AARHUS UNIVERSITY

midt
regionmidtjylland

PROJECT IN BIOINFORMATICS (10 ECTS)

Deep Learning for age and sex profiling of immune single cell data.

An image-based approach.

Author:
David Martín Pestaña
Student Number: 202303407

Supervisor:
Dr. Nicolai J. Birkbak
Associate professor

SOLIDUM PETIT IN PROFUNDIS · UNIVERSITAS ARHUSIENSIS ·

Project carried in the
Department of Molecular Medicine
Cancer Evolution Group

July, 2024

Contents

Abstract

The immune system, crucial for pathogen defense and homeostasis, suffers immunosenescence with age, a decline in immune function that increases susceptibility to diseases. This immune aging has a profound influence in human health, likely contributing to age-related diseases such as cancer. However, little is known about the inter- and intra-individual variability in immune health. Gaining insights into immune function concerning age, sex, and diseases will advance personalized decision-making in managing age-related conditions. In this study, we aimed to test the potential of Convolutional Neural Networks (CNNs) in single-cell RNA sequencing (scRNAseq) data from peripheral blood mononuclear cells (PBMCs) to find immune patterns relative to age and sex.

Single-cell data images enclosing gene expression and cell type abundance were created. Then Conv-ISA, an autoencoder CNN was implemented and trained on the images. A particularity about the data is that features were arranged close to each other based on biological reasons, rather than the traditional linear approach to omics data analysis. The findings indicate that biologically significant data arrangements improve pattern recognition as Conv-ISA successfully predicted sex. However, it faced challenges on age group predictions.

Another objective of the study was to focus on model interpretability. Using Integrated Gradients framework, we found that the model was predicting sex putting more weight into two sex dimorphic genes (ZNRF3 and SLC36A1) and the T lineage of lymphocytes, known to be different between males and females. Age prediction was done mainly based on naïve T cells, which are known to decrease with age due to thymus decay and antigen exposure throghout life.

Our study explores promising approaches for scRNAseq data analysis but also highlights the need for further refinement of neural network architectures and exploration of alternative deep learning approaches. Future research could integrate clinical data with immune patterns to enhance personalized medicine approaches for diseases with immune component.

Acknowledgements

I would like to express my gratitude to Nicolai Birkbak for giving me the opportunity to work in his group and for always motivating me to push a bit further, and to Mateo Sokač for the countless ideas to improve this work. I would also like to thank Johanne Ahrenfeldt and Randi Istrup for all their time during the meetings and their valuable suggestions. I would like to extend my thanks to every other researcher and staff member at the Department of Molecular Medicine and Bioinformatics Research Centre. They have made this first year abroad much easier; working with such kind and supportive people has been an absolute pleasure. Lastly, I could not forget my parents, family, and close friends. Although many kilometers away, I have been lucky to always have their support during this project.

1 Introduction

The immune system is a complex network responsible not only for defending against pathogens but also for maintaining the homeostasis of the organism. Broadly, the immune system is composed of an innate component, which is faster in its response but less specific, and an adaptive component, which responds through antigen-specific recognition [1].

As humans age, the immune system undergoes significant changes, characterized by an impairment in immune function known as immunosenescence [2]. This age-dependent immune decline is known to make individuals more susceptible to a wide range of diseases, as well as chronic inflammation and tissue damage [1, 2]. The role of immunosenescence in widely known diseases such as neurodegenerative disorders and malignancies has been broadly discussed, but more insight is needed as current therapeutic guidelines often overlook the individual variability of the process.

Relative to sex, males and females show significant differences in their immune responses, which correlate with susceptibility, response, and outcomes to a wide range of conditions [3]. For instance, autoimmune diseases are more common in females, while the incidence of malignancies and infections is higher in men [3]. These differences are thought to be due to a differential immune cell landscape between males and females, as well as the earlier degeneration of the thymus in males, a process believed to be a driver of the immunosenescence process [1, 4].

The immune system is shaped mostly by individual life history, with a genetic component that loses relevance with age [2]. Therefore, the inter-individual variability in immune aging is significant, which hinders the ability to standardize treatment protocols when facing age-related diseases and highlights the need to further study the dynamics of immune function over time [1, 2].

Single-cell RNA sequencing (scRNAseq) is a high-throughput technology that allows for the profiling of individual cells. This technology makes it possible to study cell heterogeneity, abundance, and expression as a whole, which is not possible with other bulk sequencing methods [5]. Information contained in scRNAseq data of immune cells has proven to be highly valuable in recent studies, allowing for the identification of new cell populations and the characterization of novel interactions within the system [6].

By studying scRNAseq data from peripheral blood mononuclear cells (PBMCs), it is possible to investigate the cellular and molecular mechanisms underlying immunosenescence in a way that traditional methods cannot achieve. This approach could identify novel age-related changes in gene expression, signaling pathways, and immune function, as well as sex-specific differences in these features. The insights gained from studying the immune system at a single-cell resolution could be crucial for understanding the implications of inter-individual variability in immunosenescence and creating better-tailored protocols for age-related diseases [7].

The analysis of the vast amount of data generated by scRNAseq technologies poses a significant challenge, often aggravated by formatting differences depending on the technology used [8]. However, deep learning, coupled with advanced computational power, offers powerful tools for the analysis of complex datasets and could help uncover relevant information from scRNAseq data [8, 9].

Convolutional Neural Networks (CNNs), which were originally developed for image recognition, could help overcome one significant difficulty of traditional biological data analysis methods. Historically, data has been arranged in a linear or tabular manner, which overlooks the relative positioning of elements within organisms, an aspect crucial for the correct functioning of biological systems [10]. The spatial awareness of CNNs could help better capture the complexities of biological data. Recent advances in this area have shown positive results when using CNNs for integrated analysis of multi-omics data arranged in a biologically significant way, capturing relevant genetic interactions and predicting crucial parameters such as the stage of malignancies [11].

Machine learning has recently revolutionized various fields such as computer vision and voice and speech recognition, to name but a few. However, the advances of artificial intelligence (AI) within medical research have progressed at a slower pace [12]. Apart from regulatory hurdles, the importance of interpretability of prediction results in health research is one of the main reasons behind the slower adoption of "black-box" machine-learning models [13, 14]. To address this, an increasing number of frameworks are being developed to interpret model results, such as DeepExplain and Integrated Gradients [15–17]. These techniques allow the evaluation of the relative importance of each predictor variable in the final prediction given by the model, conferring a degree of interpretability that facilitates the use of deep learning in omics analysis.

In the context of scRNAseq, understanding the importance of each feature in data analysis from PBMCs could determine which genes and cell types are most relevant in driving the predicted outcome, effectively pinpointing the drivers of immune aging.

Overall, profiling age in PBMCs using scRNAseq relative to sex could be a promising approach for unraveling the complexities of immunosenescence. The use of scRNAseq could provide valuable information for characterizing cellular heterogeneity and inter-individual variability in the immune system. The new insights gained, when coupled with deep learning approaches like CNNs, could contribute to developing a more targeted approach to age-related diseases.

2 Hypothesis and objectives

Understanding immunosenescence is a difficult task due to the inter-individual variability as well as the innate complexity of the immune system. The study of information contained in single cell RNA sequencing data could improve the understanding of immune aging, but its analysis is challenging due to its characteristics, mainly high dimensionality, volume and sparsity. Creating new data analysis methods combining the pattern recognition power of Convolutional Neural Networks with frameworks providing interpretability could give significant insights into the aging process within the immune system. Moreover, the power of these technologies could be enhanced by making use of its spatial awareness by training on data arranged in a biologically significant order. The newly generated knowledge could be used to improve the management of age-related affections, providing a more personalized approach to treatments.

Therefore, the aims of this project were:

1. Generate single-cell data images that accurately reflect expression levels and cell abundance, with pixels arranged in a biologically significant manner. This process will involve careful preprocessing and normalization of the data to ensure that the resulting images are suitable for analysis by CNNs.
2. Implement and train a CNN on these images to identify and extract age- and sex-dependent patterns in immune features. The CNN will be designed to handle the complexity of single-cell data, allowing it to distinguish different immune profiles across different age groups and sexes.
3. To avoid the typical black box effect of deep learning and enhance the interpretability of the model, implement integrated gradients to provide an attribution score to each feature. This approach will enable a clearer understanding of how specific features contribute to the models predictions and offer valuable insights into the biological significance of the identified patterns.

3 Methods

3.1 Single cell data processing

For this study, data was retrieved from the OneK1K cohort, a publicly available single-cell RNA sequencing dataset from over 1.27 million Peripheral Blood Mononuclear Cells (PBMCs), collected from 982 healthy individuals [18]. Cell type for each observation was predicted using CellTypist open-source tool [19, 20].

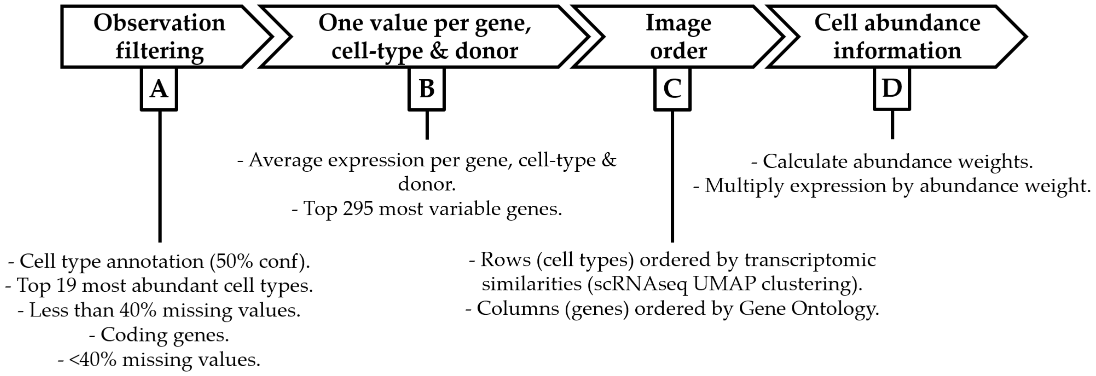

Observation filtering
One value per gene, cell-type & donor
Image order
Cell abundance information
A
B
C
D
- Average expression per gene, cell-type & donor.
- Top 295 most variable genes.
- Calculate abundance weights.
- Multiply expression by abundance weight.
- Cell type annotation (50% conf).
- Top 19 most abundant cell types.
- Less than 40% missing values.
- Coding genes.
- <40% missing values.
- Rows (cell types) ordered by transcriptomic similarities (scRNAseq UMAP clustering).
- Columns (genes) ordered by Gene Ontology.

Figure 3.1: *Schematic workflow of data pre-processing pipeline.*

A pre-processing pipeline was designed for scRNAseq data (figure 3.1). Firstly, observations were filtered to deplete low confidence observations and retain the most significant data. A confidence threshold of 50 % was set for the cell type prediction and observations from the top 19 most abundant cell types were selected. In addition, only coding genes present in more than 60 % of observations were selected for further analysis (figure 3.1-A).

Because of the theoretical pillars of convolutional neural networks, where patterns are spatially recognized, every image needed to be populated by the same combination of cell type and gene in each pixel. The single cell sequencing concept deviates from this idea, as not every sampled individual contains the same cells sequenced, so the second step of the preprocessing pipeline consisted of data reformatting steps (figure 3.1-B). The average expression for each unique combination of cell type, gene and patient was calculated, effectively creating scRNAseq pseudo-bulk data to be used as input to the framework.

Due to the architecture of the neural network, further discussed in section 3.3, data images were decided to be 19 pixels high (cell types) and 295 pixels wide (genes). Therefore, to complete the first part of the pipeline, the 295 most variable genes were selected based on their standard deviation (figure 3.1-B).

3.2 Image design

To benefit from the spatial awareness of convolutional neural networks, rows representing cell types and columns representing genes were ordered in a biologically significant manner. Cell types were arranged according to a previously reported hierarchical clustering based on transcriptional similarities [18]. To create an order for genes, the interactions between the 295 entries were analyzed using STRING tool [21] with 50 % confidence interval, no text-mining or co-occurrence and MCL clustering with inflation parameter of 4. The interacting clusters of genes were placed together in the image center, populating the edges with the non-interacting ones.

The previous steps resulted in one 19 row and 295 column image for each individual. The subsequent analysis was branched into two different strategies.

- Strategy A: Inclusion of cell type abundance.

  Averaging the data in a pseudo-bulk approach derived in a loss of details about abundance of each cell type, which could enclose relevant information. Within strategy A, weights for were calculated dividing the proportion of each cell type in each individual by the average proportion of the cell type across the dataset (equation 3.1). Expression data was then multiplied by these weights so each value in the image accounted for the cell type abundance as well as single cell expression.

- Strategy B - Exclusion of cell type abundance:

  To be able to compare the impact of accounting for cell type abundance, a second strategy that did not apply the previously described weights was created. Data was instead directly passed to the next steps.

$$Weight\ (cell\ type,\ individual) = \frac{p\ (cell\ type,\ individual)}{\overline{x}\ p\ (cell\ type)} \tag{3.1}$$

For both strategies, values were then scaled using min-max normalization. Percentile 95 value was used as maximum for each gene and cell type to avoid possible outliers weighing down the rest of the data.

The described process resulted in sets of images which were now suitable as input to the convolutional neural network.

3.3 Network architecture

A four-part autoencoder without reconstruction penalization was implemented using PyTorch framework (figure 3.2). Batches of images are inputted at each framework run and each of the four modules modify and extract information in the following steps.

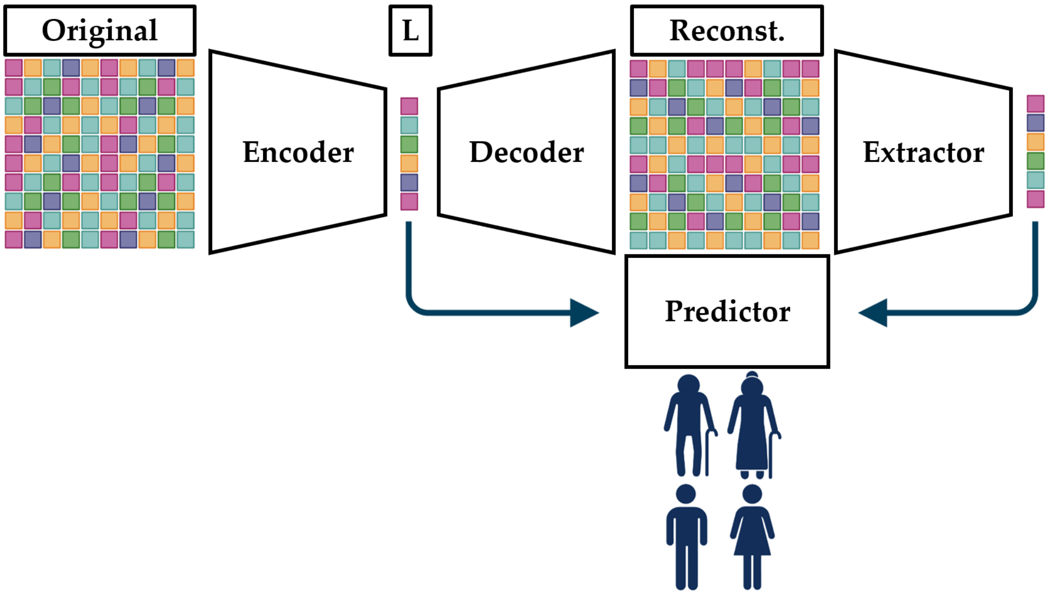


Figure 3.2: *Schematic representation of the implemented four-part autoencoder.*

- Encoder: Images are reshaped through four layers of convolutional filters, resulting in a flat vector with the aim of enclosing the most relevant information into a small, linear data unit. The result of this step is referred to as latent representation (L).

- Decoder: The decoder was designed to take as input the latent representation and, through four transposed convolutional layers, reconstruct an image of similar shape to the original. Rather than creating an image equal to the original, the goal of this module was to reconstruct the most optimal image for extracting patterns, hence the lack of reconstruction error in this step.
- Extractor: The extractor module, through four sets of convolution and max-pooling layers, aims to extract the most relevant information from the reconstructed images. Tensors are flattened into a linear vector at the end of this step.
- Predictor: The predictor module takes as input the concatenation of the latent representation (L) and the vector resulting from the extractor to account for those features encountered in the original images (latent vector) and those extracted from the reconstructed image. Through linear transformations and a softmax activation, it results in a vector with the predicted probability of each possible outcome.

The outcome predictions were then evaluated in a validation split of the dataset, using Cross Entropy Loss as cost function. In order to integrate precision and recall in one metric, as well as to have an evaluation parameter independent of the batch size, F1 score was also recorded. Making use of the latent vector (L), UMAP representations were created to evaluate the ability of ConvISA to cluster the groups together and apart from each other.

Accessory to the network, an optional hyperparameter optimization workflow was implemented using Optuna framework [22]. Upon activation of this setting, a range is set for each hyperparameter instead of a fixed value and multiple trials are run, gradually adjusting each parameter. The metric decided to optimize the parameter tuning phase was the F1 score of the validation split, as loss score is dependent on the batch size and, hence, not comparable between the optimization trials. In order to get more insight into the model, a relative importance of each hyperparameter is given as output after the Optuna framework run. Parameters that can be tuned as well as search spaces used in the subsequent trials can be seen in table 3.1.

Table 3.1: Optimizable parameters and search space used

Parameter	**Search space ranges**
Learning rate decrease start	15 - 30
Learning rate decay	10^{-7}-10^{-4}
Weight decay	10^{-8}-10^{-5}
Learning rate	10^{-6}-10^{-3}
Batch size	64 - 300
Dropout rate extractor	0.05 - 0.5
Dropout rate predictor	0.05 - 0.5
Train/test split percentage	0.5 - 0.65
Patience	5 - 10
Loss jitter	Yes/No
Channel growth between layers	*

* Discrete range of numbers, not search space

Additionally, to avoid overfitting, the following measures were taken:

1. Pruning or early stopping was included using patience
2. Growth of tensors was controlled by parameter optimization, using a discrete range of multipliers for the tensors not to grow more and make the network more complex.
3. Dropout rates were added to both extractor and predictor layers, both parameters were optimizable but neither of them could be 0 in the search space definition.

3.4 Training scenarios

In order to check if single cell data could be useful to predict immune aging relative to the sex of individuals, the combination of the three following scenarios were tested.

1. Prediction of sex or age. The ability of the framework to predict sex and age based on scRNAseq data of PBMCs was tested. Sex prediction was a binary classification problem while age prediction conformed a multi-class classification task. Age bins were created as stated in table 3.2.

2. Rescaling of the data. The original OneK1K dataset contained an unbalanced number of samples across the groups (table3.2). To study the impact of it, a second, upscaled dataset was created by resampling with replacement data from the same sex and age group. In order to avoid overfitting or having the same data image in a train and test split at the same time, distortion was applied to each of the resampled observations.
3. Separation of males and females during training. Previous studies have reported a differential single cell transcriptome between females and males. Such differences could hinder the capacity of the convolutional neural network to find age-dependent transcriptomic variation. In order to compare the performance of the framework and gather the most information, age-related trials were conducted under two conditions: with males and females separated, and with them combined.

Table 3.2: Age groups created and characteristics of cohort

Age group	Ages	Non-Resampled	Resampled
Age group 1	< 54 years old	138F I 94M	200F I 200M
Age group 2	55-64 years old	106F I 75M	200F I 200M
Age group 3	65-74 years old	164F I 130M	200F I 200M
Age group 4	> 75 years old	157F I 117M	200F I 200M

As depicted in figure 3.3, six main trials were run in total, each of it with parameter optimization enabled and the search space for each of the parameters described in table 3.1. Additionally, to study the extent to which the order of the features in a biologically significant way impacted the network performance, an extra trial was run using images in which gene and cell type order was randomized across the whole set.

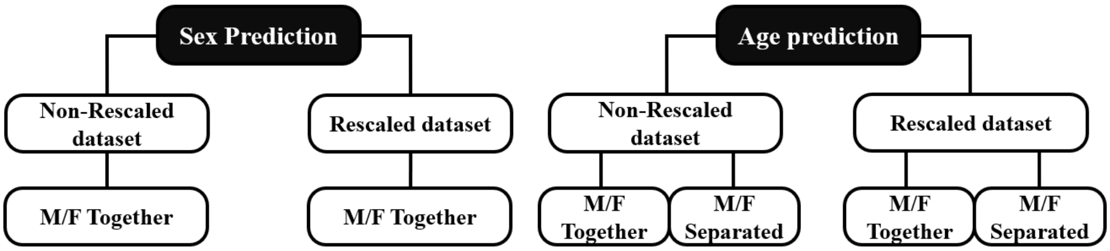


Figure 3.3: *Main trials run to test the CNN performance.*

3.5 Model interpretability

Once the parameters were tuned and the network was trained, the best performing model was saved to study the relative importance of each feature (gene and cell type) when creating the predictions. In order to achieve this, the stored model was analyzed with the Integrated Gradients method [16, 17], that based on the gradient associated to each feature relative to the predicted outcome, calculated the weight that each feature has on the prediction.

From the resulting attribution score, heatmap images were created to study which clusters of cells and genes were driving the predictions of the network.

3.6 Code availability

The resulting framework created was named Conv-ISA (Convolutions for Inference of Sex and Age) and it is fully accessible and documented as a public GitHub repository (https://github.com/Davidmrn5/convISA). The framework can be run with custom data or with the example data from OneK1K cohort from a bash command. The full run will transform the expression data into images, optimize the hyperparameters (if desired), train the network and interpret the most relevant combination of gene and cell type at driving the predictions.

In order to run it, it must be provided with files where the single cell data expression, gene and cell type ordering and clinical information for each individual are stated. In addition to that, a configuration file with the desired options is also to be present when running the framework.

The individual formatting of each of the files with examples of them as well as detailed instructions to using Conv-ISA can be accessed at the README file of the given repository.

4 Results

4.1 Model performance

The capacity of Conv-ISA to predict sex and age bins after being trained in non-resampled and resampled datasets was tested. The characteristics of the cohort in each of the datasets are shown in table 3.2. Every trial within this section was run with weighted average for cell type abundance, parameter optimization framework activated and using the search space in table 3.1.

4.1.1 Model was capable of separating males and females based on scRNAseq data.

Figure 4.1 shows the loss (Cross Entropy) relative to the training epochs for both training and validation sets, validation F1 score of the best performing model is shown below each trials loss plot. Both models reached the same apparent minimum loss in similar number of epochs, achieving a loss of 0.32 in both training and test sets. Relative to the F1 score, both training trials performed equally well, achieving a test F1 of 0.932 for the non-resampled dataset and 0.965 on the resampled dataset.

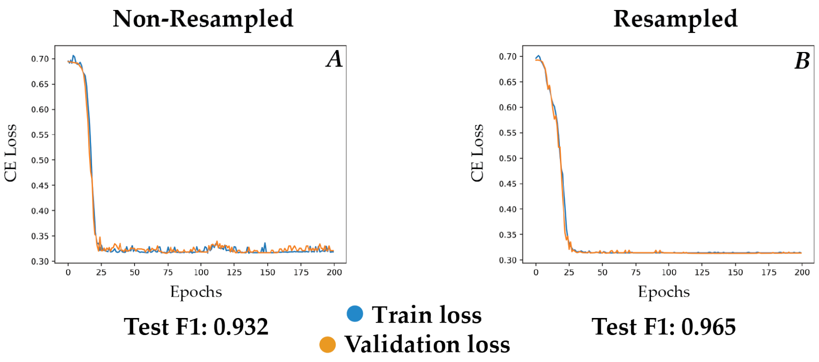


Figure 4.1: *Performance of Conv-ISA in sex prediction.*

Although no significant differences were seen in the training outcome of the resampled and non-resampled datasets, a smoother loss profile is shown in the resampled dataset. Moreover, the loss plot shows a very similar train and test loss in every area, which is indicative of little overfitting of the model, as it can achieve similar results in both splits.

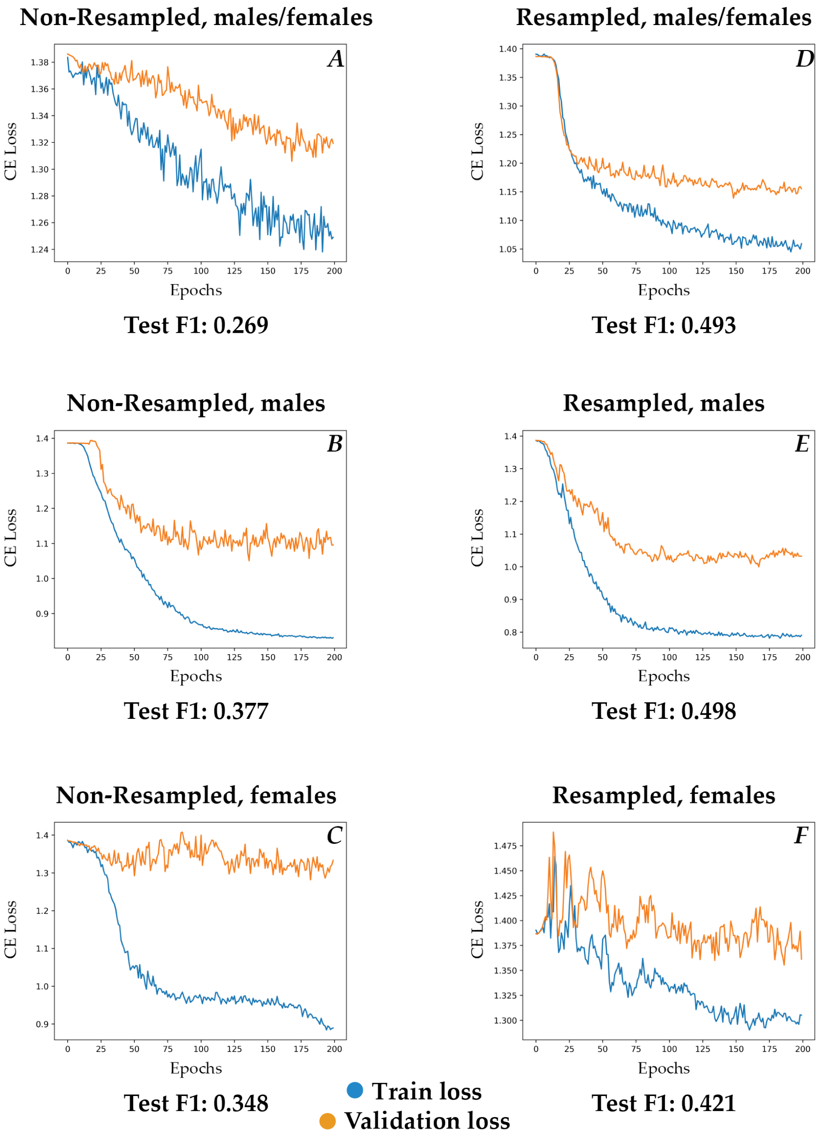
4.1.2 Prediction of age groups is less successful than sex, with worse performance on females than males.

The model showed less proficiency at predicting age groups based on the single cell landscape of individuals. The loss plots of the six trials for age prediction mentioned in the previous methods section are shown in figure 4.2. Validation F1 score for the best performing model of each trial is shown below its corresponding loss plot.

Neither of the training scenarios performed optimally, however, some observations could be made from these results. Firstly, all the results showed validation loss having a higher plateau than training loss, which could suggest overfitting of the network. The difference between training and validation loss seemed to be negatively correlated to the sample size in the sets, with higher variation in trials where sample size was smallest (males and females trained separately, non-resampled datasets).

Comparing the resampled with the non-resampled dataset when males and females were trained together (Figure 4.2-A,B), a better performance of the model was shown when resampling was done, with a cross entropy validation loss of 1.33 for the non-resampled dataset and 1.2 for the resampled dataset. The profiles of the loss plot were more optimal for the resampled dataset, as a better decreasing profile was appreciated.

When training males and females separately, the resampled dataset (Figure 4.2-E,F) showed a marginally better performance than the non-resampled one (Figure 4.2-B,C). However, there was an accused difference between male and female training performance. While models trained with data from males showed validation losses between 1 and 1.15, performance in females was never better than 1.3 in neither of the datasets. All in all, the model showed poor results at age group prediction in any of the training scenarios.



Test F1: 0.269

Test F1: 0.493

Test F1: 0.377

Test F1: 0.498

Test F1: 0.348

Test F1: 0.421

Figure 4.2: *Performance of Conv-ISA in age prediction.*

4.1.3 Latent representation of transcriptome can capture relevant patterns in sex better than age.

UMAP representation for the latent vectors encoded in the resampled datasets are shown in figure 4.3. UMAP plots of resampled and non-resampled trials showed no significant differences, henceforth only those of resampled datasets are shown.

UMAP plot in the context of sex (figure 4.3-A) shows a clear separation in the sex patterns of scRNAseq data, with clustering done more clearly for males than females, where some outliers and more spread can be seen.

Conv-ISA was less successful when identifying and encoding age-dependent patterns into the latent vector (figure 4.3-B,C,D). Although some clustering can be seen for the different age groups, there is no clear separation between these when training males and females together neither when done separately. Among the three plots related to age, the UMAP representation created for males showed a marginally better separation than the rest.

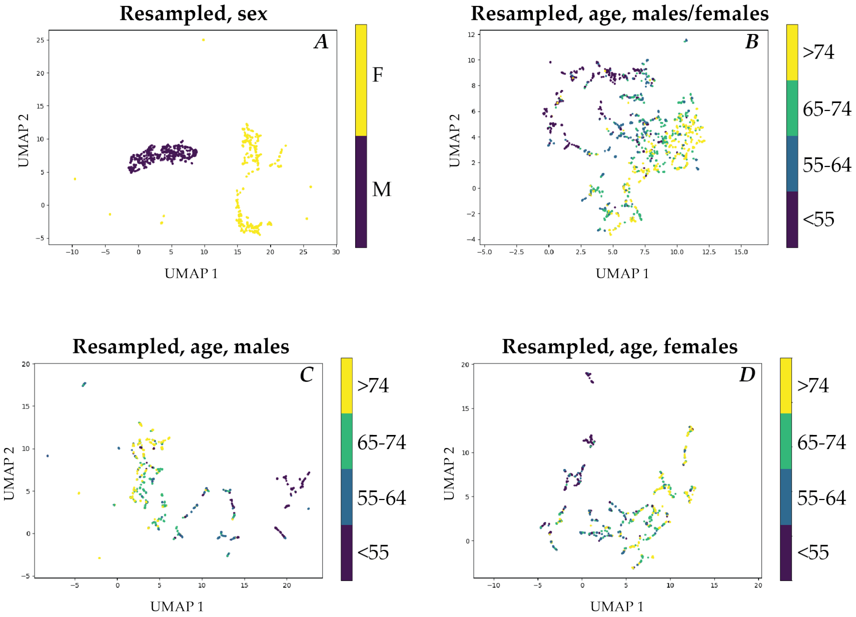


Figure 4.3: *UMAP representation of the features encoded in the latent vector in the different trials.*

4.2 Order of features inside of image

To check whether the convolutional neural network could find patterns better thanks to a biologically meaningful ordering of the data, a dataset in which features were randomized was created. Then, the performance of the model at predicting sex was compared between the randomized and the ordered dataset. The comparison between loss plots of the best performing models with randomly ordered features with its ordered counterpart is shown in figure 4.4.

The model trained on features arranged in a biologically significant order achieved a Cross Entropy loss of 0.32 for the validation split, while training in the randomly ordered image trial suffered early pruning at a two-fold higher validation loss. In addition, the model trained on randomly ordered images showed a separation between training and validation loss, sign of overfitting which was absent on the model trained on ordered features.

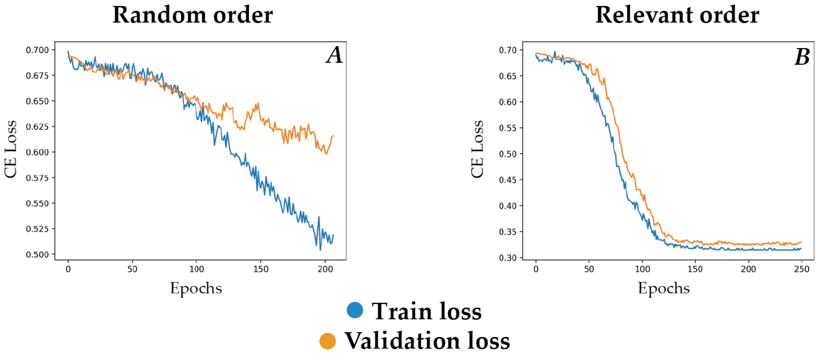


Figure 4.4: *Comparison of model performance between randomly ordered images and those ordered in a biologically meaningful way.*

4.3 Model interpretability

To know which cell types and genes were driving the model predictions the most, Integrated Gradients analysis was performed. The study was done for the resampled datasets, using the cell type abundance weights and without using the cell type abundance weights to compare which patterns did Conv-ISA detect in each scenario. Results of the analysis are shown in figure 4.5. Attribution for age prediction is shown with males and females trained together as no significant differences were observed when trained separated. Each of the heatmaps shows the absolute attribution of each feature. The bar plots within the axes represent the relative attribution for cell types and genes, respectively. The relative attribution was calculated as in equations 4.1 and 4.2 and it served as a measure of how important one particular gene or cell type was compared to the rest of genes/cell types.

$$Relative\ attribution\ (cell\ type) = \frac{\sum_{gene=1}^{genes=295} attribution\ (gene,\ cell\ type)}{Total\ attribution} \tag{4.1}$$

$$Relative\ attribution\ (gene) = \frac{\sum_{cell\ type=1}^{cell\ types=19} attribution\ (gene,\ cell\ type)}{Total\ attribution} \tag{4.2}$$

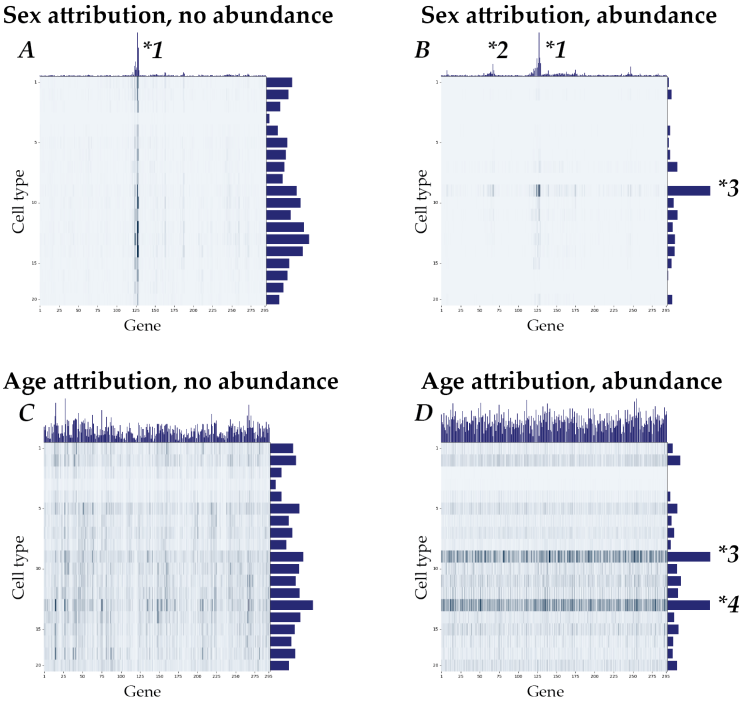


Figure 4.5: *Absolute (heatmap) and relative (axis bar plots) attribution for each of the features. Overrepresented clusters are shown with a "*".(*1) ZNRF3, (*2) SLC36A1, (*3) Naïve helper T cells and (*4) Naïve cytotoxic T cells.*

When studying the features responsible for sex prediction ignoring cell abundance (figure 4.5-A,B), a cluster of 10 genes could be seen overrepresented in the heatmap (figure 4.5-(*1)), among them, the most prominent was Zinc And Ring Finger 3 (ZNRF3). The same analysis weighting based on cell abundance seemed to be more informative, as, apart from the forementioned cluster, it also identified a second cluster of genes which had the Solute Carrier Family 36 Member 1 (SLC36A1) as main entry (figure 4.5-(*2)). In respect to cell type attribution, it showed overrepresentation of the T-cell lineage, more in particular naïve helper T cells (figure 4.5-(*3)).

For the attribution of features relative to age prediction (figure 4.5-C,D), no significant gene cluster seemed overrepresented in neither of the analysis. However, the study done accounting for cell abundance showed two cell types with a significantly higher attribution score than the rest. The two overrepresented cell types were naïve helper T cells (figure 4.5-(*3)), and naïve cytotoxic T cells (figure 4.5-(*4)).

5 Discussion

A significantly different scRNAseq landscape was found between males and females after the use of Conv-ISA, suggesting that taking advantage of the spatial awareness of convolutional neural networks on data arranged in a biologically meaningful way was a suitable strategy. The use of CNNs on images created from omics data has been discussed before [11], as well as the use of scRNAseq data for immune profiling [23]. However, to our best knowledge, no study has been performed with a similar pseudo bulk approach of single cell data of PBMCs.

In order to test the ability of the CNN to identify patterns on the data, sex and age prediction was performed. When studying sex relative expression, important differences were found between both sexes, as shown by a decent performance on sex prediction (figure 4.1) and a clear separation in UMAP plots of latent representation in males and females (figure 4.3-A).

Differences in the immune system of males and females have been previously reported, with males suffering from immunosenescence earlier in life than females [3]. This is thought to be driven by a premature thymus decay in males, which leads to lower lymphocyte counts and consequently, an earlier immune aging pattern [24]. In addition, the existence of differentially expressed genes between sexes, also known as sex dimorphic genes, is a widely known fact [25].

Our results were in line with the general knowledge, as the neural network was able to find differential patterns in the scRNAseq data of males and females. When giving interpretability to the sex prediction through integrated gradients, two gene clusters stood out, with the most prominent entries for each cluster being ZNRF3 and SLC36A1.

ZNRF3 is a ligase that participates in the Wnt pathway and has been previously reported to be related to mammalian sex determination [26, 27]. Therefore, the fact that the CNN gives the most weight to the expression of ZNRF3 for sex prediction correlates with the actual knowledge that we have on said gene.

SLC36A1 is an amino acid transporter proposed to be part of exocytosis in neurons and gut transportation [28]. The protein has been previously reported as dimorphic in mice [29] which could also explain why Conv-ISA finds sex-dependent expression patterns in the gene.

When accounting for cell type abundance and weighting the expression values based on it, the CNN gave more weight to the naïve helper T cells in order to predict sex. The naïve T cells are originated in the thymus and are one of the main effector cells for the adaptive immune response. Upon cytokine signalling from innate immune cells, such as macrophages and dendritic cells, they differentiate into different helper lineages, mainly Th1 for defense against intracellular pathogens (viruses or bacteria) or Th2 in the case of extracellular threats (parasites) [4]. A differential adaptive immune response has been previously reported between males and females, suggesting a higher Th1/Th2 activating cytokine ratio in males due to the different effect of sex hormones in the immune system [30]. Therefore, the fact that Conv-ISA gives more weight to naïve T cell abundance over other PBMCs agrees with the actual knowledge on the topic.

The performance of the network on age group prediction was suboptimal with clear signs of overfitting in every training scenario. Although tools such as early pruning, tensor size and layer number optimization were used to prevent overfitting, we suggest that the network might be too complex for the data and sample size used. Therefore, increasing the sample size and simplifying more the net could be new approaches to improve the performance of Conv-ISA [31]. One possible approach to simplification would be changing the architecture of the CNN so it predicted directly on the latent representation of the genome, instead of going through the reconstruction process, this would make it simpler and could make the predictions more accurate.

However, the CNN seemed capable of finding patterns in the data, as shown by the weights given to different cell types when accounting for their abundances. In this case, Conv-ISA seemed to predict age giving more importance to the T cells, mainly naïve helper T-cells and naïve cytotoxic T-cells. As previously discussed, the loss of a functional thymus is believed to be one of the key drivers of immunosenescence in humans [4]. The fact that the CNN gives most importance to cells coming from the T lineage would correlate with actual knowledge.

Furthermore, naïve T cells are those that have not been exposed to antigens yet. It has been suggested that, as individuals age, the increased antigen exposure during the lifespan leads to a decreased naïve/non-naïve cell ratio [32]. This could explain the higher importance that Conv-ISA gives to the abundances of naïve T cells when predicting age.

On a more methodological side, we aimed to check whether the biologically significant arranging of scRNAseq data of PBMCs within the image made a difference on the model performance. After comparing a dataset with columns and rows ordered in a random way with the order of cell types based on transcriptome and genes based on interactions, we checked that Conv-ISA performed significantly better when there was a meaningful order. This correlates with previous research carried out on CNNs based on omics data [11]. It suggests that the linear or tabular approach traditionally used for analysing biological data is suboptimal when compared to methods that can take advantage of the spatial positioning of elements. We hypothesize that the reason behind this is that organisms are highly organized systems, where every element must be in the right place for them to work, therefore, adopting more natural arrangements of data leads to better pattern finding.

6 Conclusions and future research

- The biologically significant arrangement of single cell images improved the pattern recognition ability of the convolutional neural network. Information on cell type abundance enclosed in scRNAseq data was crucial to overcome the drawbacks of averaging expression values in the pseudo-bulk approach to the analysis.
- The implementation of the CNN worked at predicting sex based on scRNAseq data from PBMCs but did not work as well for the prediction of the different age groups. It seemed to suffer from overfitting, presumably because of a too complex architecture.
- The use of methods that allow for better interpretability of the network proved to be highly valuable for giving biological sense to the predictions. The two genes given more weight for sex prediction (ZNRF3 and SLC36A1) are previously described dimorphic genes. The network also predicted sex and age based mainly on naïve T cell abundance, which are known to vary between sexes (different adaptive immune function) and ages (due to thymus decay).

Overall, the study of single cell RNA sequencing data from PBMCs using convolutional neural networks proved to be a valuable approach for identifying patterns in data arranged in a biologically meaningful way. However, the complexity of the techniques used, as well as the intricacy of the data, led to a sub-optimal accuracy at assessing the immunological age of individuals.

Looking ahead and with more time, there are many ways in which this study could be improved. Firstly, different architectures could be tried on the CNN, presumably simpler ones, to overcome the overfitting problem shown by Conv-ISA. Apart from that, the multidimensional capacities of the network were not exploited in this project; using more data layers, combining more omics techniques, or using a second layer for cell type abundance instead of the weight approach used could improve the performance of the model. In light of deep learning technologies, other approaches rather than CNNs could also be used to find complex patterns in data; foundation models leading into the use of transfer learning would be a new yet interesting approach to the analysis of immune cell scRNAseq data.

Once a good performing yet interpretable model is found, data could be used in combination with clinical data from patients. This could aid to find correlations between the immune state of each individual and parameters such as progression or response to therapy in diseases with relevant immune components such as cancer, hopefully leading to a more personalized approach to medicine.

Bibliography

(1) Weyand, C. M.; Goronzy, J. J. *Annals of the American Thoracic Society* **2016**, *13*, S422–S428.

(2) Alpert, A. et al. *Nature Medicine* **2019**, *25*, 487–495.

(3) Huang, Z.; Chen, B.; Liu, X.; Li, H.; Xie, L.; Gao, Y.; Duan, R.; Li, Z.; Zhang, J.; Zheng, Y.; Su, W. *Proceedings of the National Academy of Sciences* **2021**, *118*, DOI: 10.1073/pnas.2023216118.

(4) Sandstedt, M.; Chung, R. W. S.; Skoglund, C.; Lundberg, A. K.; Östgren, C. J.; Ernerudh, J.; Jonasson, L. *Immunity & Ageing* **2023**, *20*, DOI: 10.1186/s12979-023-00371-7.

(5) Zhu, H.; Chen, J.; Liu, K.; Gao, L.; Wu, H.; Ma, L.; Zhou, J.; Liu, Z.; Han, J.-D. J. *Science Advances* **2023**, *9*, DOI: 10.1126/sciadv.abq7599.

(6) Chu, Y. et al. *Nature Medicine* **2023**, *29*, 1550–1562.

(7) Ginhoux, F.; Yalin, A.; Dutertre, C. A.; Amit, I. *Immunity* **2022**, *55*, 393–404.

(8) Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Wang, B. **2023**, DOI: 10.1101/2023.04.30.538439.

(9) Yu, X.; Xu, X.; Zhang, J.; Li, X. *Nature Communications* **2023**, *14*, DOI: 10.1038/s41467-023-36635-5.

(10) Franke, M. et al. *Nature* **2016**, *538*, 265–269.

(11) Soka, M.; Kjær, A.; Dyrskjøt, L.; Haibe-Kains, B.; JWL Aerts, H.; Birkbak, N. J. *eLife* **2023**, *12*, DOI: 10.7554/elife.87133.

(12) Benjamens, S.; Dhunnoo, P.; Meskó, B. *npj Digital Medicine* **2020**, *3*, DOI: 10.1038/s41746-020-00324-0.

(13) Rudin, C. *Nature Machine Intelligence* **2019**, *1*, 206–215.

(14) Picard, M.; Scott-Boyer, M.-P.; Bodein, A.; Périn, O.; Droit, A. *Computational and Structural Biotechnology Journal* **2021**, *19*, 3735–3746.

(15) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. *PLOS ONE* **2015**, *10*, ed. by Suarez, O. D., e0130140.

(16) Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. In *International Conference on Learning Representations*, 2018.

(17) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks, 2017.

(18) Yazar, S. et al. *Science* **2022**, *376*, DOI: 10.1126/science.abf3041.

(19) Domínguez Conde, C. et al. *Science* **2022**, *376*, DOI: 10.1126/science.abl5197.

(20) Xu, C.; Prete, M.; Webb, S.; Jardine, L.; Stewart, B. J.; Hoo, R.; He, P.; Meyer, K. B.; Teichmann, S. A. *Cell* **2023**, *186*, 5876–5891.e20.

(21) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C. *Nucleic Acids Research* **2014**, *43*, D447–D452.

(22) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework, 2019.

(23) Mogilenko, D. A.; Shchukina, I.; Artyomov, M. N. *Nature Reviews Immunology* **2021**, *22*, 484–498.

(24) Calabrò, A.; Accardi, G.; Aiello, A.; Caruso, C.; Candore, G. *Frontiers in Aging* **2023**, *4*, DOI: 10.3389/fragi.2023.1272118.

(25) Yang, X.; Schadt, E. E.; Wang, S.; Wang, H.; Arnold, A. P.; Ingram-Drake, L.; Drake, T. A.; Lusis, A. J. *Genome Research* **2006**, *16*, 995–1004.

(26) Hao, H.-X. et al. *Nature* **2012**, *485*, 195–200.

(27) Harris, A. et al. *Proceedings of the National Academy of Sciences* **2018**, *115*, 5474–5479.

(28) Boll, M.; Foltz, M.; Rubio-Aliaga, I.; Daniel, H. *Genomics* **2003**, *82*, 47–56.

(29) Gabory, A. et al. *PLoS ONE* **2012**, *7*, ed. by Aguila, M. B., e47986.

(30) Roved, J.; Westerdahl, H.; Hasselquist, D. *Hormones and Behavior* **2017**, *88*, 95–105.

(31) Charilaou, P.; Battat, R. *World Journal of Gastroenterology* **2022**, *28*, 605–607.

(32) Li, M.; Yao, D.; Zeng, X.; Kasakovski, D.; Zhang, Y.; Chen, S.; Zha, X.; Li, Y.; Xu, L. *Immunity &; Ageing* **2019**, *16*, DOI: 10.1186/s12979-019-0165-8.