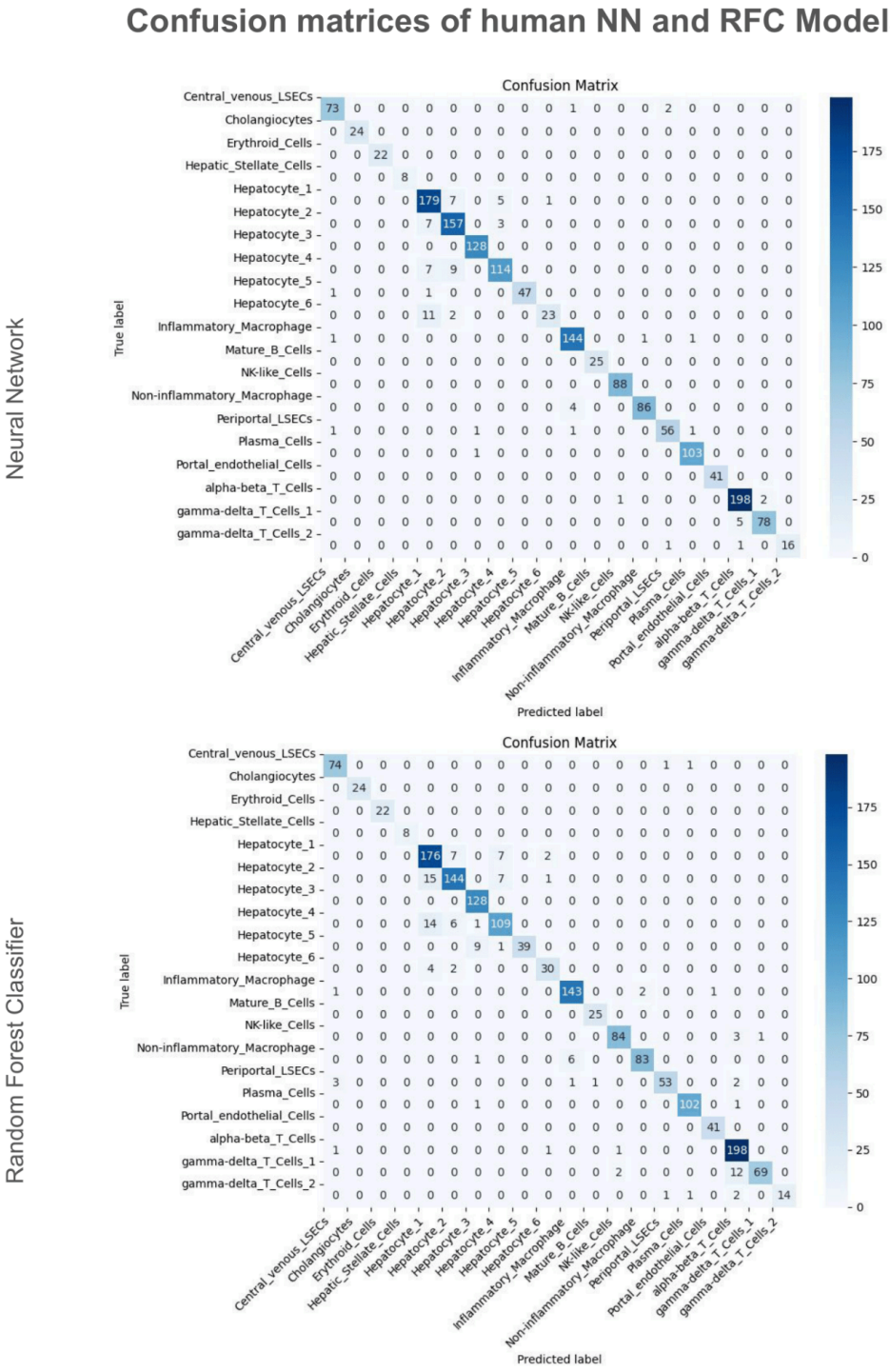
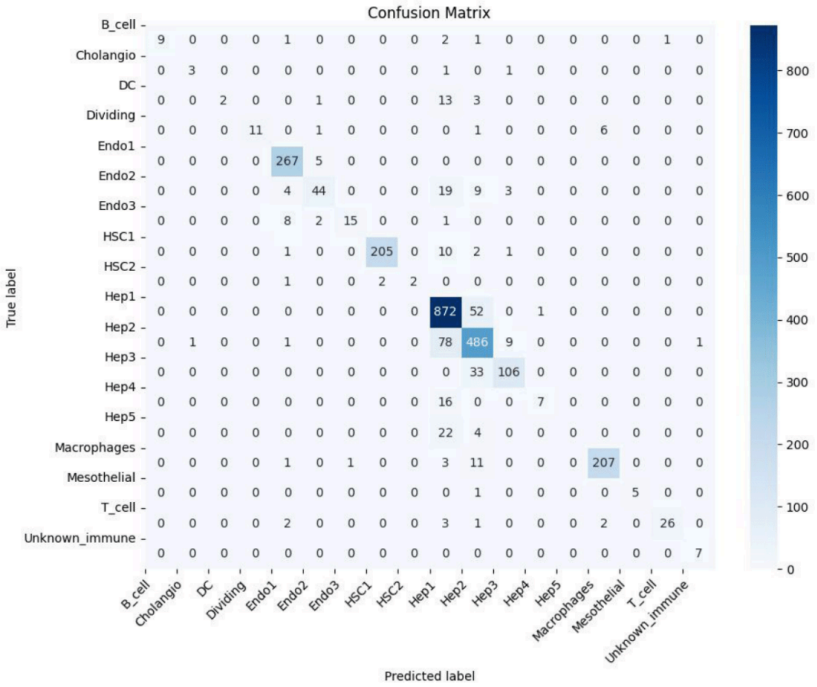


Figure 3. Confusion matrices of mouse RFC and NN models on test data reserved from the mouse training dataset. The numbers in the cells represent how many predictions of a given cell type the model made for cells of a given cell type.

Confusion matrices of mouse NN and RFC Models

Neural Network



Random Forest Classifier