






GOWDL: gene ontology-driven wide and deep learning model for cell typing of scRNA-seq data

Antonino Fiannaca [†], Massimo La Rosa [†], Laura La Paglia , Salvatore Gaglio  and Alfonso Urso 

Corresponding author: ICAR-CNR, National Research Council of Italy, Via Ugo La Malfa 153, 90146, Palermo, Italy. Tel.: (+39) 3394235272;

E-mail: antonino.fiannaca@icar.cnr.it

[†]Antonino Fiannaca and Massimo La Rosa contributed equally.

Abstract

Single-cell RNA-sequencing (scRNA-seq) allows for obtaining genomic and transcriptomic profiles of individual cells. That data make it possible to characterize tissues at the cell level. In this context, one of the main analyses exploiting scRNA-seq data is identifying the cell types within tissue to estimate the quantitative composition of cell populations. Due to the massive amount of available scRNA-seq data, automatic classification approaches for cell typing, based on the most recent deep learning technology, are needed. Here, we present the gene ontology-driven wide and deep learning (GOWDL) model for classifying cell types in several tissues. GOWDL implements a hybrid architecture that considers the functional annotations found in Gene Ontology and the marker genes typical of specific cell types. We performed cross-validation and independent external testing, comparing our algorithm with 12 other state-of-the-art predictors. Classification scores demonstrated that GOWDL reached the best results over five different tissues, except for recall, where we got about 92% versus 97% of the best tool. Finally, we presented a case study on classifying immune cell populations in breast cancer using a hierarchical approach based on GOWDL.

Keywords: Deep learning, scRNA-seq, Gene ontology, Cell typing, Marker genes

INTRODUCTION

Recent development in single-cell RNA-sequencing (scRNA-seq) technology allowed innovative discoveries in the biomedical field. scRNA-seq refers to the sequencing of single-cell genomic or transcriptomic profiles from a specific tissue of origin [1] to study quantitative variations in RNA compositions into a particular cell type. scRNA-seq has many advantages: for example, it evidences the cell heterogeneity of a specific tissue sample, both in terms of gene expression patterns inside a single cell population and in terms of the quantitative representation of different cell populations into a particular sample. Moreover, scRNA-seq characterizes rare cell types, such as rare tumor cells or hyper-reactive immune cells; it can compare different expression patterns in different cell types within different tissues or different biological conditions [2]. It is also possible to evidence and analyze different subclasses of the same cell class, for instance, different T or B cell sub-populations [3]. Indeed, the analysis of the variation of the

quantitative composition of different sub-populations of immune cells in different tissues (healthy versus disease) is a relevant biological task [4], as it represents an essential milestone for medicine in studying diseases and their treatments (i.e. immune cell behavior in cancer treatment). Automatic cell classification tasks, also known as cell typing, through clustering, classification and lineage tracing, have proven to be a valid instrument for scRNA-seq analysis [5–7]. In particular, cell typing is applied to easily and quickly characterize the different cell-type compositions in a specific tissue [7]. Cell-type classification methods can be described by two, not mutually exclusive, points of view: biological knowledge-based and computational-based. From the biological knowledge point of view, there are cell-type classification strategies that compare similarities between a single cell and a bulk or single-cell RNA-seq profile reference database. In these cell-type classification strategies, it is possible to include SingleR [8], SCINA [9], SciBet [10], scPred [11], scmap [12], CHETAH

Antonino Fiannaca is a Researcher at the High Performance Computing and Networking Institute of the National Research Council of Italy (ICAR-CNR). He received a PhD in Computer Science from the University of Palermo, Italy. He mainly works on artificial intelligence models for bioinformatics and precision medicine.

Massimo La Rosa is a Researcher at the Institute of High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). His research interests include bioinformatics, computation biology, and machine learning.

Laura La Paglia is a Researcher at the Institute of High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). Her research interests include bioinformatics, computation biology, genetics, and molecular biology.

Salvatore Gaglio is a Full Professor at the Department of Engineering, University of Palermo, Palermo, Italy. Also, he is a research associate at the High Performance Computing and Networking Institute of the Italian National Research Council (ICAR-CNR). His research interests include expert systems, artificial intelligence, and robotics.

Alfonso Urso is a Senior Researcher in systems and computer engineering at the High Performance Computing and Networking Institute of the Italian National Research Council (ICAR-CNR). His research interests are in the areas of machine learning, soft computing, and applications to bioinformatics.

Received: May 03, 2023. **Revised:** August 17, 2023. **Accepted:** September 4, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

[13], Garnett [14], scID [15] and SCSA [16]. All of these approaches, on the one hand, allow fully automatic and rapid classification of cell types on a wide variety of test datasets; nevertheless, on the other hand, they severely limit the extrapolation of cell types, which is highly dependent on the reference dataset. Furthermore, gene relationship-based classification techniques integrate information about the relationships between genes within a cell into an algorithm. Among them, it is possible to include scDeepSort [17] and sigGCN [18]. Finally, the a priori knowledge-based techniques exploit knowledge of marker genes or, more generally, biologically relevant to a case study. SCSA [16] and scCATCH [19] belong to this category. From the computational point of view, the exponential growth in the number of cells and samples has led to the adaptation and development of several methods for automatic cell identification. Some of these approaches exploit statistical models [10, 16]. Moreover, recently developed artificial intelligence approaches based on machine learning and deep learning architectures have been demonstrated to learn complex relationships and integrate existing knowledge and big data. Among them, two possible approaches relate to the learning strategy. In unsupervised learning, the input not labeled data are grouped together based on their similarities or differences, and each cluster is given a class label. Cell-type clustering algorithms using these techniques are scmap [12], scCATCH [19] and scID [15]. The other approach is supervised learning, where the training data are labeled, and the algorithm assigns directly test data to a specific category. Cell-type classification algorithms using that approach are SingleR [8], scClassify [20], scAnnotatR [21] and ACTINN [22].

This paper presents a new computational architecture for cell-type classification called gene ontology-driven wide and deep learning (GOWDL). According to the above-mentioned concepts, it exploits the knowledge provided by a set of reference genes implementing a supervised classification model with no need for a preliminary clustering procedure. GOWDL is based on two main features to be exploited to build our cell-type predictor. First of all, we consider the semantic similarity among genes in terms of functional proximity. For this reason, we compute a similarity distance based on the gene functional annotations provided by the Gene Ontology (GO) repository [23]. Secondly, we take advantage of the focused knowledge supplied by a set of marker genes, also called bio-relevant genes, that characterize as a signature of a specific cell type. The proposed computational model, then, integrates both kinds of data using a wide and deep learning (WDL) architecture [24], where the deep neural network represents the supervised classification system. It is implemented through a customized version of a convolutional neural network (CNN) [25], dealing with functional information. In contrast, the wide layer is implemented with a linear model that analyzes the gene signature. Original WDL, which considers a feed-forward deep neural network as the deep layer, has been used for cell typing in the work of [26].

This paper represents an extension and an improved version of our previous work [27]. The main improvements of this work are the following: we introduce a novel computational layer, namely the kernel dataset creation (KDC) layer, that allows embedding into the input dataset some functional annotations; we test our predictor with external independent datasets; we compare with 12 other state-of-the-art predictors; we present a case study about the classification of closely related T cells in breast tissue using a hierarchical structure of our classifier.

MATERIALS AND METHODS

Datasets

We selected different datasets containing scRNA-seq expression profiles from several human tissues to evaluate the ability of the proposed architecture to classify different cell types in several contexts. We used data from both solid and liquid tissues, including blood, breast, kidney, lung, pancreas and melanoma. Datasets are split into training sets and test sets to validate the proposed classifier with external data never seen by the model during the training phase, including the cross-validation procedure. We did not consider a breast tissue test set because we introduced a specific case study involving immune cell populations in breast cancer. Datasets were downloaded from several repositories such as Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/>), Human Kidney Atlas database (kidneycellatlas.org), Human Cell Landscape (CHL), Immune Cell Atlas (<http://immunecellatlas.net/>), Single Cell Portal (https://singlecell.broadinstitute.org/single_cell). The leading information about the datasets is summarized in Table 1, whereas in Supplementary materials S2 (Tables T1 and T2), we reported more details, including the number of cells, the number of genes and the cell types for each train and test dataset. All data were normalized using the ‘Log normalize’ function implemented in the Seurat package [28]. Bio-marker genes were considered for each tissue to characterize a specific gene signature associated with a particular cell type using the CellMatch database [19].

CellMatch is a comprehensive cell taxonomy reference database for healthy and cancerous tissues, providing information on human and mouse marker genes in different cell types and related cell subtypes.

Proposed pipeline

Figure 1 shows our proposed pipeline. Here, we briefly describe it, then each block will be detailed in the following subsections. The initial preprocessing step is composed of three consecutive stages. The number of cells and genes is filtered due to different criteria. Then the filtered dataset is split into two training sets with the same number of cells: one with the expression values of only the relevant genes and the other with the remaining genes. After the preprocessing, we implement the learning model through the GOWDL architecture, which integrates a customized version of a convolutional neural network (CNN layer) and a linear model (wide layer). GOWDL uses GO terms to encode functional annotations into the network.

Preprocessing

Training datasets have been preprocessed as follows. Preprocessing comprises three consecutive steps: cell filtering, gene filtering and dataset splitting (Figure 1). In cell filtering, we removed cells belonging to underrepresented cell types (<5%), as done in [17]. Then we removed cells belonging to cell types that do not contain any marker gene provided by the CellMatch DB. In gene filtering, we considered only the genes having GO annotations for each sample. GO provides a framework and a set of concepts, called GO terms, for describing the functions of gene products from all organisms. All genes not present in GO are discarded. This filtering step is essential for further analysis, which involves the computation of a functional distance among GO terms. In the last preprocessing stage, we split the filtered dataset into two new datasets. One dataset contains all the expression values of the genes marked as relevant genes according to CellMatch. The other

Table 1. Overview of 11 chosen scRNA-seq human tissue datasets, six for training and five for external testing.

Dataset	Tissue	Source	Dimensions cells * genes	# of Cell-types	Cell-types
Training	Blood	[29] retrieved from Single Cell Portal: SCP345	13 316 * 21 814	7	T cell, CD14+CD16- monocyte, CD14+CD16+ monocyte, Natural killer cell, Naive B cell, Dendritic cell, Memory B cell
	Breast	[30] retrieved from Single Cell Portal: SCP1106	24 271 * 28 118	16	Myeloid, Epithelial cell, CD8+ T cell, T helper cell, Plasma cell, CAF, B cell, T Reg, T cell, Endothelial cell, T cell Cycling, Follicular Helper, PVL, Myoepithelial, Natural killer cell, Natural killer (NKT) T cell
	Kidney	[31] retrieved from Kidney Cell Atlas: mature human	7803 * 33 694	7	Natural killer cell, T cell, B cell, CD8+ T cell, Neutrophil, Natural killer T (NKT) cell, Mast cell
	Lung	[17] retrieved from Human Cell Landscape	24 051 * 20 021	11	AT2 cell, Macrophage, Endothelial cell, T cell, Dendritic cell, Mast cell, Muscle cell, B cell, Basal cell, Epithelial cell, Monocyte
	Melanoma	[26] retrieved from GEO: GSE72056	4645 * 23 686	6	T cell, B cell, Macrophage, Endothelial cell, Cancer-associated fibroblast, NK cell
	Pancreas	[32] retrieved from GEO: GSE84133	8569 * 20 125	6	Alpha cell, Acinar cell, Ductal cell, Beta cell, Delta cell, Mesenchymal cell
External testing	Blood	[33] retrieved from 3k PBMCs, 10x Genomics	12 786 * 32 738	5	T cell, CD14++CD16- monocyte, Natural killer cell, Naive B cell, Dendritic cell
	Kidney	[34] retrieved from GEO: GSE169285	22,964 * 25 720	4	Natural killer cell, T cell, B cell, Natural killer T (NKT) cell
	Lung	[35] retrieved from ENA: PRJEB52292	129 340 * 33 538	7	Macrophage, Endothelial cell, T cell, B cell, Dendritic cell, Basal cell, Monocyte
	Melanoma	[36] retrieved from GEO: GSE115978	7186 * 23 686	5	T cell, B cell, Macrophage, Endothelial cell, Cancer-associated fibroblast
	Pancreas	[37] retrieved from GEO: GSE81547	2544 * 15 123	5	Alpha cell, Acinar cell, Ductal cell, Beta cell, Delta cell

dataset contains the expression values of the remaining genes. This operation is necessary because, as we will see in the following sections, our computational model needs two different types of inputs.

GOWDL architecture

Our proposed architecture extends the WDL model [24, 26]. Original WDL is a hybrid network composed of two computational modules. The deep module comprises a deep feed-forward neural network, whereas the wide module is implemented through a generalized linear model. A concatenation operator merges both modules' output weights.

Our version of the WDL, GOWDL, takes as input the expression values of the relevant genes and the expression values of the remaining genes (see previous Section). The former is fed to the generalized linear model to exploit the gene signatures that characterize the available cell types. The latter is fed to the gene ontology-driven CNN (GOCNN) and will be enriched with functional annotations. GOCNN replaces the feed-forward deep neural network of the original WDL and takes advantage of the CNN defined in [38]. In particular, in [38], authors demonstrated that for a similar classification problem, CNN outperformed other traditional machine learning algorithms such as SVM, KNN, Random Forest and FFN. CNN is usually used to extract features that share topological or proximity properties, such as groups of close pixels in an image [39, 40]. In our case, the features represented by the expression values of the genes in a cell do not have any evident proximity or similarity relationships among each other. For this reason, we introduced the KDC layer to define a GOCNN.

The detailed layer architecture of GOWDL is shown in Figure 2. The input of GOCNN is the expression matrix of the non-relevant genes; then, the first layer is the KDC layer responsible for arranging the input matrix according to a similarity metric based on GO annotations. The KDC layer is better explained later on in this section. The following layers comprise three dropout layers, a 1D convolution with kernel size and stride equal to the size k of the kernel dataset, a max pooling, a flatten and a final dense fully connected layer with ReLU activation. Conversely, the wide layer takes the expression matrix of the relevant genes as input, and they are merged with the output of the GOCNN using a concatenate layer. This way, the expression of relevant genes is directly fed into the output layer, preceded by another dropout layer, a dense, fully connected network with a number of neurons equal to the number of cell types to predict, and softmax activation. The network is trained by backpropagation, considering categorical cross-entropy loss and Adam optimizer.

The KDC layer, depicted in Figure 3, can rearrange the set of features of the expression matrix so that each input gene is surrounded by its closest genes according to a functional similarity metric. To do that, we applied the GOGO algorithm [41] that, starting from the GO terms of each gene, computes a functional distance among them, called GO distance. Then, for each input gene, it is possible to obtain a list of its neighbors, sorted by increasing values of the GO distance, which we call a genes dictionary. For a given kernel size k , then, the input expression matrix is rearranged so that each gene is surrounded by its $k-1$ closest genes (bottom part of Figure 3). This way, using the same size k for the CNN kernel and considering a k stride, the CNN kernel is always centered on one gene and its neighbors,

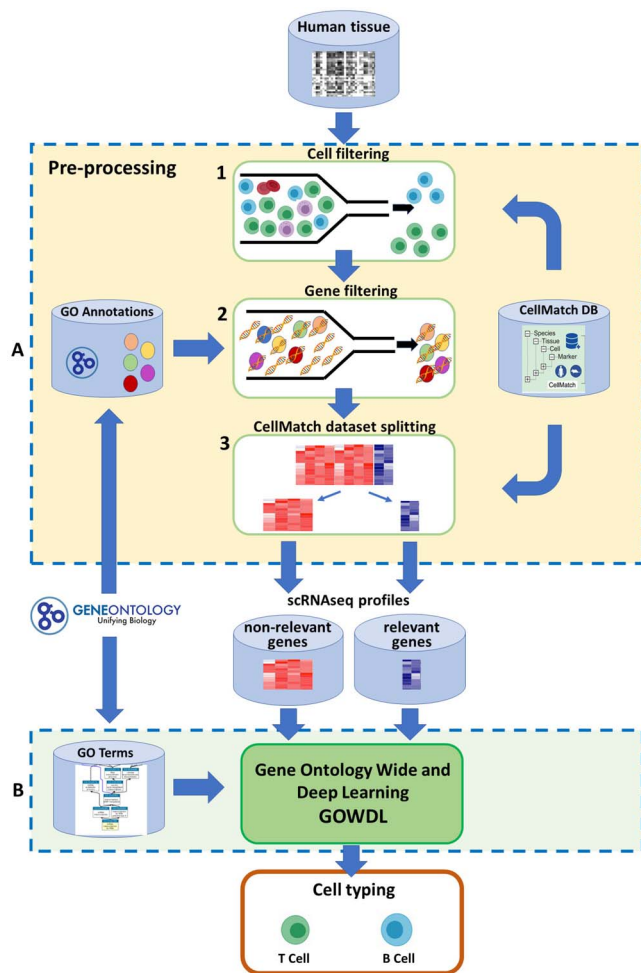


Figure 1. Proposed pipeline. The cell typing pipeline is reported starting from scRNA-seq expression data of a human tissue dataset. The main components are (A) the preprocessing ((1) cell filtering, (2) gene filtering, (3) CellMatch dataset splitting) and (B) the proposed GOWDL model. A: In (1), different cells are filtered according to their abundance and the presence of cell marker genes. The green and blue-colored cells are filtered in this step as they satisfy both cell filtering criteria. In (2), annotated genes are filtered. The oval-colored shapes represent different GO annotations. The DNA symbols represent different genes. The DNA symbols within oval shapes represent annotated genes. In (3), the filtered dataset is split into two new different datasets: the former (red matrix) consists of expression values of genes marked as relevant according to CellMatch db, and the latter (blue matrix) is formed by expression values of not relevant genes. B: After the preprocessing step, the scRNAseq profiles of the two different datasets are processed into the GOWDL model, producing a cell-type classification as the final output.

preserving their functional proximity. GOWDL is implemented in Python 3.9.7 using the Keras framework [42] with Tensorflow 2 backend [43].

RESULTS

Classification results have been computed in terms of several scores, including accuracy, precision, recall, F1 score and Matthew correlation coefficient (MCC) [44] (Eq. 1–5 in [Supplementary material S1](#)). Internal validation has been done through 10-fold cross-validation over the selected training datasets ([Table 1](#)). GOWDL confusion matrices for each training dataset are reported in [Supplementary material S1](#) (Figures S1–S6). In our previous work [27], we showed that the proposed approach can classify

Table 2. Results of 10-fold cross-validation over training datasets.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC (%)
Blood	95.57	95.72	95.57	95.60	92.60
Breast	91.50	91.67	91.50	91.49	87.22
Kidney	92.53	92.77	92.53	92.55	92.54
Lung	95.10	94.89	95.10	94.80	84.09
Melanoma	99.58	99.58	99.58	99.55	98.33
Pancreas	98.40	98.40	98.39	98.39	97.52

cell types better rather than using the GOCNN or the linear layer in a separate way. We also showed how GOWDL outperformed the original WDL model for cell typing [26]. During the cross-validation, we aimed at optimizing some hyper-parameters, such as the number of neurons of the CNN layer, the size of the kernel for the kernel dataset (see [Section 2.4](#)), the number of epochs and the optimizer for minimizing the loss function during the learning phase. We adopted a grid search strategy and reported the best-found parameters in [Supplementary Material S1](#) ([Table S1](#)). Here we report in [Table 2](#) the results, averaged per fold, corresponding to the best parameters obtained during the 10-fold cross-validation procedure of the complete GOWDL pipeline. For each tissue and every score, we got results more excellent than 90%, except for MCC for breast and lung (about 87% and 84%, respectively).

Comparison with other classifiers

We compared our proposed classifier with 10 other state-of-the-art tools for cell classification: SingleR [8], SCINA [9], Scibet [10], scPred [11], scmap [12], CHETAH [13], scClassify [20], Garnett [14], scAnnotatR [21], ACTINN [22]. These classifiers, briefly presented in [Section 1](#), implement different computational models and learning strategies and represent a well-balanced benchmark, as already done in the work of [21]. Moreover, we compared GOWDL with two general-purpose tree-based classifiers, namely XGBoost [45] and CatBoost [46]. All classifiers were trained and tested with the same train and external test datasets, summarized in [Table 1](#), including pancreas, lung, kidney, melanoma and blood tissues. Concerning the test datasets, we considered only the cells belonging to the same cell types of the training sets ([Table T2](#) of [Supplementary material S2](#)). For each tissue, we trained the GOWDL model with the best hyper-parameters given by cross-validation over the whole training dataset. We used default parameters for the other classifiers. In order to run all the experiments, we adapted the R scripts developed in the work of [21]. Tables with complete results of 12 classification algorithms over the external datasets, tissue by tissue, are reported in [Supplementary Materials S2](#) (Tables T3–T7). Here, we displayed results as a series of boxplots ([Figure 4](#)). We produced a chart for each prediction score, namely accuracy, precision, recall, F1 score and MCC; each boxplot represents the average results over the five tissues for a given tool. From the chart, it is noteworthy that our predictor provides very stable results regardless of the tissues. The area of the corresponding boxplot is always narrow for every score. That means there is a low variance concerning the tissue because the lines representing the first quartile (upper line), median (middle line) and third quartile (bottom line) are close. Other tools, such as scClassify, although they reach some very high scores, like MCC, have a wider area. Globally, GOWDL and Scibet provide the best performances, with our predictor reaching

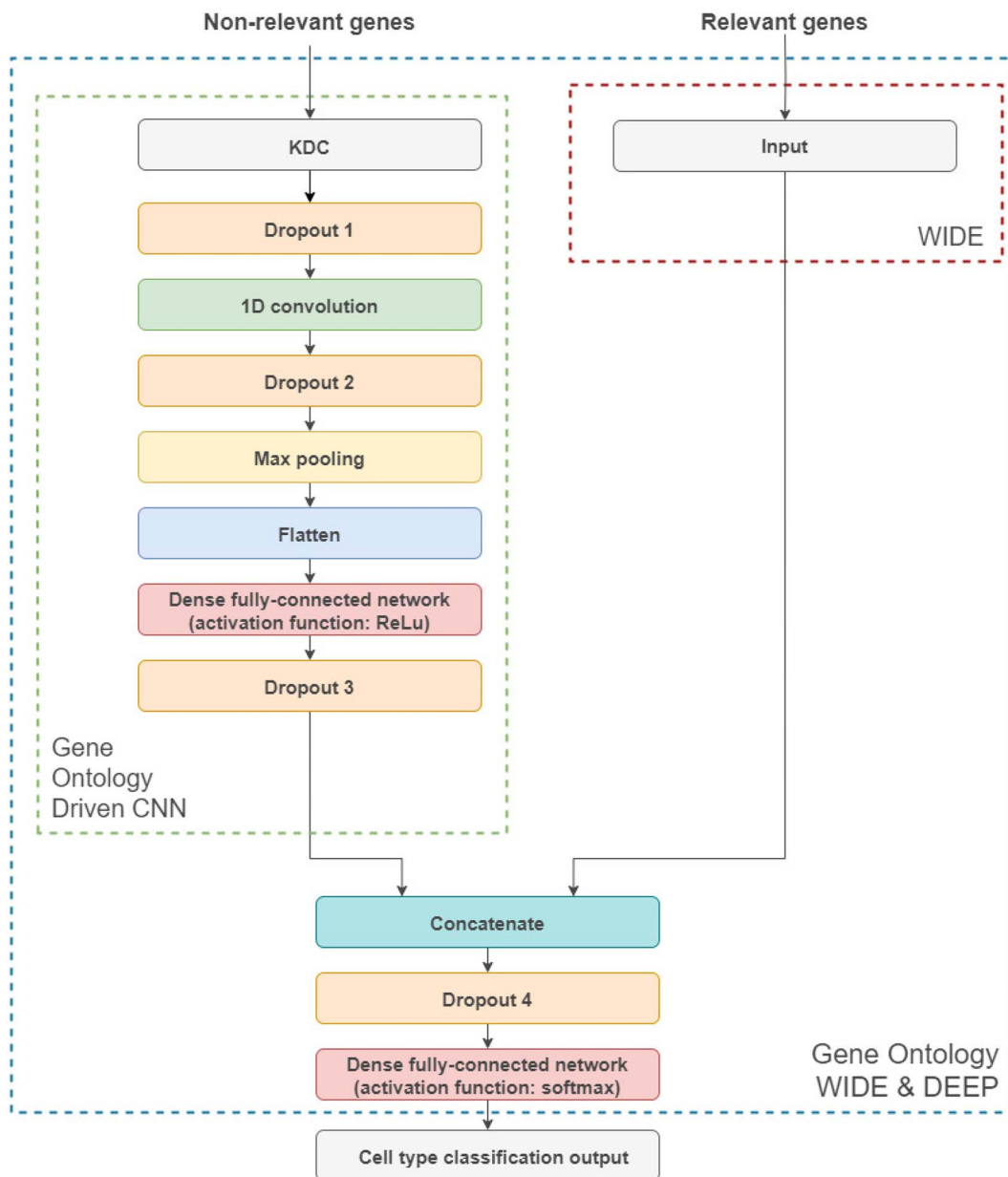


Figure 2. GOWDL model architecture. It takes as input two RNA-seq expression data matrices. The green dotted box shows the gene ontology-driven CNN (GOCNN), composed of a kernel dataset creation (KDC) layer and a convolutional neural network.

higher median values, except for the recall score (Figure 4(C)). Generally speaking, improving precision without hurting recall in multiclass classification problems with unbalanced datasets is challenging since, to increase the false positive for the minority class, the true positives are also often increased, resulting in a decreased recall. For this reason, we used the F1 score measure, which provides a way to combine both measures into a single score that incorporates the properties of precision and recall. 4(D) clearly shows that the proposed algorithm performs better than the others regarding the F1 score.

In Figure 5, we presented a detailed bar plot that shows the classification scores for every predictor, considering each tissue separately. From the graph, we can see that GOWDL, Scibet and ACTINN reach overall the best scores, ranging from about 80% for kidney to 95% for blood. In contrast, other tools, like SingleR and scClassify, have achieved interesting results for some tissues

(e.g. melanoma and pancreas). Still, if we look at the boxplot (Figure 4), they are very tissue-dependent. From this point of view, experimental results demonstrated cell types of lung tissue are the most difficult to predict when dealing with external data, with GOWDL reaching about 75% for accuracy and recall and about 70% for precision, F1 score and MCC. However, GOWDL is less affected by this performance drop concerning the other predictors considering all the classification scores. Also, to test the advantage introduced by a proximity measure among genes, we compared the GOWDL algorithm with a modified version without the gene ontology-driven component. Results in Supplementary materials S2 (Tables T8) show how GOWDL performs better in accuracy with all the tissues and almost always reaches the best results in terms of F1-score and MCC. Figure 6 shows ROC curves and the corresponding AUC scores for five tissues. We calculate them for the proposed algorithm and compare it with

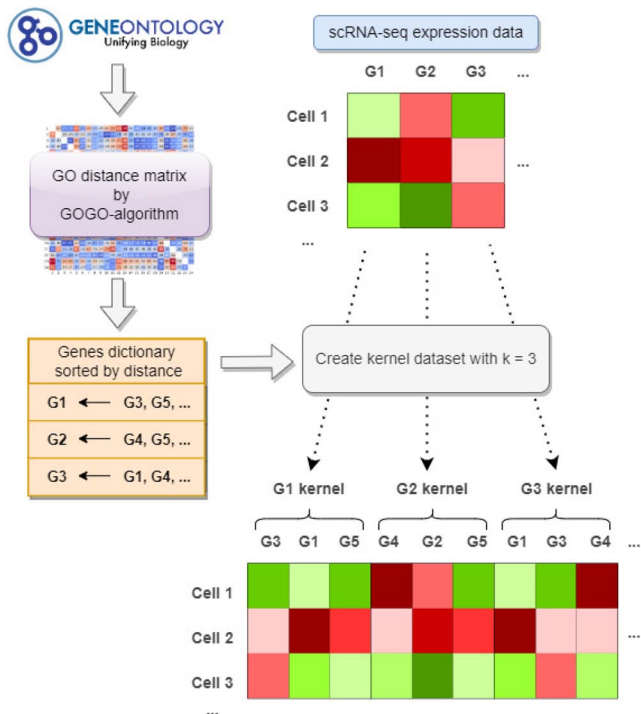


Figure 3. KDC schema. Here it is reported the creation of a gene dictionary sorted by distance (on the left) that is used to transform the original input data in a kernel dataset (on the right).

the other five algorithms that provide a probability value for their prediction, i.e. SingleR, SCINA, scPred, scAnnotatR, ACTINN, XGBoost and CatBoost. GOWDL can reach a higher value of AUC for each tissue, but the kidney, where ACTINN performs slightly better (0.98 versus 0.97). Considering the AUC average value over all tissues, the GOWDL algorithm results best.

Case study on immune cells population: a hierarchical approach on T cell classification

Immune cells are fundamental defenders of innate and adaptive immunity, protecting the host against endogenous and exogenous agents, including malignant cells. Immunosurveillance mechanisms are essential both in normal and cancer tissues. Different cell sub-types of the immune system include T cells, B cells, dendritic cells, monocytes and macrophages. They share distinct duties in regulating innate and adaptive immune functions [47]. Studying these distinct cell classes is essential to understand the molecular mechanisms of the immune system in normal tissue and cancer disease [48]. Indeed the immune cell profiles of different tissue types, such as normal and cancer tissue, can indicate a status of immune activation in the tumor, which includes both pro- and anti-inflammatory features [49]. Moreover, among the different immune cell populations, the T cells are considered a parent cell class, including other sub-classes, for instance, CD4+ T cells, Cd8+, memory T and Regulative T cells [30]. Their composition in terms of quantity and cell sub-types characterizes a specific immune profile, reflecting the heterogeneity of the disease, from morphology to molecular alterations, with specific genotype-phenotype correlations [50]. Therefore, a deeper study on breast cancer T cell type classification could help estimate patients' prognoses and outcomes in breast cancer to design targeted therapies.

Considering the above, it seems interesting to exploit the proposed GOWDL algorithm in a different configuration that could emphasize T cell types more. We introduce a hierarchical version of the GOWDL algorithm called Hier-GOWDL to reach this goal. This approach represents a hierarchical problem as a set of independent flat classification problems by performing a two-level analysis. Hier-GOWDL considers a first level that predicts all the ancestor labels and a second level that classifies all the corresponding descendant labels. In classifying immune cell populations, the two levels of Hier-GOWDL are trained on generic T cells as ancestor labels and T cell sub-types as descendant labels, respectively. We can perform this analysis because among the 16 cell types contained in the Breast tissue dataset (Table T1 in Supplementary materials S2), we used to train the GOWDL model, there are six T cell sub-types, i.e. CD8+ T, T helper, T Reg, T cell Cycling, Natural killer T and 'other T' cells, that are undefined T cells sub-types. This way, the Hier-GOWDL model of the first level is trained with 11 cell types (we add a generic 'T cell' type that aggregates all the T cells to the other 10 different cell types), and the second one is trained with the six T cell sub-types. The 'Flat' GOWDL model is the same one learned with 10-fold cross-validation (see Section 3) using the full 16-class breast dataset. To compute the classification results, we considered all the T cells as a unique class for comparison with level 1 of Hier-GOWDL; we then considered only the six T cells sub-types for comparison with level 2 of Hier-GOWDL.

To compare the Hier-GOWDL approach to the 'flat' GOWDL model, we produced the confusion matrices shown in Figure 7. As for the flat GOWDL model, we started from the confusion matrix obtained during the 10-fold cross-validation process, and then we aggregated as a unique T cell class the six T cell sub-types (Figure 7(A)) to perform the comparison with the confusion matrix obtained at the first level of the Hier-GOWDL (Figure 7(B)). On the other hand, we extracted the sub-matrix with only the six T cell sub-types (Figure 7(C)) to allow the comparison with the confusion matrix obtained at the second level of the Hier-GOWDL (Figure 7(D)). With regards to the first two confusion matrices (Figure 7(A) and (B)), we can notice how they are very similar, with Hier-GOWDL capturing more true T cells, even if some other cells are misclassified, such as Epithelial cells. It is noteworthy, in turn, the behavior of the models for the classification of the six T cell sub-types (Figure 7(C) and (D)). Hier-GOWDL can capture more true cells in almost each T cell type, including T Reg, T helper, T cell cycling and generic T cell classes. That could confirm the hypothesis that a hierarchical structure for learning specific sub-classes provides better classification results. This performance improvement was evident when we computed the classification scores of both models.

To estimate the performances of the hierarchical multilabel classification method, we used the average accuracy and three measures defined in [51] i.e. the hierarchical precision, hierarchical recall and hierarchical f-score (Eq. 6–8 in Supplementary material S1). Table 3 reports the comparison results. The first two rows report the higher level (11 classes), whereas the last two refer to the lower level (six T cell sub-type classes). The Hier-GOWDL approach performs slightly better in both cases than the proposed GOWDL model: more in detail, at the first level, we have a few advantages in using the hierarchical approach, with a deviation lower than 0.2 percent in terms of hierarchical F1-score, whereas, at the second level, the gain is more significant, with a deviation greater than 1 percent in terms of hierarchical F1-score. As expected, we recorded the main advantages at level 2 because the model has to distinguish among

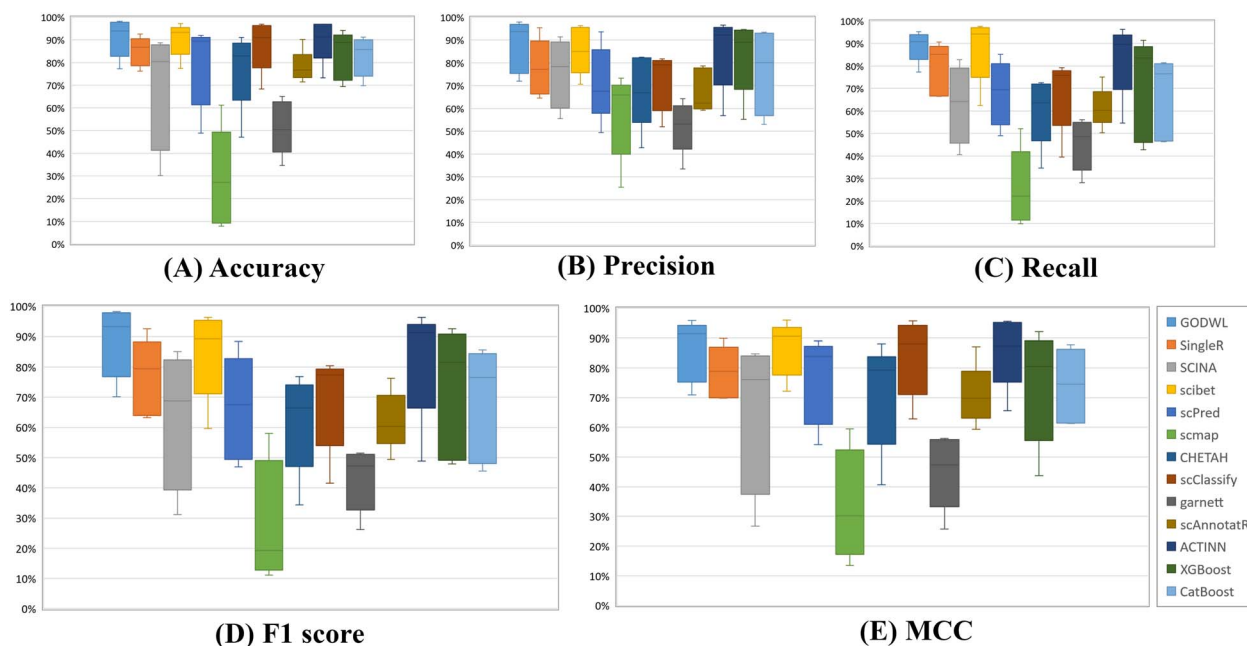


Figure 4. Comparison among GOWDL and other 12 cell typing algorithms, regarding five performance measures. Boxplots are calculated over five datasets belonging to different tissues.

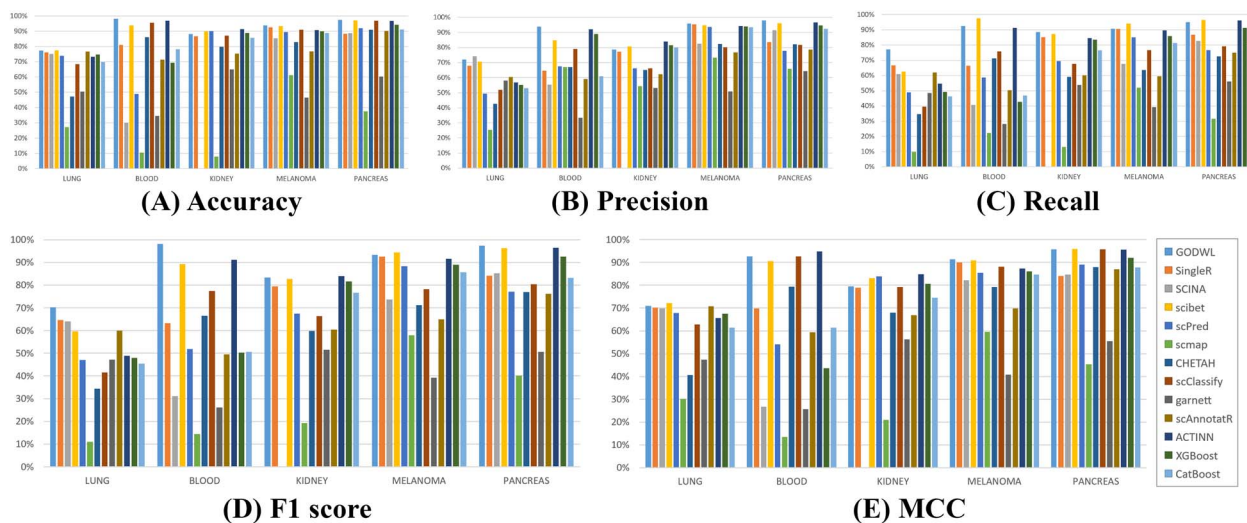


Figure 5. Comparison among GOWDL and other 12 cell typing algorithms, over five different tissues.

six classes instead of considering all the 16 classes as done by the ‘flat’ GOWDL algorithm. In general, this case study shows that, in the future, the discovery of new gene markers related to different subtypes could improve the proposed flat algorithm’s performance, which can be easily adapted in a hierarchical model to investigate the classification of cell types and subtypes deeper.

CONCLUSION

We presented GOWDL, a gene ontology-driven WDL architecture for classifying cell types in scRNA-seq data. As demonstrated in this work, the proposed algorithm exploits the advantages of information extracted from the Gene Ontology resource and from the CellMatch database to predict cell types in different tissues. Based on the WDL model, the proposed architecture allowed us to

combine this information and perform equal to or better than the other 12 state-of-the-art algorithms on five independent scRNA-seq datasets belonging to five different human tissues. Also, looking at the near future, accordingly to the prevision of increasing availability of scRNA-seq data and marker genes, we introduced a case study to extend GOWDL at the cell sub-types level with a hierarchical version of the proposed algorithm. As discussed in this work, we believe the proposed algorithm can be considered a valuable contribution to the cell type classification task, as it performs equal to or better than the other 12 state-of-the-art algorithms in cell type classification. Indeed our model can be integrated into more complex bioinformatics pipelines developed for specific biomedical tasks, which are addressed to understand the behavior and the variation of distinct phenotypes associated with different physiological and pathological conditions, such as cancer pathology.

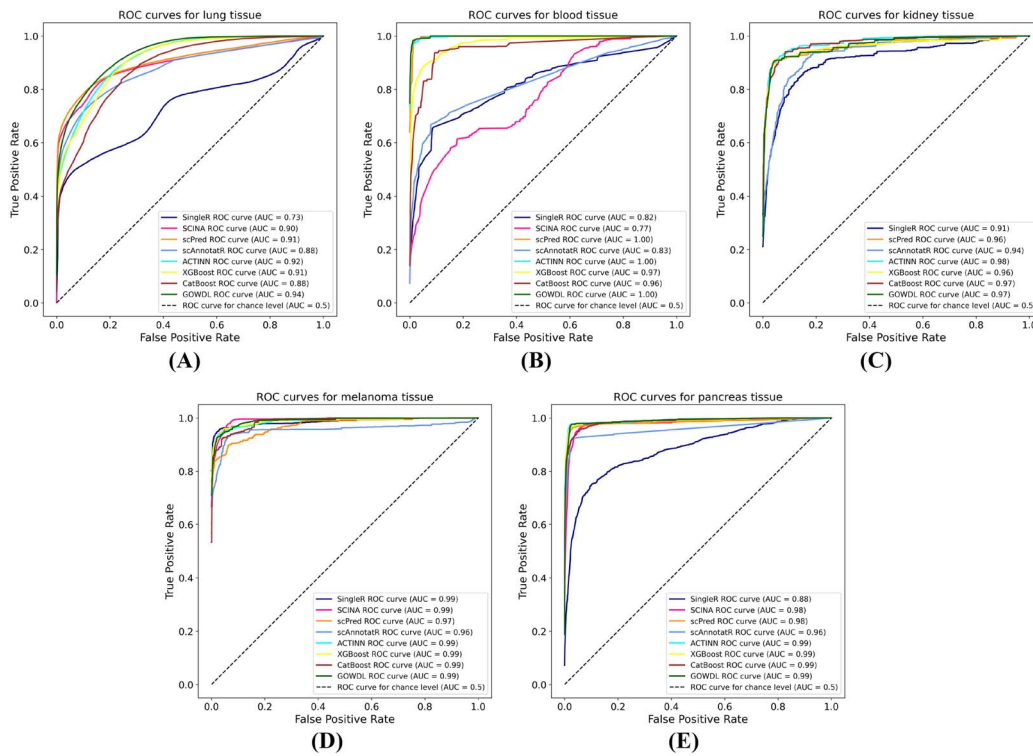


Figure 6. ROC curves and the corresponding AUC scores for five tissues. We compared the proposed algorithm with the other five algorithms that provide a probability value for their prediction, i.e. SingleR, SCINA, scPred, scAnnotatR and ACTINN.

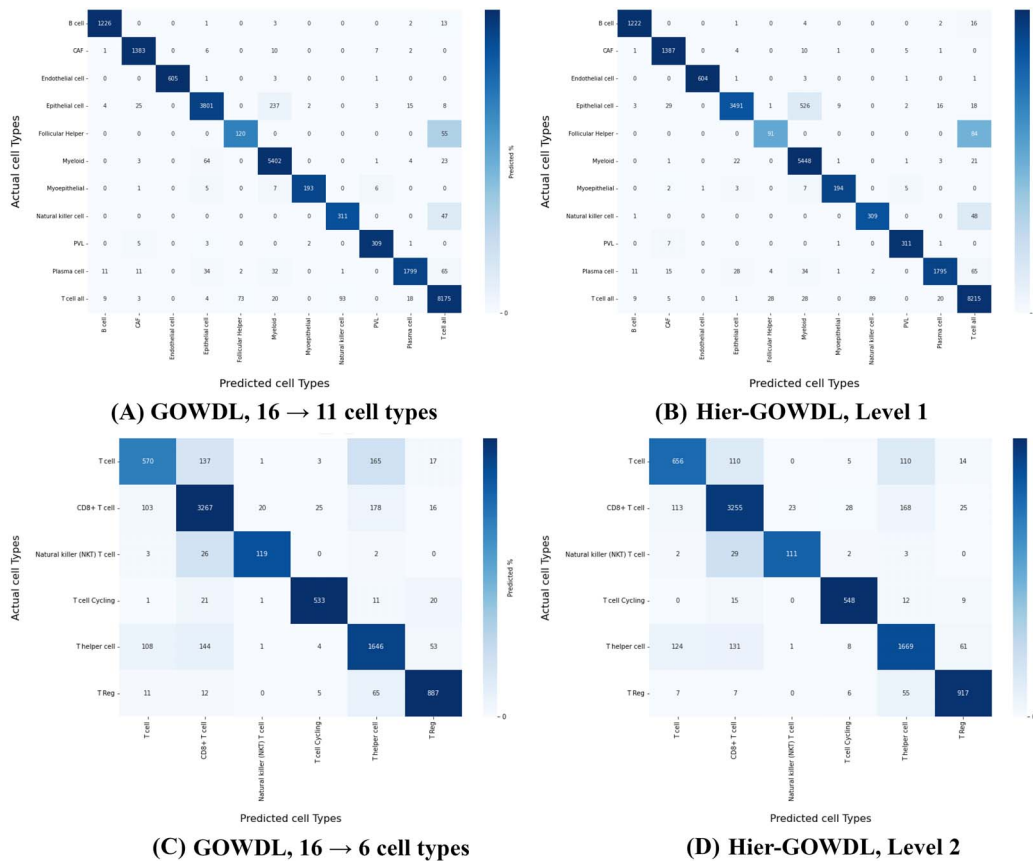


Figure 7. Confusion matrices for the immune cells population case study. At the top, there is (A) the rearranged version of the original confusion matrix of the GOWDL algorithm trained with the Breast dataset, where all T cells are considered as a unique 'T cell all' class, against (B) the confusion matrix of the first level of the Hier-GOWDL algorithm. At the bottom is (C) the submatrix containing only the T cells sub-types of the original confusion matrix of the GOWDL algorithm trained with the Breast dataset, against (D) the confusion matrix of the second level of the Hier-GOWDL algorithm.

Table 3. Comparison between the standard and the hierarchical version of the GOWDL algorithm for Breast dataset. 'Acc', 'hP', 'hR' and 'hF' stand for T cells accuracy, hierarchical precision, hierarchical recall and hierarchical F1-score. To perform the comparison with the first and second level of Hier-GOWDL, we got results of GOWDL on the Breast dataset. Then we aggregated 16 cell types to 11 (with all T classes together) and extracted six cell types from 16 (only T cell sub-types), respectively.

Algorithm	Cell types	Level	Acc (%)	hP (%)	hR (%)	hF (%)
GOWDL	16 → 11	-	98.22	97.48	97.38	97.43
Hier-GOWDL	11	1	98.30	97.24	97.86	97.55
GOWDL	16 → 6	-	85.90	81.54	80.23	80.87
Hier-GOWDL	6	2	87.01	82.08	81.99	82.04

Key Points

- We present GOWDL, a deep learning model for cell-type prediction using scRNA-seq data.
- GOWDL implements a hybrid computational model that considers both gene functional annotations, using an ontology-driven CNN, and the gene signature that characterizes the cell types, using a generalized linear model.
- We compared the performances of GOWDL with 12 state-of-the-art classifiers, considering five independent test sets belonging to different tissues and including several cell types. GOWDL outperformed the other classifier and obtained the most stable results across each test set.
- We present a case study on immune cell populations where we propose a two-level hierarchical configuration of GOWDL, namely Hier-GOWDL, to classify closely related T cells. Hier-GOWDL reached slightly better results than the flat configuration, especially at the second classification level, which deals with the identification of close sub-types of T cells.

SUPPLEMENTARY DATA

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

AUTHOR CONTRIBUTIONS STATEMENT

A.F.: project conception, pipeline design and implementation, discussion, assessment, and writing. M.L.R.: project conception, design of experiments, implementation, discussion, assessment, and writing. L.L.P.: project conception, biological expertise, discussion, assessment, and writing. S.G.: discussion, assessment, and writing. A.U.: discussion, assessment, writing, and funding. All authors read and approved the final manuscript.

FUNDING

This work was funded by the National Research Council of Italy [DBA.AD005.225 – NUTRAGE Project].

DATA AVAILABILITY

Preprocessed data and source code of GOWDL are available at <https://github.com/BCB4PM/GOWDL>.

Original training and test datasets are available on their proper public repositories, as shown in Table 1.

REFERENCES

1. Tang F, Barbacioru C, Wang Y, et al. Mrna-seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;**6**(5): 377–82.
2. Nguyen A, Khoo WH, Moran I, et al. Single cell rna sequencing of rare immune cell populations. *Front Immunol* 2018;**9**:1553.
3. Stewart A, Ng JC-F, Wallis G, et al. Single-cell transcriptomic analyses define distinct peripheral b cell subsets and discrete development pathways. *Front Immunol* 2021;**12**:602539.
4. Ding J, Smith SL, Orozco G, et al. Characterisation of cd4+ t-cell subtypes using single cell rna sequencing and the impact of cell number and sequencing depth. *Sci Rep* 2020;**10**(1):1–11.
5. Abdelaal T, Michielsen L, Cats D, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol* 2019;**20**(1):194.
6. Javier Diaz-Mejia J, Meng EC, Pico AR, et al. Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Research* 2019;**8**:296.
7. Zhao X, Shuang W, Fang N, et al. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets. *Brief Bioinform* 2020;**21**(5):1581–95.
8. Aran D, Looney AP, Liu L, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**(2):163–72.
9. Zhang L, Luo D, Zhong X, et al. SCINA: semi-supervised analysis of single cells in Silico. *Genes* 2019;**10**(7):531.
10. Li C, Liu B, Kang B, et al. SciBet as a portable and fast single cell type identifier. *Nat Commun* 2020;**11**(1):1818.
11. Alquicira-Hernandez J, Sathe A, Ji HP, et al. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol* 2019;**20**(1):264.
12. Kiselev VY, Yiu A, Hemberg M. Scmap: projection of single-cell RNA-seq data across data sets. *Nat Methods* 2018;**15**(5):359–62.
13. Kanter de JK, Lijnzaad P, Candelli T, et al. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res* 2019;**47**(16):e95–5.
14. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* 2019;**16**(10): 983–6.
15. Boufea K, Seth S, Batada NN. Scid uses discriminant analysis to identify transcriptionally equivalent cell types across single-cell rna-seq data with batch effect. *iScience* 2020;**23**(3):100914.
16. Cao Y, Wang X, Peng G. Scsa: a cell type annotation tool for single-cell rna-seq data. *Front Genet* 2020;**11**:490–0.
17. Shao X, Yang H, Zhuang X, et al. Scdeepsort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res* 2021;**49**(21):e122–2.
18. Wang T, Bai J, Nabavi S. Single-cell classification using graph convolutional networks. *BMC Bioinformatics* 2021;**22**(1):364–4.
19. Shao X, Liao J, Xiaoyan L, et al. Scatch: automatic annotation on cell types of clusters from single-cell rna sequencing data. *IScience* 2020;**23**(3):100882.
20. Lin Y, Cao Y, Kim HJ, et al. scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. *Mol Syst Biol* 2020;**16**(6).
21. Nguyen V, Griss J. scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data. *BMC Bioinformatics* 2022;**23**(1):44.

22. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. *Bioinformatics* 2020;**36**(2): 533–8.
23. Carbon S, Douglass E, Good BM, et al. The gene ontology resource: enriching a Gold mine. *Nucleic Acids Res* 2021;**49**(D1): D325–34.
24. Cheng H-T, Koc L, Harmsen J, et al. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, pages 7–10, New York, NY, USA, 2016. ACM.
25. Albawi S, Mohammed TA, and Al-Zawi S. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. IEEE, 2017.
26. Wilson CM, Fridley BL, Conejo-Garcia JR, et al. Wide and deep learning for automatic cell type identification. *Comput Struct Biotechnol J* 2021;**19**:1052–62.
27. Coppola G, Fiannaca A, La Rosa M, et al. A Gene Ontology-Driven Wide and Deep Learning Architecture for Cell-Type Classification from Single-Cell RNA-seq Data. In: *Engineering Applications of Neural Networks*. Springer, 2022, 323–35.
28. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**(13):3573–87.
29. *Single Cell Portal. Study: ICA: Blood Mononuclear Cells*, 2022, accessed: 12-2022.
30. Wu SZ, Roden DL, Wang C, et al. Stromal cell diversity associated with immune evasion in human triple-negative breast cancer. *EMBO J* 2020;**39**(19):e104063.
31. Stewart BJ, Ferdinand JR, Young MD, et al. Spatiotemporal immune zonation of the human kidney. *Science* 2019;**365**(6460): 1461–6.
32. Baron M, Veres A, Wolock SL, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst* 2016;**3**(4): 346–360.e4.
33. 10x Genomics. *3k PBMCs from a Healthy Donor, Single Cell Gene Expression Dataset by Cell Ranger 1.1.0*, 2016, accessed: 11-2022.
34. Menon R, Otto EA, Berthier CC, et al. Glomerular endothelial cell-podocyte stresses and crosstalk in structurally normal kidney transplants. *Kidney Int* 2022;**101**(4):779–92.
35. Madissoon E, Oliver AJ, Kleshchevnikov V, et al. A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nat Genet* 2023;**55**(1):66–77.
36. Jerby-Arnon L, Shah P, Cuoco MS, et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell* 2018;**175**(4):984–997.e24.
37. Martin Enge H, Arda E, Mignardi M, et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* 2017;**171**(2):321–330.e14.
38. Canakoglu A, Nanni L, Sokolovsky A, Ceri S. Designing and Evaluating Deep Learning Models for Cancer Detection on Gene Expression Data. In: *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer, 2020, 249–61.
39. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018;**9**(4):611–29.
40. Raitoharju J. Convolutional neural networks. In: *Deep Learning for Robot Perception and Cognition*. Elsevier, 2022, 35–69.
41. Zhao C, Wang Z. GOGO: an improved algorithm to measure the semantic similarity between gene ontology terms. *Sci Rep* 2018;**8**(1):15107.
42. Chollet F, et al. *Keras*. <https://keras.io>, 2015.
43. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z. *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. Software available from tensorflow.org.
44. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Mining* 2017;**10**(1):35.
45. Chen T and Guestrin C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–94, New York, NY, USA, 2016. ACM.
46. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, and Gulin A. Catboost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6639–49, Red Hook, NY, USA, 2018. Curran Associates Inc.
47. Degnim AC, Brahmbhatt RD, Radisky DC, et al. Immune cell quantitation in normal breast tissue lobules with and without lobulitis. *Breast Cancer Res Treat* 2014;**144**:539–49.
48. Goff SL, Danforth DN. The role of immune cells in breast tissue and immunotherapy for the treatment of breast cancer. *Clin Breast Cancer* 2021;**21**(1):e63–73.
49. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature* 2000;**406**(6797):747–52.
50. Annaratone L, Cascardi E, Vissio E, et al. The multifaceted nature of tumor microenvironment in breast carcinomas. *Pathobiology* 2020;**87**(2):125–42.
51. Cerri R, Pappa GL, Carvalho ACPLF, Freitas AA. An extensive evaluation of decision tree-based hierarchical multilabel classification methods and performance measures. *Comput Intell* 2015;**31**(1):1–46.