Performance of Supervised Classifiers on Mouse Cell Atlas single-cell RNA Sequencing Data for Cell Annotation

Aarathi Raghuraman[1], **Malabika Sen**[1], **Nure Tasnina**[1]
[1] Virginia Tech Computer Science Department
raarathi@vt.edu, malabika@vt.edu, tasnina@vt.edu

1. Introduction

Cell type identification is a key component of any downstream analysis of sc-RNA seq data (Luecken and Theis, 2019). Identifying cell types can provide valuable information on emergence of new cell types and lineage tracing for known cell types (Svensson, Vento-Tormo and Teichmann, 2018). Commonly, cell type identification is achieved either through manual annotation, automated unsupervised clustering or supervised classification. There has been a significant amount of algorithms developed for automatic cell type identification given the cumbersome nature of manual annotation. Based on a survey of 22 such classification algorithms (Abdelaal et al., 2019), and brief research on automated clustering methods we believe there is room for improvement. Despite the vast number of datasets (27) used to measure annotation accuracy, Abdelaal et. al.'s analysis lacks testing on a truly diverse dataset like that of the Mouse Cell Atlas (Han et al., 2018). In light of these shortcomings with the individual cell type identification methods and validation datasets, we will study the effects of the top performing supervised cell type annotation methods on the Mouse Cell Atlas Dataset by Han et. al.

1.1. Benchmarking automatic Cell Identification Methods

Abdelaal et. al. performs a series of experiments using supervised classifiers and prior-knowledge classifiers to identify cell populations in a wide range of datasets that differ in number of cells, features, cell annotation levels and protocols (Abdelaal et al., 2019). The authors start by dividing their experiments based on Intra-dataset setups, train and test lie within the same dataset, and Inter-dataset setups, train and test lie in different datasets. The authors evaluate the performance of each classifier based on accuracy, F1 score and scalability, but the deepest dataset only contains 92 cell populations (AMB), the largest dataset only contains 65,943 cells and the largest number of features only include 21,952 features (Table 1). To truly test a classifier on accuracy, F1 score and scalability, we need to use a larger and deeply annotated dataset such as the Han et. al. Mouse Cell Atlas which has 233,994 cells (reduced version, Methods 2.1 & 2.2) across 39 different tissue types, 760 cell-types and 39,855 features (Han et al., 2018; Butler, 2018).

Dataset	Description	No. of cells	No. of genes	No. of cell populations
Tabula Muris (TM)	Whole Mus musculus	54,865	19,791	55
Allen Mouse Brain (AMB)	Primary mouse visual cortex	570	12,627	3/16/92
Zheng 68K	PBMC	65,943	20,387	11
Mouse Cell Atlas (MCA) (Butler, 2018)	Across 42 Mouse tissue types	233,994	39,855	760

Table 1: Datasets of Interest

While our objective is to train and test within the MCA dataset, the diversity of the tissue types make it important to not only consider high performing algorithms from the Intra-datset experiments, but also the Inter-dataset experiments. The Inter-dataset experiment of interest is the one where the authors trained on a pairwise combination of three, 34 cell type deep, brain datasets and tested on the one left-out 34 cell type brain dataset.

From Figure 1, we observe that apart from the Inter-dataset experiment on Deep Annotation Levels, $\text{SVM}_{rejection}$ and SVM perform relatively well on all relevant experiments, making them the first choice for evaluation on the MCA dataset. singleCellNet, scmapcell, scPred and ACTINN are some of the other well performing algorithms. The biggest downfall of singleCellNet for our application is the high compute time; scPred fails to produce a result for Complex datasets experiment; this leaves us with scmpacell and ACTINN. Given the recent rise in various cell atlas development initiatives leading to availability of large datasets and advances in deep learning based methods, we decided to explore the performance of ACTINN on the MCA dataset instead of scmapcell (Henry et al., 2014; Zappia, Phipson and Oshlack, 2018; Zheng and Wang, 2019).

Figure 1: Summary of Results of experiments run by Abdelaal et. al.: Rows represent the classification algorithms and columns represent the experiment conducted. The different levels of green represent Good, Intermediate or Poor Performance as described in the legend. Grey boxes indicate that an algorithm could not be run for the given experiment. The experiments of utmost significance to our objective are marked by orange dots

2. Materials and Methods

2.1. Dataset

We used the Mouse Cell Atlas (MCA) dataset developed by Han et. al (Han et al., 2018) to study the performance of supervised classifiers.

The MCA dataset was developed owing to the establishment of a low-cost and robust single-cell technology called Microwell-Seq. After collecting samples from a variety of tissues-origins like uterus, bladder, ovary, etc. the data was processed through Microwell-seq to create an atlas of all major mouse cell types. The whole dataset consisted of more than 400,000 single cells originating from more than 50 mouse tissues and cultures. Using literature study, they annotated the cells based on specific markers present in them. They were then able to identify more than 800 cell types grouped into 98 major clusters using 60,000 cells sam-

pled from the complete dataset. We use this annotated dataset to study the performance of the 2 classifiers: SVM & ACTINN. The MCA consists of rich information and all of the data has been generated using a single platform. The consistency in technology makes it a useful asset and this is one of the main reasons why we chose the MCA for our evaluation.

2.2. Data Pre-processing

As a part of this experiment to evaluate the performance of supervised classifiers, we will consider only the annotated part of the dataset which makes up to $\sim 234,000$ cells. This annotated dataset were also assigned a cluster ID in the original study, during which traditional quality control and filtration were already performed for sanitizing the data. As a result, we need not repeat those steps. Next, we use the Seurat package (Butler, 2018) to perform the pre-processing steps. We perform the standard log-normalization fitted to a scale of 10,000. We also select the top 1000 highly variable genes (HVG). For large datasets, selecting HVG based on variance mean ratio (VMR) has been found to be effective for downstream analyses. Further, we also remove the mitochondrial expression per cell so that they do not skew the cell based gene expression.

Finally, we reduce the filtered dataset to its top 75 principal components. As seen in Fig.2, the standard deviation is almost constant after the 75th principal component. Thus, for our experiment we perform analysis on the $\sim 234,000$ cells with a reduced feature space of the first 75 principal components.

To run one of the supervised classifiers we needed smaller number of cells. Hence, we perform a stratified sampling where we sub-sampled 10% of the cells from the above mentioned preprocessed dataset. We got on average 400 cell-types in this sub-sampled dataset. The stratified sampling method allows us to preserve the class frequencies in the smaller dataset.

The Han dataset has batch-specific datasets each of which contains cells from a single tissue. After keeping only the batches for which we have cell-type annotations there remained 65 batches.

2.3. Supervised Classifiers

2.3.1. Support Vector Machine (SVM)

We consider SVM with Linear and Gaussian Kernel for this study. For each of the following experiments we did 5-fold cross validation to validate the performance of the model.

- **Batch-Specific Tissue:** We ran SVM with Linear Kernel on each batch of cells. Here, the cells are the samples and the raw gene expression values are the features. The number of genes were around two to fifteen times the number of cells for the batch tissues which led us to the conclusion that SVM with Gaussian Kernel will not give a good result, rather it might give us a complex overfitted model. That is the reasoning behind choosing Linear Kernel for this dataset.
  We used scikit-learn's (Pedregosa et al., 2011) LinearSVC() method with $C = 1$ for this purpose.

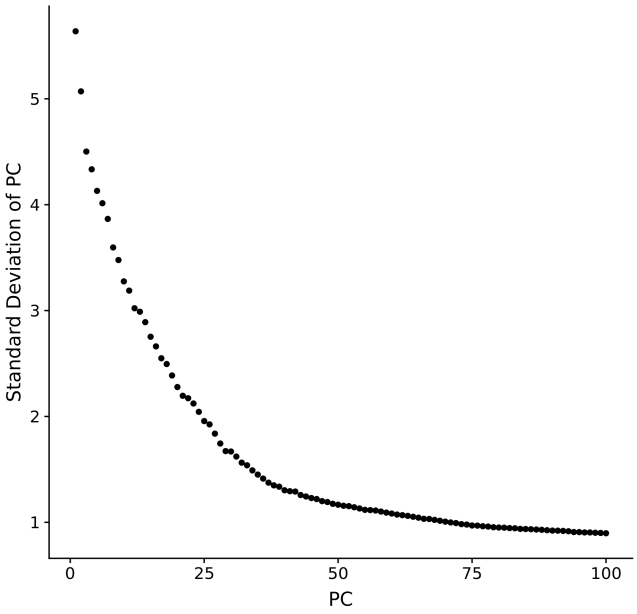

Figure 2: ElbowPlot of the standard deviation measured against the Principal Components

- 10% **Stratified Sub-sample:** We ran SVM with Gaussian Kernel on 10% cells (i.e. 24,000 cells) of the whole preprocessed dataset. In this experiment, cells are the samples and 75 principle components derived from PCA of gene expression values of 1000 highly variable genes are the features. As the number of samples is way higher than the number of features we implemented SVM with Gaussian Kernel as in such cases it is proven to do better than or same as SVM with Linear Kernel.
  We used scikit-learn's SVC() method with $C = 1$ and $Kernel = RBF$ for this experiment.

- **Whole Preprocessed Dataset:** We tried running SVM with Linear Kernel on the whole preprocessed dataset with $234,000$ cells and 75 principle components. We used scikit-learn's Lin-

earSVC() method with $C = 1$ and $C = 10$ for this experiment. However, the model did not converge for this dataset with provided value for the hyper-parameter C after running for 8 hours. Given the time constraint, we could not tune the hyper-parameter to see if it converges for any other values.

2.3.2. Automated Cell Type Identification using Neural Networks (ACTINN)

ACTINN is a 4 layer neural network built using tensorflow with 3 linear hidden layers with 100, 50 and 25 nodes in order. The number of nodes of the last layer corresponds to the number of cell types. Each hidden layer uses a Recti-Linear Unit activation function (ReLU) with the last layer using a softmax function. The model uses Adam optimizer with cross-entropy as the loss function (Ma and Pellegrini, 2020). Using ACTINN we perform 3 different experiments:

- 10% **Stratified Sub-sample:** We ran ACTINN with a learning rate of 0.01 and batch size of 128 for 50 epochs on 10% sub-sample of the whole preprocessed dataset with $234,000$ cells and 75 PCA components (Methods 2.2). Evaluations were performed on 10-fold cross-validation & training on 90% of the dataset, reported as validation accuracy, and the held-out 10% of the dataset was used to find testing accuracy; $21,060$ cells were used for cross-validation & training and $2,340$ cells were used for testing.

- **Whole Preprocessed Dataset:** One of the areas where ACTINN shined was its ability to scale well, which was true when compared to SVM on the MCA dataset as well. We were able to run the model on the whole dataset (234000 cells with 75 PCA components) with a learning rate of 0.0001 & 0.01 and batch size of 128 for 50 epochs. We split the dataset similar to the 10% sub-sample dataset with 90% of the dataset for cross-validation & training and 10% of the dataset for testing; $210,594$ cells were used for cross-validation & training and $23,340$ cells were used for testing.

- **Tissue Type Annotation:** For the last experiment, we ran ACTINN on the Whole Preprocessed Dataset with Tissue Type as the desired output labels. In this run, we used a learning rate of 0.01 with batch size of 128 for 50 epochs. Given that there are only 39 tissue types with every tissue type apart from Fetal_Kidney (only 11 count) having atleast 500 cells to represent, we expect a much higher performance. Evaluations were performed on 10-fold cross-validation & training on 90% of the dataset, reported as validation accuracy, and the held-out 10% of the dataset was used to find testing accuracy; $210,594$ cells were used for cross-validation & training and $23,340$ cells were used for testing.

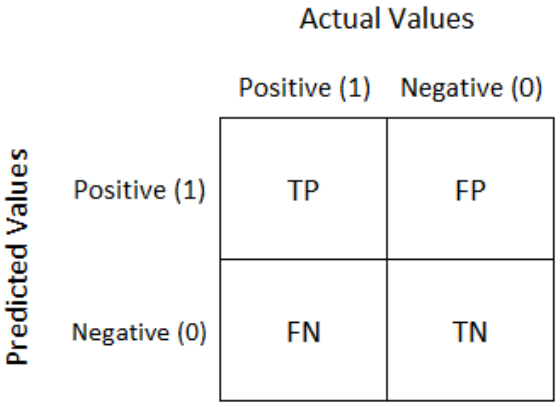

Figure 3: Confusion Matrix

2.4. Evaluation Metrics

2.4.1. Confusion Matrix

A confusion matrix is a concise and easy to interpret visualization for any classification model's predictions. The rows in the confusion matrix represent the predicted values and the columns represent the actual values for each of the different labels, in our scenario this will be cell types. The confusion matrix for a binary classification problem is given by Figure 3, where True Positive (TP) represents all positive predictions that were correctly predicted, True Negative (TN) represents all negative predictions that were correctly predicted, False Positive (FP) represent all positive predictions that were incorrectly predicted and False Negative (FN) represent all negative predictions that were incorrectly predicted. Based on the confusion matrix, you can find accuracy, precision, recall & F1 score as described in the next subsection. To adapt the binary confusion matrix for multi-class predictions, we simply increase the number of rows and columns to represent the multiple cell types. When visualizing the confusion matrix now, each row explains the distribution of the predictions for the true cell type defined by the row. Note: In our figures we normalized the counts with respect to each true label.

2.4.2. Accuracy, Precision, Recall & F1 Score

The Accuracy of the model is a common way to interpret its success, but often does not explain the com-

plete distribution of the predictions for each cell type.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

In order to better capture the FP & FN, we use the F1 score which is a harmonic mean of precision and recall as defined below.

$$\text{Precision} = \frac{TP}{TP + FP}$$
$$\text{Recall} = \frac{TP}{TP + FN}$$
$$\text{F1 score} = \frac{2 * Recall * Precision}{Recall + Precision}$$

The F1 Score represents both precision and recall in one value. In order to account for our multi-class scenario the F1 score is calculated as one vs. all, that is, for each cell type the true positives are represented by the correctly identified cell types and the false positives are a sum of all the wrongly identified cell types. We can do the same for negative values to compute the mean or median F1 Score. Note: All metrics are either represented as a fraction between 0 and 1 or a percentage.

3. Results

3.1. Preprocessed Dataset

Fig. 4 gives a plot of the ~ 800 cell-types distributed across the $\sim 234,000$ cells.
The 2 cell-types TS_Mrpl12_high(Trophoblast-Stem-Cell) and ES_Actb_high(Embryonic-Stem-Cell) make up close to 10% of the dataset. As seen in Fig. 4, the distribution count of the data is unbalanced across the cell-types. We will see how this may play a role in our chosen models' prediction ability in the following sections.

3.2. SVM

3.2.1. Classification per batch

We have 65 batches each containing cells from a single tissue. For SVM with Linear Kernel, we have performed classification on each of the 65 batch specific tissues seperately. Hence, we have 65 trained SVM model at the end each having the power of determining cell types of cells coming from a particular type of tissue. Fig. 5 shows the accuracy achieved for the different batch tissues.
Liver_2 has the highest accuracy, close to $\sim 99\%$. with only 3 cell-types and 229 cells with 16706 genes in it.

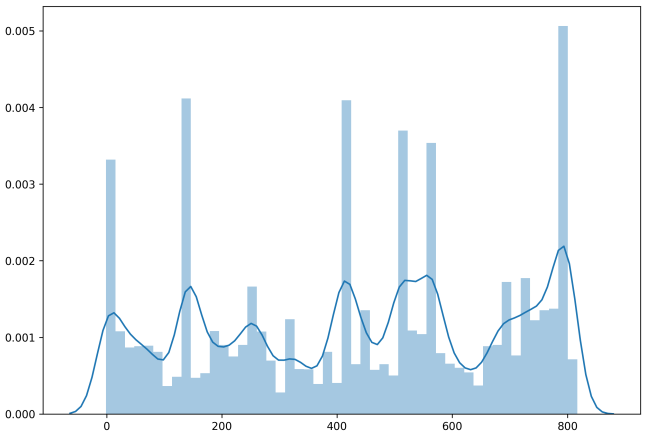

Figure 4: Distribution Plot of the pre-processed dataset as a function of annotated cell-types

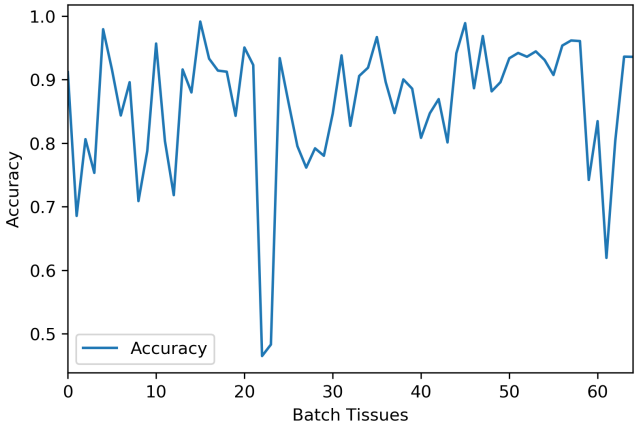

Figure 5: Accuracy for different batch tissues. Along X-axis we have the index for 65 batch-tissues.

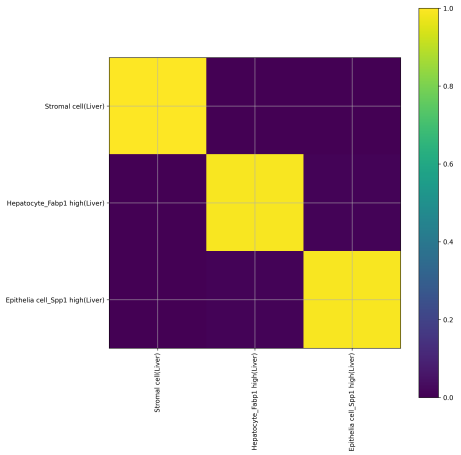

Figure 6: Confusion Matrix for Liver_2 batch tissue

The worst performing batch was MammaryGland.Lactation_1 at an accuracy of $\sim 46\%$. Mamma-

ryGland.Lactation_1 has one of the largest number of cell-types i.e. 89 and 6633 cells with 12731 genes. 80% of the batches has accuracy of more than 80%. The distribution of cells across the different cell types in the batches was not balanced. Hence, we took the macro-average F1 scores to make a better evaluation of the performance of SVM.

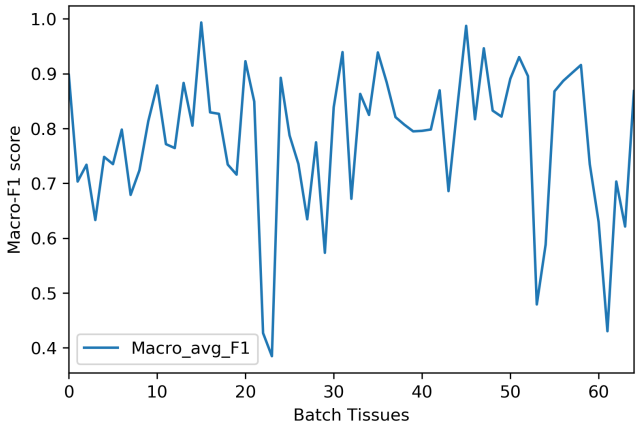

Figure 7: Macro-average F1 scores across the batch-tissues

Fig.7 shows the macro-average F1 scores achieved for different batch tissues. Liver_2 which has the highest accuracy has shown the highest macro-average F1 score of 0.9933. The MammaryGland.Lactation_2 has shown the worst macro-average F1 score of 0.38. The direct correlation between accuracy and macro-average F1 score suggests that SVM is not being dominated by the cell-types having higher populations. However, for these five batches Prostate_1, Prostate_2, TrophoblastStemCells_1, Testis_1, EmbryonicStemCells_1 we found extreme inconsistency between the value of accuracy and F1-score. For all of these batches, accuracy is 90% whereas the the macro-average F1-score ranges from 0.47 to 0.67.
We looked for cell-types showing low F1 scores in these batches. We first investigated the possibility that the model might have misclassified cells from cell-types having low populations. However, this had been ruled out as we found cell-types with lowest population showing high F1 score of 90%.
The second idea was to figure out if the misclassified cell-types have high correlation in gene expression values with any other cell types which might lead to the mis-classification. This possibility has been initially supported by Testis_1 batch tissue which has 18 cell types and 2212 cells in total. The two cell-types Leydig-cell(Testis) and Pre-Sertoli cell_Cst9 high(Testis) have the lowest population in Testis_1 batch. As is evident, the cell population is almost same

Cell-Type	Population	F1-score
Leydig cell(Testis)	25	0.91
Pre-Sertoli cell_Cst9 high(Testis)	20	0

Table 2: Two lowest populated cell-types in Testis_1 batch with their F1 score

for these two cell-types but the F1 scores varies significantly. To understand why Pre-Sertoli cells are being misclassified but Leydig cells are not, we looked into the top 5 correlation scores for these two cell types which is given in Table 3 and Table 4. We figured out that the highest correlation value for Leydig cells is 0.94 whereas for Pre-Sertoli cells all the top 5 correlation scores are 0.98 or higher. This means that there is a high probability of Pre-Sertoli cells being mis-classified as one of these 5 cell-types when the model only takes gene expression values into account.

Cell-Type	Correlation
Pre-Sertoli cell_Ctsl high(Testis)	0.941
Sertoli cell(Testis)	0.933
Elongating spermatid(Testis)	0.927
Pre-Sertoli cell_Cst9 high(Testis)	0.926
Macrophage_Lyz2 high(Testis)	0.917

Table 3: Top 5 correlation scores along with cell types for Leydig-cell(Testis)

Cell-Type	Correlation
Elongating spermatid(Testis)	0.999
Preleptotene spermatogonia(Testis)	0.991
Macrophage_Lyz2 high(Testis)	0.989
Sertoli cell(Testis)	0.989
Spermatids_1700016P04Rik high(Testis)	0.984

Table 4: Top 5 correlation scores along with cell types for Pre-Sertoli cell_Cst9 high(Testis)

3.2.2. 10% Stratified Sub-samples

We took 10% sub-sampled dataset of the original preprocessed dataset where each of the sub-samples contain 10% of total cells. After running SVM on these 10 different datasets we got average accuracy of 65%. In Fig.8 we show the confusion matrix for 25 cell-types which have the most number of cells. One observation we made from here is, TS_Rps28_high(Trophoblast-Stem-Cell) celltype is being mis-classified as TS_Mrpl12_high(Trophoblast-Stem-Cell).

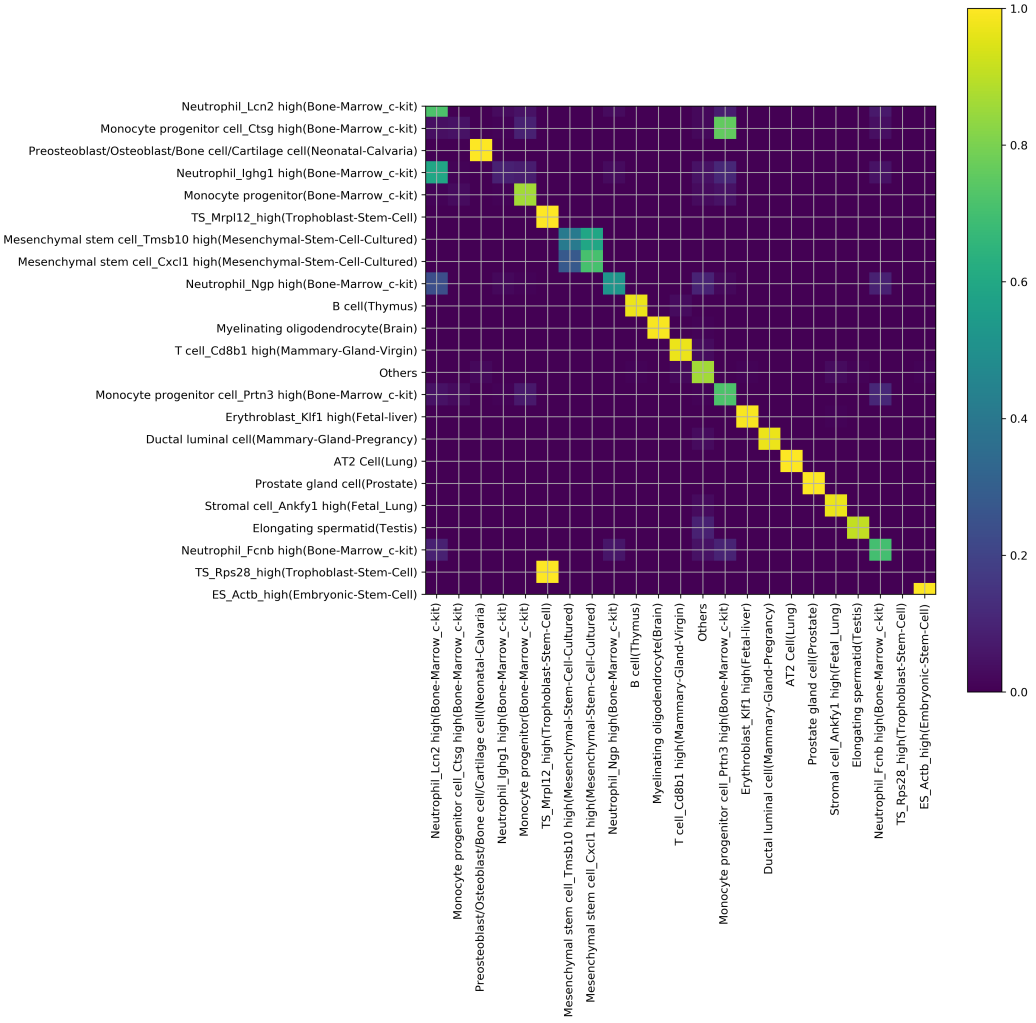

Figure 8: Confusion matrix for top 25 cell-types (by cell count) in one of the stratified sub-samples having 10% of the cells.The rows indicate the actual cell types and the columns indicate the predicted cell types. The legend indicates the normalized counts (Methods 2.4.1). Used SVM here.

3.3. ACTINN

3.3.1. On Preprocessed Dataset

Upon running ACTINN, we got an overall validation accuracy of 64.69% over 10 folds and test accuracy of 64.33% & median test F1-score of 0.61 for a learning rate of 0.0001. To see if changing the learning rate has an effect on performance, we retrained ACTINN with a learning rate of 0.01, which resulted in an overall validation accuracy of 67.99% over 10 folds and test accuracy of 68.34% & median test F1-score of 0.65. Since the accuracy increased by 4% we performed all further analysis on the ACTINN model with learning rate of 0.01.

Comparing the performance of ACTINN on MCA to experiments in Abdelaal et. al.'s paper, we know for Zheng68K dataset (largest dataset), the highest F1-score achieved was by $\text{SVM}_{rejection}$ of 0.92, but the model only captured 61.8% of the cells. ACTINN on the other hand received 0.74 as the F1-score considering all the cells. Although the F1-score on the MCA dataset is 0.65, this is comparable to the F1-score of 0.74 on the Zheng68K data as our MCA dataset is 3.5 times larger and 70 times deeper in cell types.

To further understand if there is a pattern in the high and low F1 scores by cell types we subsetted the top 10% of the cell types ordered by F1 score and all 0 F1 scores. For each subset of cell types we investigate the associated tissue type counts. For the top 10% of the cell types we found the following tissue type count:

We did not notice any alarming numbers for the tissue counts for the top performing cell types. We then investigated the tissue counts for all cell types with an F1-score of 0.

We noticed that there are an alarming number of

Tissue Type	Count
Placenta	10
Pancreas	8
Bladder	5
Lung	5
Neonatal-Muscle	5

Table 5: Top 5 Tissue Counts for top 10% cell types ordered by F1-scores

Tissue Type	Count
Mammary Gland Lactation	68
Embryonic-Mesenchyme	13
Bone-Marrow	12
Neonatal-Skin	10
Lung	8

Table 6: Top 5 Tissue Counts for cell types with F1-score of 0

Mammary Gland Lactation tissue represented in the low performing cell types. Further investigating the cellnames of the Mammary Gland Lactation tissue type that were incorrectly predicted revealed an abundance of Secretory alveoli cell associated cells. Based on this information, we plotted the pairwise correlation of the mean gene expression data for each cell type. From the plot, we notice a significantly

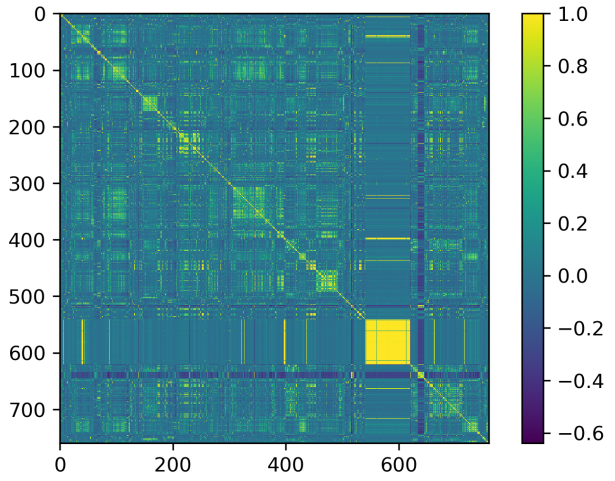

Figure 9: Pairwise Correlation Matrix by Cell Types: Each cell type is represented by the numbers between 0 and 760 on the x & y axes. The legend extends from -0.6 to 1, representing the correlation values.

large box of yellow values around cell types indexed from 550 to 650. The yellow values indicate a correlation value very close to 1. Investigating these cell types lead us to the same cellnames we saw performing poorly with F1 scores of 0, that is, the Secretory alveoli cells of tissue type Mammary Gland Lactation. The high correlation in gene expression among these cell types explain the low F1 score found for them. Removing All Mammary Gland Lactation related cell types and all cell types with less than 10 representative cells, gives us a median F1-score of 0.71 which is a significant improvement from 0.65. Since the number of true cell-types is very large (> 800), we decided to plot the confusion matrix for the top 25 cell-types present in the pre-processed dataset. Fig. 10 shows the confusion matrix for the top 25 selected cell-types. All celltype predictions that fall outside the top 25 cell types are collated as 'Others'. From the plot we can see a clear diagonal suggesting a high true positive rate. The accuracy for this subset of the test dataset was found to be more than the whole dataset at around $\sim 77\%$. The median F1-score of the top 25 cell types by count was 0.71, which is also significantly higher than the overall F1-score of 0.65. Another observation of interest are the count values for the pair Mesenchymal stem cell_Cxcl1 high(Mesenchymal-Stem-Cell-Cultured) & Mesenchymal stem cell_Tmsb10 high(Mesenchymal-Stem-Cell-Cultured) and the pair TS_Rps28_high(Trophoblast-Stem-Cell) & TS_Mrpl12_high(Trophoblast-Stem-Cell). The first pair suggest a correlation between the cell types as the normalized count lies around 0.5. The second pair shows a preference towards TS_Mrpl12_high(Trophoblast-Stem-Cell) cell types as there are more predictions that are of TS_Mrpl12_high(Trophoblast-Stem-Cell) celltype than TS_Rps28_high(Trophoblast-Stem-Cell) cell-type. This is as also seen in the SVM model.

Overall ACTINN was able to produce results that are comparable to the analysis performed by Abdelaal et. al., but this does not suggest ACTINN as a good classifier. An F1 score of 0.65 is low, but considering the scale of the dataset by size and annotations, it is a decent F1 score. Moreover, ACTINN was able to run in 1.16 hours compared to SVM that did not produce a result. ACTINN's ability to produce a result when compared to SVM and the method's ability to produce higher results when high correlation values are removed, makes its a good classifier.

3.3.2. 10% Stratified Sub-sample

For ACTINN, running on the stratified 10% sub-sample resulted in a drastic reduction in accuracy. The mean validation accuracy was 32.62% and the test accuracy was 32.90% with an F1-score of 0.24. This experiment shows us that with the high number of cell

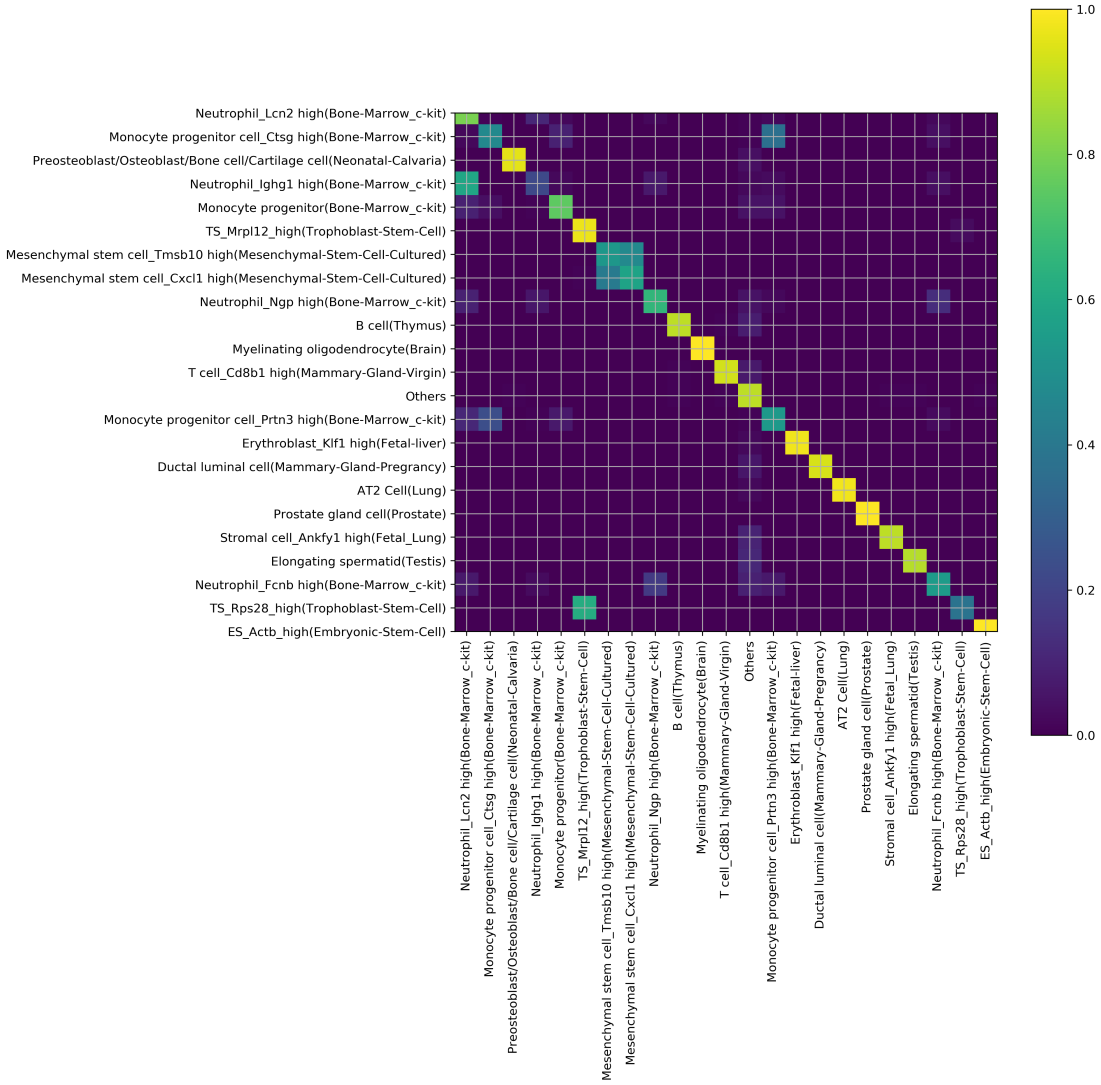

Figure 10: Confusion Matrix for the top 25 cell-types, by count. The rows indicate the actual cell types and the columns indicate the predicted cell types. The legend indicates the normalized counts (Methods 2.4.1)

types and low population per cell type, ACTINN fails to identify cell types correctly and compares to the study performed Abdelaal et. al.

3.3.3. Tissue Type Annotation

The idea behind tissue type annotation is to potentially create a 2 step ensemble method to predict cell types, that is, first predict the tissue type from the PCA reduced count matrix and then predict the cell based on tissue specific models. To get started on this idea, we trained ACTINN to predict Tissue Type Annotation using the preprocessed dataset and found the validation accuracy to be 95.11% and test accuracy to be 95.28% with an F1 score of 0.95. Given the low number of cell types, we plotted a confusion matrix to better understand the distribution of the predicted cell types.

From the plot we notice that two pairs of cell types do not have accurate predictions: 1) True Fetal Kidney & Predicted Kidney and 2) True Bone Marrow & Predicted Peripheral Blood. The low cell count for the Fetal Kidney Tissue explains the missed predictions for Kidney, but the True Bone Marrow Tissue Type has over 500 cells. Another interesting fact about the second pair is that there are other Bone Marrow Tissue Types: Bone-Marrow_c-kit & Bone_Marrow_Mesenchyme, but the Bone Marrow tissues are instead predicted as Peripheral Blood. To investigate, we analyzed the correlation between tissue types. We noticed some correlation between tissue types indexed by the numbers 5 through 12. Taking a closer look, the tissue types in the area are all Fetal type with Fetal Kidney being one of them. From the correlation matrix nothing else is striking, but when we looked at the correlation of the Bone-Marrow Tis-

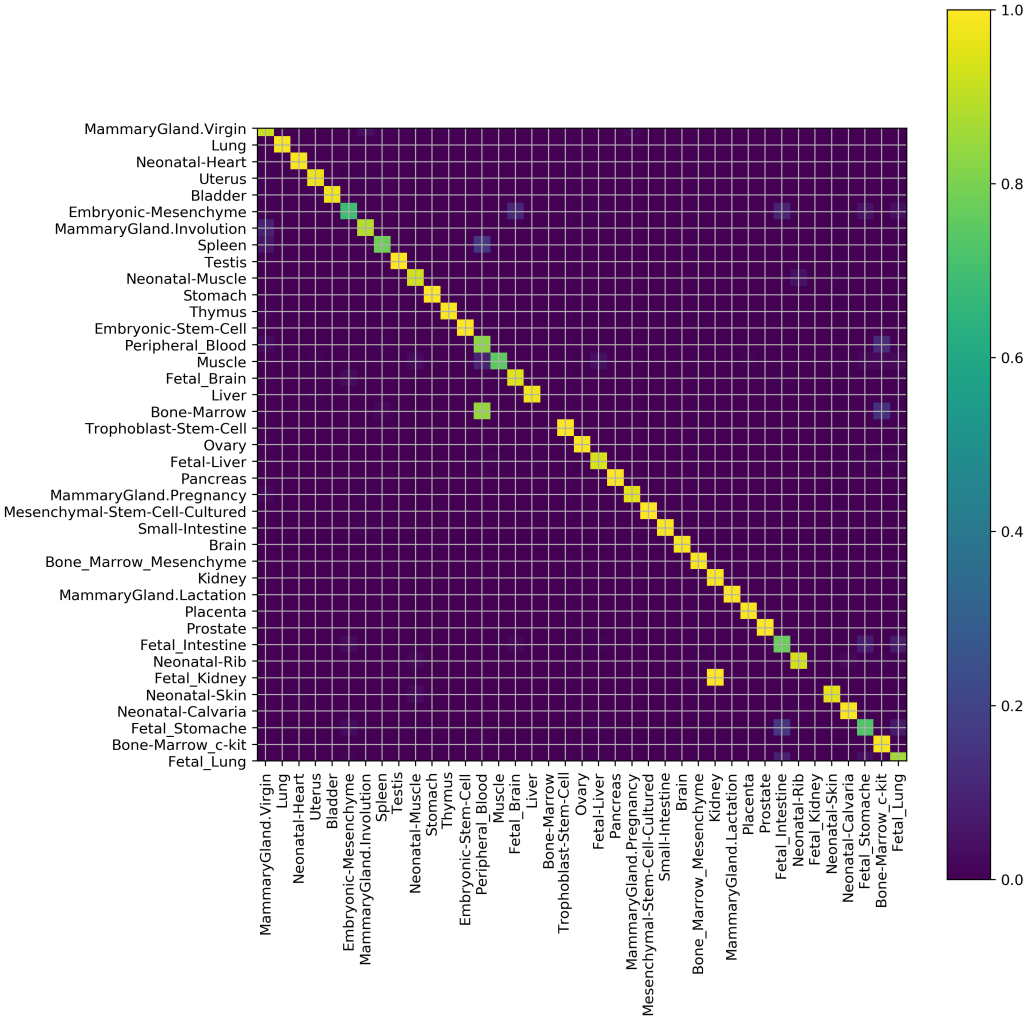

Figure 11: Confusion Matrix for Tissue Type Annotation. The rows indicate the actual tissue types and the columns indicate the predicted tissue types. The legend indicates the normalized counts (Methods 2.4.1)

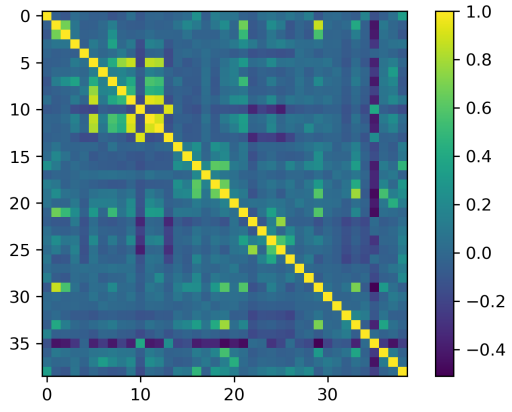

Figure 12: Pairwise Correlation Matrix by Tissue Types: Each tissue type is represented by the numbers between 0 and 39 on the X & Y axes. The legend extends from -0.5 to 1, representing the correlation values.

sue type, it was most correlated to Peripheral Blood Tissue. It is not alarming that Bone-Marrow Tissue is correlated to Peripheral Blood Tissue, but it is alarming that it is not equally correlated to the other Bone Marrow Tissue types. Overall, the ACTINN supervised classifier is an excellent choice for Tissue Type Annotation.

3.4. Scalability

Given the size and depth of annotation of the MCA dataset, scalability became very relevant. Initially we were unable to even pre-process our dataset given the size and ultimately found the recent mid-April release of the Seurat 3.1 pipeline for MCA (Butler, 2018). As for the models, SVM took on average 10 mins for 5-fold cross-validation on each batch specific tissue dataset, but failed to computer when run on the Whole processed dataset; ACTINN took an 1.16 hours for 10-fold cross-validation on the Whole processed

dataset and about 50mins on the Tissue Type Annotation. ACTINN scales much better than SVM but is comparable in performance.

4. Discussion

We studied the performance of the 2 supervised classifiers: SVM and ACTINN on the MCA dataset. Due to the large number of cell counts and annotation types, we wanted to see if there is a pattern between the number of cell counts or cell types which might lead to better or worse classification results. While the Liver_2 batch tissue classification is the best performing, it has one of the lowest cell counts at 261. Similarly, MammaryGland.Lactation_1 which had the worst accuracy surprisingly had one of the largest cell counts present. This result disregarded our assumption that the number of counts will positively affect classification results. Instead, it was interesting to see that most of the misclassified cell-types were due to high correlation of mean gene expression between labels. Both the SVM and ACTINN model had trouble handling Mammary Gland Lactation cell types due to the high correlation. Comparing the models, Abdelaal et. al. found SVM to significantly outperform ACTINN in terms of the F1 score during their evaluations, we noticed the same for the 10% stratified sub-sample scenario, where SVM had an average accuracy of 65% and ACTINN has an average accuracy of 32%. That being said, SVM did not produce any result when run on the complete dataset, but ACTINN acheived 68% accuracy. This suggests that neither SVM or ACTINN can serve as the sole classifier and maybe an ensemble approach will yield a better result. To analyze the potential of an ensemble approach we trained ACTINN on tissue type labels and acheived an accuracy of 95% with the low performing samples explained by high correlation. The accuracy of SVM for cell type prediction per batch tissue type was on average 86%. This is promising for implementing an ensemble approach, but no conclusions can be made about the ensemble approach until further analysis of the combined model performance is conducted. We hope to study a combined approach in the near future to enable better prediction accuracies for large and diverse datasets such as the MCA dataset. With the rise in single-cell RNA sequencing for Cell Atlas development, our study & future work will serve as a stepping stone for robust models in terms of accuracy & scalability (Zheng and Wang, 2019).

References

Abdelaal, T., L. Michielsen, D. Cats et al. 2019. "A comparison of automatic cell identification methods for single-cell RNA sequencing data." *Genome Biol* 20:194.
URL: *https://doi.org/10.1186/s13059-019-1795-z*

Butler, A., Hoffman P. Smibert P. et al. 2018. "Integrating single-cell transcriptomic data across different conditions, technologies, and species." *Nat Biotechnol 36, 411–420 (2018). .*
URL: *https://doi.org/10.1038/nbt.4096*

Han, Xiaoping, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye et al. 2018. "Mapping the mouse cell atlas by microwell-seq." *Cell* 172(5):1091–1107.

Henry, V. J., A. E. Bandrowski, Pepin A-s, B. J. Gonzalez and Desfeux A. 2014. "OMICtools: an informative directory for multi-omic data analysis." *Database* .
URL: *https://doi.org/10.1093/database/bau069*

Luecken, Malte D and Fabian J Theis. 2019. "Current best practices in single-cell RNA-seq analysis: a tutorial." *Molecular systems biology* 15(6).

Ma, Feiyang and Matteo Pellegrini. 2020. "ACTINN: automated identification of cell types in single cell RNA sequencing." *Bioinformatics* 36:533–538.
URL: *https://doi.org/10.1093/bioinformatics/btz592*

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg et al. 2011. "Scikit-learn: Machine learning in Python." *Journal of machine learning research* 12(Oct):2825–2830.

Svensson, V., R. Vento-Tormo and S. A. Teichmann. 2018. "Exponential scaling of single-cell RNA-seq in the past decade." *Nature Protocols* pp. 599–604.

Zappia, L., B. Phipson and A. Oshlack. 2018. "Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database." *PLoS Comput Biol* 14.
URL: *https://doi.org/10.1371/journal.pcbi.1006245*

Zheng, Jie and Ke Wang. 2019. "Emerging deep learning methods for single-cell RNA-seq data analysis." *Quantitative Biology* 7(4):247–254.