Predict the severity and the state of COVID-19 based on the GSE158055 scRNA-seq dataset

Alexander Espig
10037532

Hauke Schüle
10004972

Phuong Mai Pietzonka
10061325

Abstract

The field of NLP is advancing in various domains, including biomedicine. Foundation models are being used to process biological data such as proteins, DNA, RNA, and gene expression values. These models have diverse applications, including disease prediction.
In this project, we utilized the foundation model scGPT to generate embeddings from single-cell RNA sequences of patients with and without COVID-19. We then applied various techniques to process these embeddings and predict the state and severity of the disease. Some of these techniques include clustering, feedforward neural networks, and multi-head classification.
While our baseline clustering approach resulted in poor alignments with true labels, our classification models performed well, with the multi-head classifier achieving the best balance between state and severity classification. Performance improved with larger datasets and we observed that different cell types contributed unequally to the performance of our classification task.
These results highlight the potential of foundation models in biomedical research and emphasize the importance of data quantity and cell type selection in improving disease prediction.

Code: https://github.com/HaukeGS/biomedicine_project/

1 Introduction

Recent advances in the NLP field, such as foundation models, have led to breakthroughs in many downstream tasks and enhanced the ability of models to discern patterns and entity relationships within language (Srivastava et al., 2023). These foundation models, trained on large datasets, have demonstrated strong performance not only in natural language processing but also in analyzing "sentences" composed of proteins, DNA, RNA, and gene expression values (Hao et al., 2024). Their capabilities enable the processing of single-cell RNA sequencing (scRNA-seq) data, providing expression data at the individual cell level (Chen et al., 2019). This performance is evident in tasks such as cell type annotation, perturbation response prediction, and gene network inference (Cui et al., 2024).

Single-cell transcriptomics, in particular, has provided valuable insights into cellular behavior and its relation to disease. These insights are especially useful in understanding conditions like autoimmune diseases and neurodegenerative diseases (Bossel Ben-Moshe et al., 2019; Lee and Lee, 2020). A notable example is the work done by Lee and Lee (2020), where gene expression data from blood samples were processed using deep neural networks to predict the onset of Alzheimer's disease.

In the case of COVID-19, severe infection is characterized by significant changes in the pulmonary immune response, which can be traced at the single-cell level through specific alterations in the transcriptome (Chua et al., 2020; Liao et al., 2020). Studies involving bronchoalveolar lavage and peripheral blood mononuclear cells have shown that COVID-19, particularly in severe cases, disrupts immune system functionality across various immune cell compartments, including monocytes, natural killer cells, dendritic cells, and T cells (Wilk et al., 2020; Lee et al., 2020). However, despite these promising insights, scRNA-seq data is notoriously noisy due to the biological variability inherent in single-cell measurements (Kiselev et al., 2019; Chen et al., 2019), making data interpretation and analysis challenging.

In this project, we aim to develop a model capable of detecting the previously mentioned changes at the single-cell level and classifying the states of COVID-19 using scRNA-seq data. To achieve this, we utilized the foundation model scGPT to generate low-dimensional representations of the original data, so called embeddings. We then developed multiple models to predict both the severity and state of the disease in COVID-19 patients, training these models from scratch with different architectures using the generated embeddings.

2 Related Work

Deep learning models have been widely applied in biomedical research, particularly in disease classification based on gene expression data. Many studies have explored various neural network architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Feed-Forward Neural Networks (FFNs), for analyzing high-dimensional biological datasets.

In Lee et al. (2020), they observed that through the combination of single-cell RNA sequencing and single-cell proteomics of whole-blood and peripheral-blood mononuclear cells to determine the severity of COVID-19 between mild and severe cases. The biological data was preprocessed using `mkfastq` (Cell Ranger 10x Genomics), and low-quality cells (those with low gene variability) were filtered out. The values from the generated data were then normalized, and principal component analysis (PCA) was conducted. Finally, the cells were clustered using unsupervised clustering (with a resolution of 0.5) for classification purposes.

Tabares Soto et al. (2020) investigated the use of machine learning and deep learning techniques to classify cancer types using microarray gene expression data. Their work compared Logistic Regression and CNNs for classification and employed k-fold cross-validation to validate model performance. Similar to their approach, our study utilizes deep learning to classify disease severity, but instead of cancer classification, we focus on predicting COVID-19 severity using embeddings generated by scGPT. Unlike Tabares Soto et al. (2020), we employ an FFN, which is better suited for handling structured feature representations,

Ravindran and Gunavathi (2023) provided a survey on deep learning techniques for gene expression data classification, discussing various models, including FFNs, CNNs, Autoencoders (AEs), and RNNs. The study includes multiple FFN-based approaches used for cancer classification, demonstrating that FFNs are a good method for analyzing high-dimensional biological data. Several FFN-based models reviewed in the survey achieved high classification accuracy, further supporting their effectiveness in disease prediction tasks. Given that scGPT embeddings capture biologically relevant information in a structured format, FFNs are a good choice for classifying COVID-19 severity. Additionally, FFNs offer a computationally efficient (Beskopylny et al., 2020) alternative to more complex architectures like CNNs and RNNs, making them a practical option for our classification task.

3 Dataset & Data preparation

To classify whether a cell originates from a healthy individual or a COVID-19 patient, and to further differentiate between states of the disease or its severity, we will utilize the GSE158055 scRNA-seq dataset (GSE, 2021). This dataset was employed in the study by Zhang et al. (2023a), which aimed to identify condition-specific and cell type-specific regulons across diverse cell types.

The dataset comprises anonymized patient information, including age, sex, and city of residence. However, the most critical aspects are the patient's COVID-19 status and disease severity, which is categorized into three groups: severe/critical, mild/moderate, and a control group. Additionally, the dataset provides details on the patient's COVID-19 test results, reflecting different disease stages: control (uninfected), progression (active infection), and convalescence (recovery phase).

In total the dataset is composed of 1462702 single cell samples, which were gathered from 196 individuals in 284 sequencing samples. The distribution of the labels 'severity' and 'disease state' can be seen in figure 1. 165000 single cell samples were from the control group. 700968 single cell samples were from patients that have shown mild or moderate symptoms and 596734 samples were from patients with severe or critical conditions. 787987 samples were from patients, who were already in a convalescing state, while 509715 samples were from patients, who's illness were still progressing.

This metadata is integrated with single-cell RNA sequencing (scRNA-seq) data, allowing for a detailed examination of gene expression at the individual cell level. The dataset contains 64 distinct cell types, recorded across different patient conditions. However, the distribution of these cell types is highly imbalanced. Some appear as frequently as 227,948 times, while others are observed only 17 times.

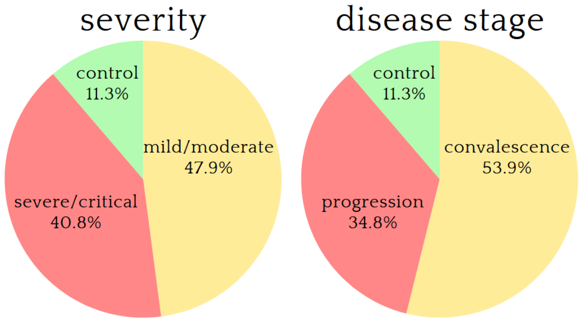

Figure 1: Distribution of labels in percent

4 Approach

First we needed to create embeddings with a pretrained foundation model for our single cell sequencing data. For this we used the single-cell generalized pretrained transformer scGPT. Then we created a baseline for the task by utilizing unsupervised clustering techniques. Following this we used a simple feedforward neural network to validate that the embeddings contained information and to verify our hypothesis that we can classify the severity of the illness by using the generated embeddings. We refined this process and experimented with the hyperparameters and data distributions to increase the F1-Score and the accuracy of our deep learning model.

4.1 scGPT Embeddings

To efficiently train a classifier on single cell patient data, we needed to create embeddings with a foundation model. This way the foundation model encodes important characteristics in a low dimensional vector with signed floating point numbers, decreasing the needed memory and disk space for storing, transmitting and processing the data, as well as enabling a classification head to focus on the critical information in the sequenced single cell sample.

We implemented an embedding pipeline as jupyter notebook on Kaggle, since scGPT had unsolvable dependency issues on Google colab, our preferred platform. For the preprocessing, we only considered the 1800 most variable genes and used a log1p normalization, like the tutorial did. We had to split the dataset into 25 smaller chunks, to fit into memory, processing them one at a time. For each chunk, we generated embeddings, appended the results to a pickle file on disk, and freed any unnecessary memory before moving to the next chunk.

This way we created two main embedding datasets. One with only 5 chunks for validating approaches and one with all chunks for generating the results on the whole dataset, with sizes of 584 MB and 2.85 GB respectively.

4.2 Data preprocessing

After generating the embeddings from the single-cell RNA sequencing data, we observed a significant imbalance in label distribution for disease severity: 700.968 cells labeled as mild/moderate, 596.734 as severe/critical, and only 165.000 as control. To ensure balanced model training, we applied downsampling, adjusting the number of samples per label to a uniform count. This process ensured that all cell types maintained an identical distribution across severity labels, meaning each cell type contributed an equal number of samples for mild/moderate, severe/critical, and control cases. This balance is crucial for preventing model bias towards more frequent labels and improving generalization.

Then we split the dataset, using 80% for training and 20% for testing, while maintaining a similar distribution of severity labels and cell types across both sets. This approach preserves the dataset's diversity of the cell types, ensuring that the model learns meaningful patterns across different conditions.

4.3 Baseline

To establish a reference model for comparing our classification model, we utilized clustering techniques. This allowed us to assess the effectiveness of our classification approach by benchmarking it against patterns identified through unsupervised learning.

As our clustering technique, we used k-means, a simple and efficient algorithm, which is well-suited for large datasets. Its simplicity makes it an effective choice for establishing a baseline to compare our classification model with.

In order to find an optimal number of clusters (k) for our k-Means, we were using the elbow criterion. This is necessary because k-Means does not automatically determine the number of clusters. The elbow criterion is a widely used technique, commonly referenced in research such as in Azar et al. (2013).

To assess the effectiveness of k-means and establishing a baseline, we used the Silhouette Score

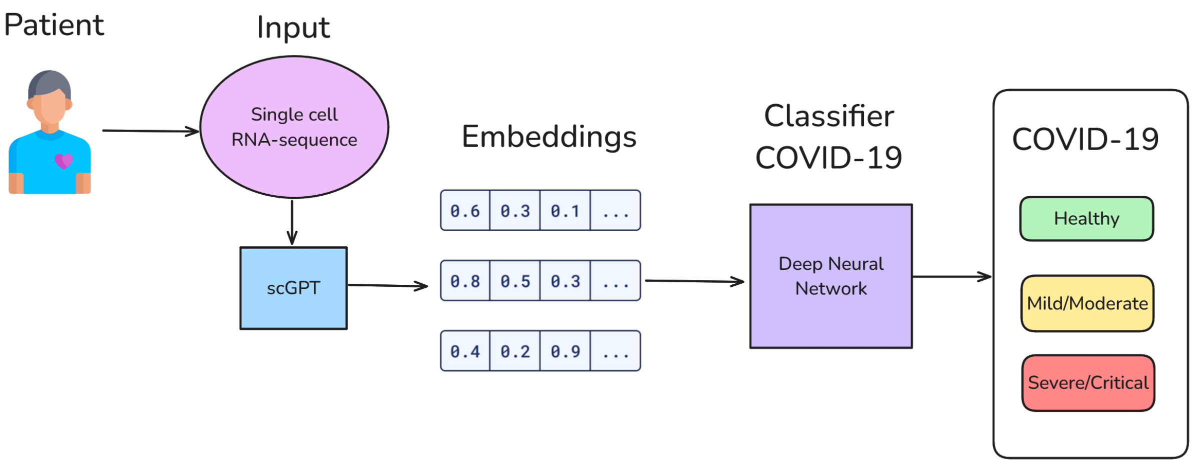

Figure 2: **COVID-19 Severity Classification Using scRNA-seq and Deep Learning**: Our approach classifies the state and severity of COVID-19. We use single-cell RNA sequencing data as input for the foundation model scGPT, which generates embeddings. These embeddings are then processed by a deep neural network to predict the severity of COVID-19, classifying patients as healthy, mild/moderate, or severe/critical.

and the Adjusted Rand Index (ARI). The Silhouette Score measures cluster cohesion and separation, indicating how well data points fit within their assigned cluster and ARI quantifies the similarity between clustering results and ground truth.

Using the elbow method, we determined the optimal number of clusters to be k = 2. With this value, we proceed to initialize k-Means for clustering.

4.4 Deep Learning Classification

We constructed a deep neural network capable of processing the embeddings generated by scGPT and classifying different COVID-19 states and severity levels. The decision to use a deep neural network was inspired by Lee and Lee (2020). However, in our case, we utilized information derived from scRNA-seq data instead of blood cells and focus on a different disease. An overview of our main approach is presented in Figure 2.

Feed-forward classifier For our deep neural network, we implemented feedforward layers to identify hidden patterns in the data and interpret the information for making predictions. The feedforward layers contain between 512 and 2048 neurons, considering that the generated embeddings have a dimensionality of 512. After each layer, we apply the LeakyReLU activation function with $\alpha = 0.2$ to regulate values within the network while preserving negative values. A more detailed display of the layers can be seen in figure 3.

The classifier has three output nodes, each representing a distinct class. For severity prediction, the classes are control, mild/moderate, and severe/critical, as we can observe in Figure 2. For state prediction, the model classifies samples into control, progression, and convalescence.

These raw logits got passed onto the Cross-Entropy-Loss function that is used by the Adam optimizer to update the weights.

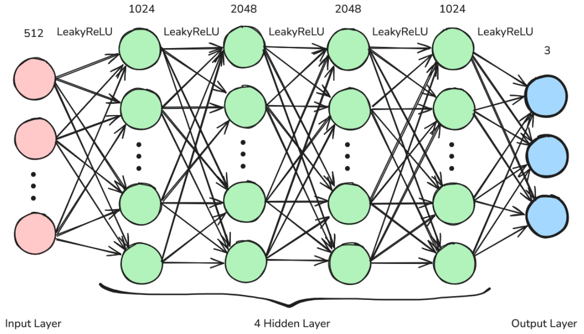

Figure 3: Feedforward Neural Network with number of nodes per layer at the top

Multi-head classifier We hypothesized that the labels for both classification tasks were highly related. This hypothesis was confirmed by analyzing the correlation between labels in the dataset. Based on this observation, we extended our feedforward classifier architecture by adapting it for multi-head classification, inspired in the concept of multi-task learning (Zhang et al., 2023b).

As shown in Figure 4, we introduced two separate feedforward layers with LeakyReLU activation. Each layer serves as a classification head: one head predicts disease severity, while the other classifies the patient's state. This architecture benefits from shared learning between both classification tasks by utilizing a common input encoder while allowing each head to specialize in its respective classification.

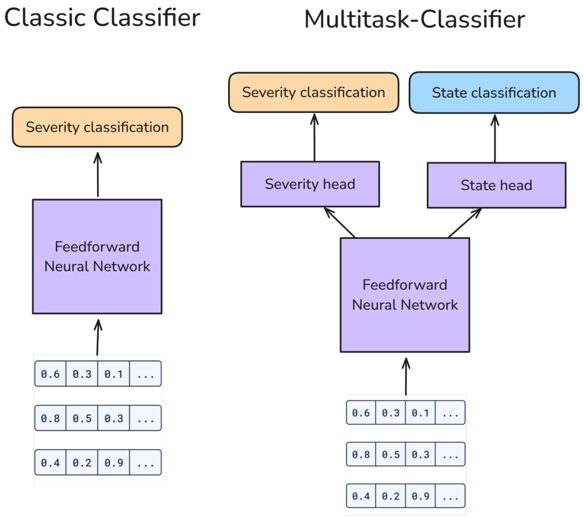

Figure 4: **Architecture of classifiers**: We propose two different architectures for classification. The classic classifier (left) focuses on predicting either the severity or the state of the disease independently. In contrast, the multitask classifier (right) shares a common encoder but incorporates two separate classification heads—one for severity and one for state prediction.

4.5 Data distribution

A relevant hyperparameter to consider for our model is the data distribution, specifically the balance across labels and cell types. As noted by Kiselev et al. (2019); Chen et al. (2019), single-cell RNA sequence data can be challenging to interpret and analyze due to the high variability inherent to single-cell data. This variability makes it difficult to detect and interpret anomalies effectively. However, we hypothesized that some single-cells are more sensitive to diseases as COVID-19, the changes produced by a disease in single-cells are different across all the types and these changes can be found in different levels. Because of this, we consider this the question: "What is more relevant to the performance of the model: More different cell types with less samples per cell type or less, but more prevalent, cell types with more training samples per cell types?"

To investigate this question, we developed an algorithm for the dataset creation. For a given total number of samples and a minimum number of samples per cell type, it samples a subset that meets the specified criteria, while choosing more prevalent cell types over less prevalent ones.

With these subsets we conducted more experiments of which the results can be seen in 5.

Additionally we examined the different cell types for their performances at the same number of samples and displayed the best and worst five cell types for our classifier task. For this experiment we limited the number of samples per celltype to a maximum of 3000 samples and only considered cell types with a minimum number of 500 samples, with an equal label distribution, as always.

5 Results

In this section, we present and analyze the results of our approaches, starting with our baseline using clustering techniques.

Clustering as baseline Using PCA to visualize the results of the k-Means algorithm we can see in Figure 5 very clear separated regions. However taking the previous metrics into account the clustering is not as good as it seems. The Adjusted Rand Index (ARI) for our clustering is -0.0046, indicating that the clustering performance is no better than random and suggesting that the clusters formed by k-Means do not align well with the true labels. Additionally, the Silhouette Score of 0.2024 reveals that the clusters are not well-separated, further indicating that the clustering results are suboptimal.

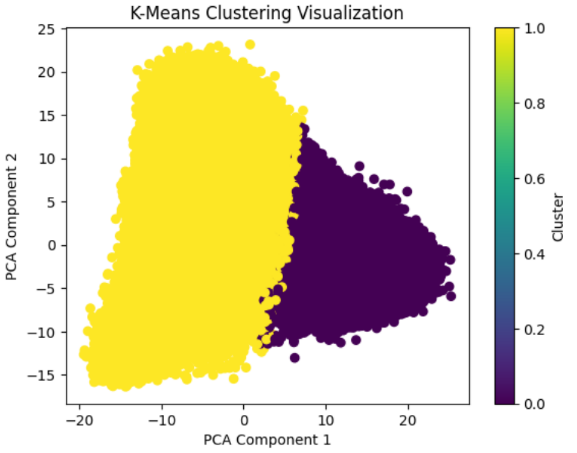

Figure 5: PCA with k-Means

Visualizing the severity of our embedded data as seen in Figure 6, we can see, that the three groups are scattered and well mixed.

To further analyze the data, we calculated the accuracy of our k-Means clustering by comparing

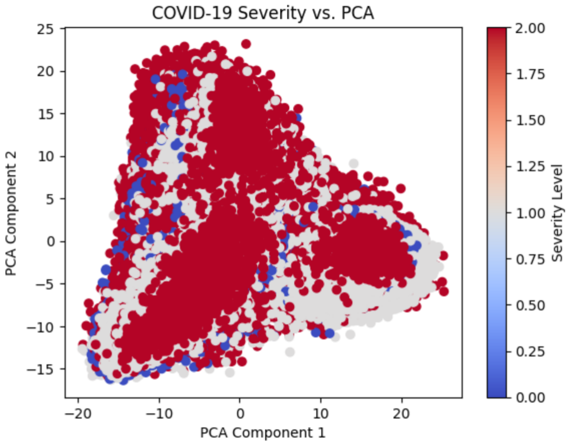

Figure 6: PCA with COVID-19 Severity

its alignment with the true severity labels. Specifically, we determined the majority class within each cluster and assigned it as the predicted label. We then computed the accuracy, which resulted in a value of 0.5025. This outcome is consistent with the other metrics, reinforcing that the clustering performance is suboptimal and only slightly better than random.

Architecture comparison We compared the different architectures introduced in section 4.4. To ensure a fair evaluation, we downsampled all cell types to 3,000 samples where possible. We then balanced the distribution of labels within each cell type. For cell types with fewer samples in a particular label class, we performed additional downsampling within the cell type to maintain a balanced 1-1-1 ratio. For example, if a cell type that had in total 4000 samples, but it had only 300 samples with control as label, we used a total of 900 samples for that cell type, ensuring an even distribution across all classes. However, the overall label distribution remained equal across cell types to prevent bias in the model.

	Severity-Classification		**State-Classification**	
Approach	F1-Score	Accuracy	F1-Score	Accuracy
State Classifier	-	-	.892	.891
Multihead-Classifier	.874	.874	.839	.841
Severity Classifier	.835	.833	-	-

Table 1: **Comparison of architectures** For this experiment, we set a maximum of 3,000 samples per cell type and used the same dataset, split with a fixed seed to ensure reproducibility. We evaluated the model's performance using the macro F1-score and accuracy across all labels.

We applied this data distribution to both classifiers: the classic architecture (used separately for state and severity classification) and the multi-head classifier (which predicts both simultaneously). All classifiers were tested on the same dataset, with slight adjustments in label distribution depending on the specific classification task. As shown in Table 1, the multi-head classifier benefited from the shared encoder, achieving superior performance compared to the severity classifier and slightly outperforming the state classifier. This improvement may be attributed to the training process, where the data distribution prioritized severity labels over state labels. We made this decision to maintain a balanced dataset for performance comparison.

Size of datasets comparison We then experimented with different sample filtering thresholds to evaluate whether increasing the number of samples for certain cell types while keeping others underrepresented could enhance model performance. To investigate this, we conducted experiments with sample limits of 1,000, 3,000, and 10,000, using the classic architecture for severity classification.

Approach	F1-Score	Accuracy
Severity Classifier		
With filter 1000	.800	.798
With filter 3000	.835	.833
With filter 10000	.888	.887

Table 2: **Comparison of dataset sizes** For this experiment, we maintained a balanced distribution of labels across all cell types. The only modification was adjusting the maximum possible number of samples per cell type, referred to as "With filter." We evaluated the model's performance using the macro F1-score and accuracy across all labels.

As shown in Table 2, the model's performance improved as the number of available samples increased. These results suggest that the model can effectively leverage larger datasets, even if some cell types remain underrepresented

Wide vs deep cell type representation As mentioned in 4.5, we conducted more classification experiments on subsets of the data with the same number of overall samples, but with different distributions of cell types. We conducted these experiment series three times: with 90000, 60000 and 30000 total number of samples. The results can be seen in the tables 3, 4 and 5 respectively.

The results of the experiment series with 90000 samples show not much of a difference between the different data distributions. This suggests that

Cell type distribution	F1-Score	Accuracy
90000 samples		
29 CTs á 3103 samples	.869	.867
22 CTs á 4090 samples	.865	.862
18 CTs á 5000 samples	.851	.850
15 CTs á 6000 samples	.848	.848
9 CTs á 10000 samples	.866	.863
6 CTs á 15000 samples	.851	.850

Table 3: **Comparison of cell type distributions.** Each time the subset consisted of 90000 samples. Less than 3103 samples per cell type were not possible while maintaining an equal label distribution.

with a certain number of training examples, the classifier can still generalize well enough, even when only trained with as few as six different cell types. Interestingly enough the F1-score and the accuracy show no improvement for six cell types with 15000 samples over the same six cell types with 10000 samples.

Cell type distribution	F1-Score	Accuracy
60000 samples		
35 CTs á 1714 samples	.852	.850
30 CTs á 2000 samples	.849	.847
20 CTs á 3000 samples	.849	.847
15 CTs á 4000 samples	.842	.841
12 CTs á 5000 samples	.845	.844
10 CTs á 6000 samples	.840	.839
6 CTs á 10000 samples	.852	.851

Table 4: **Comparison of cell type distributions.** Each time the subset consisted of 60000 samples. Less than 1714 samples per cell type were not possible while maintaining an equal label distribution.

The results of the experiment series with 60000 samples show a small drop in F1-score and accuracy when training with a medium number of cell types and samples. This suggests that it benefits the classifier to either consider most of the cell types with some samples or only the few most prevalent ones with a lot of training data, but not the configuration in between, when training with this amount of training data.

Cell type distribution	F1-Score	Accuracy
30000 samples		
37 CTs á 810 samples	.792	.789
30 CTs á 1000 samples	.797	.795
15 CTs á 2000 samples	.795	.793
10 CTs á 3000 samples	.780	.779
7 CTs á 4285 samples	.808	.807
6 CTs á 5000 samples	.797	.796
5 CTs á 6000 samples	.818	.814
3 CTs á 10000 samples	.819	.817

Table 5: **Comparison of cell type distributions.** Each time the subset consisted of 30000 samples. Less than 810 samples per cell type were not possible while maintaining an equal label distribution.

As we expected, the results of the table 5 with total number of samples of 30000 show worse performances in all configurations than with 60000 or 90000 samples. So more training data result in a better classifier, even when discarding some cell types with not enough samples.

Performance per cell type As mentioned in 4.5, we analyzed the different cell types for their performances, which is displayed in table 6. It can clearly be seen that some cell types are much more suitable for the classification task than others, as the range in F1-score for the best and the worst cell type, shows. It seems that there are three cell type in particular that are not well suited for our classification task, as they show a much lower performance than the others.

Cell type	Accuracy	F1-Score
Top 5		
Mono_c4-CD14-CD16	.926	.927
Mono_c2-CD14-HLA-DPB1	.910	.911
Mono_c3-CD14-VCAN	.901	.903
DC_c2-CD1C	.896	.899
T_CD8_c06-TNF	.893	.894
Bottom 5		
T_CD8_c13-HAVCR2	.820	.820
B_c03-CD27-AIM2	.832	.791
DC_c4-LILRA4	.744	.733
B_c05-MZB1-XBP1	.683	.653
Mega	.643	.576

Table 6: **Comparison of cell type performances.** Each cell type had between 500 and 3000 samples for this evaluation.

6 Conclusion

In this project, we explored diverse architectures, techniques, and data distributions to predict disease severity and state using single-cell RNA sequences. We observed that generating embeddings through foundation models preserves valuable information about cells and their states in relation to disease. Furthermore, we found that deep neural networks provided a better understanding of these embeddings compared to clustering for predicting disease state and severity. Additionally, we observed that using a multi-head architecture improved training for the severity-classification in expense of having a worse performance for the state-classification compared to the separate classic architectures.

Another interesting finding was that, as expected, increasing the number of training samples improved model performance. Despite the unbalanced representation of different cell types, both accuracy and F1-score increased as the number of available samples per cell type grew. This suggests that having more training data helps the model generalize better, even in cases where certain cell types are underrepresented.

On the other hand the experiment series on the cell type distributions suggest that the performance increase stops at a certain point of increasing the amount of training data per cell type. Furthermore covering more cell types with less training data per cell type seems to show not much of a difference to only training on the most prevalent cell types with more training data.

Finally, the examination of classification performance per cell type have shown that there are better and worse suited cell types for the classification task. Further work could be exploring this discovery and selecting a curated set of cell types to further improve the model performance.

References

2021. Geo accession viewer. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158055`. Accessed: 2025-01-17.

Ahmad Taher Azar, Shaimaa Ahmed El-Said, and Aboul Ella Hassanien. 2013. Fuzzy and hard clustering analysis for thyroid disease. *Computer Methods and Programs in Biomedicine*, 111(1):1–16.

Alexey Beskopylny, Alexandr Lyapin, Nikita Beskopylny, and Elena Kadomtseva. 2020. Comparison of the efficiency of neural network algorithms in recognition and classification problems. *E3S Web of Conferences*, 224:01025.

Noa Bossel Ben-Moshe, Shelly Hen-Avivi, Natalia Levitin, Dror Yehezkel, Marije Oosting, Leo A. B. Joosten, Mihai G. Netea, and Roi Avraham. 2019. Predicting bacterial infection outcomes using single cell rna-sequencing analysis of human immune cells. *Nature Communications*, 10(1):3266.

Geng Chen, Baitang Ning, and Tieliu Shi. 2019. Single-Cell RNA-Seq technologies and related computational data analysis. *Front Genet*, 10:317.

Robert Lorenz Chua, Soeren Lukassen, Saskia Trump, Bianca P. Hennig, Daniel Wendisch, Fabian Pott, Olivia Debnath, Loreen Thürmann, Florian Kurth, Maria Theresa Völker, Julia Kazmierski, Bernd Timmermann, Sven Twardziok, Stefan Schneider, Felix Machleidt, Holger Müller-Redetzky, Melanie Maier, Alexander Krannich, Sein Schmidt, Felix Balzer, Johannes Liebig, Jennifer Loske, Norbert Suttorp, Jürgen Eils, Naveed Ishaque, Uwe Gerd Liebert, Christof von Kalle, Andreas Hocke, Martin Witzenrath, Christine Goffinet, Christian Drosten, Sven Laudi, Irina Lehmann, Christian Conrad, Leif-Erik Sander, and Roland Eils. 2020. Covid-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nature Biotechnology*, 38(8):970–979.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. 2024. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491.

Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. 2019. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282.

Jeong Seok Lee, Seongwan Park, Hye Won Jeong, Jin Young Ahn, Seong Jin Choi, Hoyoung Lee, Baekgyu Choi, Su Kyung Nam, Moa Sa, Ji-Soo Kwon, Su Jin Jeong, Heung Kyu Lee, Sung Ho Park, Su-Hyung Park, Jun Yong Choi, Sung-Han Kim, Inkyung Jung, and Eui-Cheol Shin. 2020. Immunophenotyping of covid-19 and influenza highlights the role of type i interferons in development of severe covid-19. *Science Immunology*, 5(49):eabd1554.

T. Lee and H. Lee. 2020. Prediction of alzheimer's disease using blood gene expression data. *nature*.

Mingfeng Liao, Yang Liu, Jing Yuan, Yanling Wen, Gang Xu, Juanjuan Zhao, Lin Cheng, Jinxiu Li, Xin

Wang, Fuxiang Wang, Lei Liu, Ido Amit, Shuye Zhang, and Zheng Zhang. 2020. Single-cell landscape of bronchoalveolar immune cells in patients with covid-19. *Nature Medicine*, 26(6):842–844.

U Ravindran and C Gunavathi. 2023. A survey on gene expression data analysis using deep learning methods for cancer diagnosis. *Progress in Biophysics and Molecular Biology*, 177:1–13.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan

Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Reinel Tabares Soto, Simon Orozco Arias, Victor Romero-Cano, Vanesa Segovia, José Rodríguez-Sotelo, and Cristian Jiménez Varón. 2020. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, 6:e270.

Aaron J. Wilk, Arjun Rustagi, Nancy Q. Zhao, Jonasel Roque, Giovanny J. Martínez-Colón, Julia L. McKechnie, Geoffrey T. Ivison, Thanmayi Ranganath, Rosemary Vergara, Taylor Hollis, Laura J. Simpson, Philip Grant, Aruna Subramanian, Angela J. Rogers, and Catherine A. Blish. 2020. A single-cell atlas of the peripheral immune response in patients with severe covid-19. *Nature Medicine*, 26(7):1070–1076.

Lin Zhang, Hafumi Nishi, and Kengo Kinoshita. 2023a. Single-cell rna-seq public data reveal the gene regulatory network landscape of respiratory epithelial and peripheral immune cells in covid-19 patients. *Frontiers in Immunology*, 14.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023b. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.