IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received December 30, 2020, accepted January 12, 2021, date of publication January 19, 2021, date of current version January 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3052923

Cell Subtype Classification via Representation Learning Based on a Denoising Autoencoder for Single-Cell RNA Sequencing

JOUNGMIN CHOI[1], JE-KEUN RHEE[2], AND HEEJOON CHAE[1]
[1]Division of Computer Science, Sookmyung Women's University, Seoul 04310, Republic of Korea
[2]School of Systems Biomedical Science, Soongsil University, Seoul 06978, Republic of Korea

Corresponding author: Heejoon Chae (heechae@sookmyung.ac.kr)

This work was supported in part by the Bio and Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT), Korean Government, under Grant 2019M3E5D3073365, in part by the Agenda Project of the Rural Development Administration, Republic of Korea, under Grant PJ0143072019, and in part by the National Research Foundation of Korea (NRF) funded by MSIT, Republic of Korea under Grant NRF-2018R1C1B6005304.

ABSTRACT Identification of single-cell subtypes is one of the fundamental processes required to understand a heterogeneous population composed of multiple cells, based on single-cell RNA sequencing data. Previously, cell subtype identification was mainly carried out by dimension reduction and clustering approaches that grouped cells with similar expressed profiles together. However, for high robustness to noises and systematic annotation of the subtype in each cell, supervised classification approaches have been widely used. Recently, deep neural network (DNN) models have been widely presented in various fields, including biology. By capturing the composite relationship between sample features and target outcomes, a DNN model enables significant performance improvements in biological data mining analyses. In this paper, we constructed a DNN model, called scDAE for single-cell subtype identification combined with representative feature extraction using a multilayer denoising autoencoder (DAE). The feature sets were learned by the DAE and were further tuned by fully connected layers using a softmax classifier. The model was compared against four state-of-the-art cell subtype identification methods and two conventional machine learning algorithms. From multiple tests, scDAE significantly outperformed competing methods especially on data sets having a large number of cell subtypes and noises. Extracted cell features from the proposed model were clearly clustered with respect to subtype. The results of the experiments indicated that our proposed model is effective in identifying single-cell subtypes and molecular signatures representative of each cell subtype. scDAE is publicly available at https://github.com/cbi-bioinfo/scDAE.

INDEX TERMS Cell subtype, classification, gene expression, scRNA-seq, single-cell.

I. INTRODUCTION

Gene expression profiling technologies such as microarray and RNA sequencing have allowed the investigation of the gene expression levels of tens of thousands of genes simultaneously. By measuring the transcriptome levels of genes, we can identify the differentially expressed genes in a specific disease, search enriching gene sets in a biological group, and construct gene regulatory networks. However, the conventional methods detect gene expression levels using bulk cells, that is, it is impossible to explore gene expression profiles at the single-cell level [1]. Individual cells are composed of heterogeneous subtypes in a given tissue, and gene expression levels show variations, even within the same cell subtype. Thus, the precise expression profiling of individual cells and the accurate annotation of cell subtypes is essential to elucidate the understanding of biological systems.

In recent years, the development of single-cell RNA-sequencing (scRNA-seq) is leading to facilitate a novel in-depth biological founding. For example, scRNA-seq helps to understand cell lineages and pathogenesis [2]–[4]. In particular, within the aspect of cancer genomics, scRNA-seqs have been widely used to resolve tumor evolution processes, to segregate primary and metastatic tumors, to investigate

The associate editor coordinating the review of this manuscript and approving it for publication was Carmen C. Y. Poon.

tumor immune infiltration, to develop clinical application strategies, and so on [5]–[9]. To date, cell subtype composition has been estimated by several deconvolution methods from bulk gene expression profiles [10]–[12], but the advent of scRNA-seq not only directly detects cell composition but also determines gene expression levels in each cell subtype.

The subgrouping of the cells based on scRNA-seq data has been mainly carried out by unsupervised learning, such as principal component analysis (PCA) or other clustering approaches. For instance, RaceID proposed resolving different cell subtypes in a complex mixture based on identified cell clusters by k-means clustering [13], and SNN-Clip presented single-cell transcriptomes clustering with a shared nearest neighbor graph construction [14]. To uncover multiple layers of biological populations in scRNA-seq datasets, DendroSplit developed an interpretable clustering framework based on a separation score using feature selection [15], while SIMLR [16] and MPSSC [17] employed multiple kernel learning and spectral clustering to learn cell-to-cell similarities among heterogeneous populations of samples, respectively. RAFSIL implemented a random forest model in an unsupervised way to apply similarity learning for exploratory analysis of cell subtype discovery [18], and SinNLRR was developed as a scRNA-seq cell subtype detection method, which identified non-negative and low-rank representations of gene expression matrix from all candidate subspaces [19]. However, despite the advantage of not requiring specific cell subtype information but using the expression pattern of marker genes, most of the unsupervised cell subtype identification suffer when the subtype-specific marker genes are poorly selected due to lack of prior knowledge [20].

Recently, several supervised approaches were developed for the characterization of individual cell subtypes. Scmap selected top N residuals as informative features from a fitted linear model capturing the relationship between gene expression value and dropout rate of existing reference dataset [21]. Based on these features, to predict a cell subtype, projection of a new cell was performed to identify the most similar cell subtype. CaSTLe performed feature selection based on the highest mean expression and mutual information between the genes and cell subtypes [22]. Then, a classification model based on XGBoost was constructed, improved by transfer learning, and classified the target data. ScPred decomposed a gene expression matrix using a singular value decomposition method to identify important features and a support vector machine (SVM) model that was trained as a classification model [23]. CHETAH created a classification tree model based on reference profiles having average gene expression values for each cell subtype [24]. By traversing the tree, the input cell is classified based on the similarity with each node. Garnett [25] and CellAssign [26] performed cell subtype assignment based on the user-specified marker gene set for each cell subtype with raw scRNA-seq read counts, where Garnett trained an elastic-net regression-based classifier and CellAssign identified cell subtypes by computing a probabilistic assignment for each cell to a cell subtype. Although the proposed methods have shown reasonable performance based on the conventional machine learning and statistical models, they are limited in processing raw data that requires careful engineering procedures to transform the raw dataset into a suitable representation form that machine learning systems can understand [27], [28] or require an additional process to select markers for each cell subtype manually.

To solve this issue, deep neural network (DNN) models have been presented. These models can automatically learn informative features of the input within the latent space, where each layer of DNN captures patterns of the raw input data in different perspectives by optimizing the objective function [28]. Several studies employed a DNN model due to the ability of the latent feature extraction. For example, ADAGE and an ensemble ADAGE integrated diverse gene expression data and predicted the involvement of biological processes based on low-level gene expression differences without requiring prior knowledge [29], [30]. A DNN approach using a variational autoencoder (VAE) trained on pan-cancer RNA seq data identified specific patterns of gene expression data and profiled a biologically relevant latent space [31], [32]. For studies using scRNA-seq data, Lin, *et al.* employed DNN models combining prior biological information to reduce the dimensions of expression values, and evaluated performance by comparing their model to prior clustering and dimensionality reduction methods [33]. To perform fundamental analysis for single-cell transcriptome dataset, the single-cell variational inference (scVI) model was introduced for probabilistic representation [34] and ACTINN implemented a DNN model to assign each cell a cell type [35]. To improve the problem of clustering scRNA-seq data caused by low RNA capture rate, scDeepCluster integrated a zero-inflated negative binomial model with a clustering loss function to optimize clustering explicitly [36]. However, identification of individual cell subtypes employing deep neural networks based on scRNA-seq data still needs improvement.

In this paper, we present a simple, but significantly more stable DNN-based cell subtype classification model utilizing the scRNA-seq dataset, named scDAE. Due to the high level of noise introduced by technical biases that vary across cells, such as amplification bias and library size differences, and dropout events caused by low RNA capture rate, misclassification of cell subtypes can occur. This could affect downstream analysis significantly and lead to the false interpretation of the results. The impact from these noises becomes more significant as the number of cell subtypes increases. Recently, several approaches to characterize the individual cell subtypes have been developed, and analysis for single-cell transcriptome dataset composed of multiple cell subtypes has been performed. But still, these studies suffer from the high level of noises due to the increased number of subtypes. To prevent this issue, a cell subtype prediction model robust to noise is needed.

We implemented a denoising autoencoder (DAE) to transform the high-dimensional scRNA-seq data into

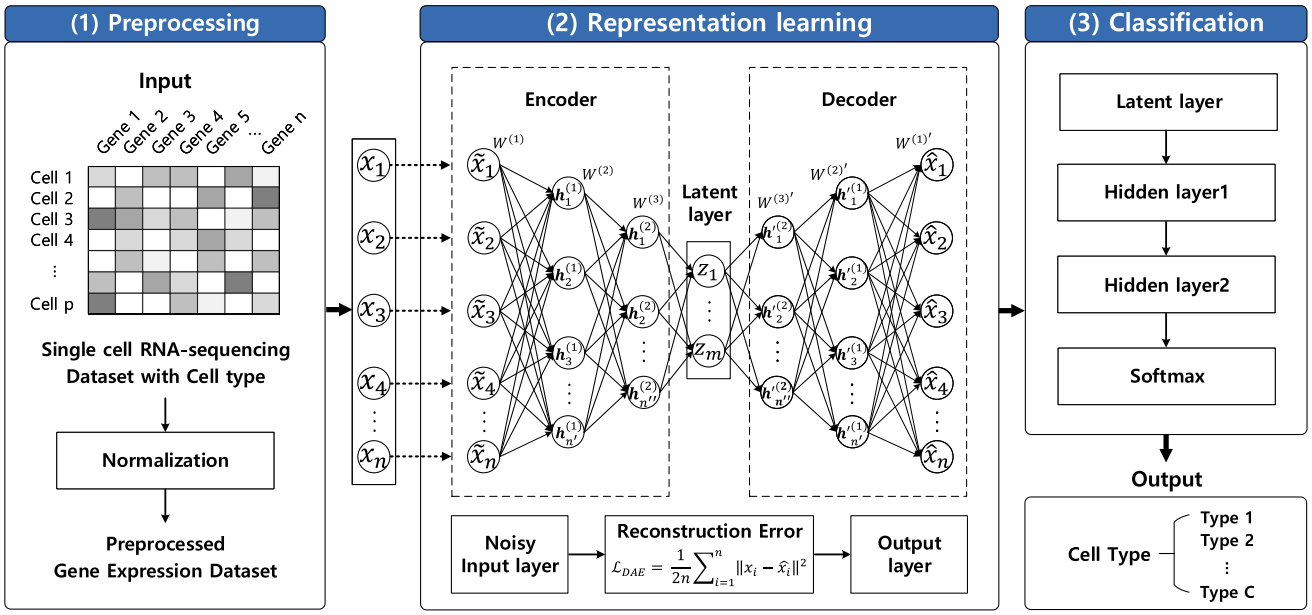

FIGURE 1. Illustration of the proposed cell subtype classification model based on a DNN using scRNA-seq data. The classification procedure consists of three main phases: (1) preprocessing to normalize the data for accurate prediction, (2) representation learning through multi-layered DAE to extract latent features and (3) classification of cell subtypes by FC neural network with softmax layer based on extracted latent features.

low-dimensional data to extract informative representations and fully connected (FC) neural network models with softmax layers for cell subtype classification. We evaluated the performance of our proposed method with the state-of-the-art classifier models. Our comparison results indicate that the proposed model provided the highest classification performance and successfully extracted latent features related to the multiple cell subtypes.

II. METHODS

In this section, we introduce multiple steps to extract latent features from scRNA-seq datasets and describe the details of the model structure to predict cell subtypes. A flowchart is shown in Fig 1.

A. PREPROCESSING

Recently, it was highlighted that scRNA-seq datasets typically reflect biological heterogeneity and technical biases [37]. To eliminate the effects originating from these issues, raw scRNA-seq read counts were pre-processed using R package DESeq2 [38]. First, we removed genes with no count in any cell. Second, size factors were calculated, and read counts were normalized by library size. For the last step, read counts were log-transformed.

B. REPRESENTATION LEARNING

To extract informative signatures from the genes in a scRNA-seq dataset, which are robust to the high level of noises caused by the technical biases from current scRNA-seq protocols, we designed and implemented multi-layered DAEs. The DAE is an addition of a noising layer to a regular autoencoder (AE), where the AE is a symmetrical neural network learning a compact representation of the input data by reconstructing output as close as possible to the input. As the number of hidden nodes is smaller than the number of input nodes, the latent features of the input data can be extracted by minimizing reconstruction errors [39]. The AE consists of encoding and decoding layers, where encoding layers perform a deterministic mapping with a non-linear activation function, transforming input node x of raw data into a latent representation in an unsupervised way. Conversely, decoding layers try to reconstruct the original input from the extracted features by minimizing reconstruction errors.

Given a set of original input data $x \in R^n$, where n is the dimension of data, an encoder tries to convert x to $\tilde{x}$ by adding Gaussian noise to the input to obtain a feature representation by learning the approximation function:

$$z = h_{W,b}(\tilde{x}) = f_e(W_e \cdot \tilde{x} + b_e), \tag{1}$$

where W is a weight matrix, b is a bias term, $f_e(\cdot)$ is a non-linear activation function and e represents the encoder, whereas the decoder reconstructs the original signal as close as possible to the uncorrupted original input x:

$$\hat{x} = f_d(W_d \cdot z + b_d), \tag{2}$$

where d represents the decoder. For activation function, empirically-selected exponential linear units (ELUs) [40] and the tanh function [41] were applied. During the training phase, the proposed DAE was trained to minimize the reconstruction error, which was formulated as follows,

$$\mathcal{L}_{DAE} = \frac{1}{2n} \sum_{i=1}^{n} \left\| x_i - \hat{x}_i \right\|^2, \tag{3}$$

where x_i and $\hat{x_i}$ are the original input and the reconstructed output, respectively. Allowing the model to robustly reconstruct the output from partially destroyed input, DAE separates signals from noises and learns the latent features capturing the distribution of the training dataset.

C. CLASSIFICATION

We constructed a FC neural network followed by a softmax [42] layer for the final cell subtype classification phase. Informative features learned from the representation learning procedure were delivered as an input, and the posterior probability of the i_{th} cell subtype was estimated through the softmax function S_i:

$$S_i = \frac{e^{f_i}}{\sum_{i=1}^{C} e^{f_i}}, f_i = WX + b, \tag{4}$$

where f_i is a logit computed from input X from the FC layer, C is the number of cell subtypes, and W and b are a weight matrix and a bias vector of FC layer, respectively. The weights are trained by minimizing cross-entropy, defined as follows:

$$\mathcal{L}_{SM} = -\sum_{i=1}^{n} y_i \cdot log(\hat{y}_i), \tag{5}$$

where y is the correct target label and $\hat{y}$ is a predicted label. After training the FC layers with the softmax classifier based on learned features, we performed additional fine-tuning (FT) to adjust the weights of the trained model for improving the prediction outcome, where we simultaneously minimized the reconstruction error and loss from the FC layers as follows:

$$\mathcal{L}_{FT} = \frac{1}{2n} \sum_{i=1}^{n} \left\| x_i - \hat{x_i} \right\|^2 - \sum_{i=1}^{n} y_i \cdot log(\hat{y}_i) \tag{6}$$

To prevent overfitting, we applied dropout [43], by randomly removing a few nodes to ensure that they had no effect on network decisions and that the network learned more robust features. L2 regularization was also added to the loss function. We used the RMSprop algorithm [44] and the adaptive optimization algorithm, Adam [45] for DAE and FC layers training, respectively. Our proposed model based on a DNN was built by Tensorflow library (Version 1.8.0) [46] and is publicly available at https://github.com/cbi-bioinfo/scDAE.

III. RESULTS

A. EXPERIMENTAL DESIGN

1) DATASET

To rigorously evaluate model performance, we obtained the total 21,679 number of cells for 78 cell subtypes across four scRNA-seq datasets that are publicly available with cell subtype annotations from Gene Expression Omnibus [47] or ArrayExpress [48]. They were organized into two groups of comparable datasets. The number of samples and classes used for training and testing datasets is shown in Table 1 and SupplementaryTable S1. The pancreas group dataset profiled islets of Langerhans cells generated from pancreas tissue: alpha (α), beta (β), delta (δ), and gamma (γ) cells.

TABLE 1. Datasets used for scDAE optimization and evaluation.

Group	Tissue	Number of cell subtypes	Number of cells	Dataset
Pancreas	Pancreatic tissue	4	10,134	GSE81608 [49] GSE84133 [50] GSE85241 [51] E-MTAB-5061 [52]
Mouse cell atlas	Bladder	16	2,746	GSE108097 [53]
	Neonatal muscle	27	4,873	
	Lung	31	3,926	

It consists of four datasets generated from different platforms, where [49]–[51] were produced on SMARTer, inDrop, CEL-Seq2 platform, and [52] was based on the Smart-Seq2 protocol. Cells with the annotations "not applicable" and "contaminated" were removed due to uncertainty, which resulted in a total of 10,134 cells. Based on these datasets, the proposed model was optimized and the performance was compared to the other methods.

The second group consists of three datasets covering major mouse organs, which are bladder, lung, and neonatal muscle, obtained from a mouse cell atlas by Microwell-Seq [53]. Bladder and neonatal muscle dataset contained 16 subtypes of 2,746 cells and 27 subtypes of 4,873 cells, respectively, and lung dataset had 31 cell subtypes with 3,926 cells. It was used for testing the robustness of our model. For each dataset, we randomly selected 70% of samples as a training dataset and 30% of samples as a testing dataset.

2) MODEL OPTIMIZATION

The hyperparameters in scDAE including the depth of hidden layer, number of hidden nodes, learning rate, training epochs, and dropout rate were optimized, and each experiment was repeated ten times. For optimization, we randomly selected 70% of samples as a training data set and 30% of samples as a test data set from the pancreas group described in the Experimental data section. The training set was used for unsupervised pre-training of DAE, as well as, training the FC layers. For the hidden layers and nodes, the proposed DAE architecture was composed of an encoder of two hidden layers, each with 1000 and 500 nodes; and a decoder of two layers, each with 500 and 1000 nodes, and 125 for latent representations (1000-500-125-500-1000), which achieved the best average accuracy (Table 2). Two FC layers, each with 125 hidden nodes, were trained for a final classifier. The learning rate of pre-training and fine-tuning was set to 0.001, the dropout rate was set to 0.3 for fine-tuning, and the corruption level was set to 0.4 for the first layer of the encoder in DAE, showing the best average accuracy. The maximum training epochs for pre-training and fine-tuning were set as 3000 and 1000, respectively. Accuracy results from the experiments with different parameters are shown in the SupplementaryTable S2.

In addition to optimizing the number of hidden nodes, the learning rate, and training epochs, we also performed

TABLE 2. Performance of scDAE under different numbers of hidden nodes and latent representations.

		Number of nodes			
Encoder	layer 1	500	1000	2000	3000
	layer 2	250	500	1000	1500
Latent representation		60	125	250	375
Decoder	layer 1	250	500	1000	1500
	layer 2	500	1000	2000	3000
Average accuracy		98.42%	98.93%	98.91%	98.86%

TABLE 3. Average classification performance results conducting 10-fold cross validation based on the pancreas dataset.

Methods	MCC	Precision	Recall	F1 score	P-value
scDAE	0.9859	0.9916	0.9913	0.9914	-
SVM	0.9845	0.9906	0.9905	0.9905	$P < 10^{-2}$
ACTINN	0.9766	0.9864	0.9857	0.9859	$P < 10^{-5}$
CaSTLe	0.9795	0.9876	0.9875	0.9875	$P < 10^{-5}$
scPred	0.9303	0.9702	0.9571	0.9608	$P < 10^{-10}$
RF	0.9206	0.9659	0.9518	0.9560	$P < 10^{-10}$
CHETAH	0.6573	0.7537	0.7392	0.6879	$P < 10^{-12}$

experiments to find the optimal depth of a hidden layer. We constructed multiple models having different numbers of hidden layers. From the experiments, increasing the number of hidden layers from one to many and repeating that ten times for each model, resulted in a model having two layers showed the best average accuracy of 98.93%, compared to 98.83% and 98.89% for a one-layered model and three-layered model, respectively. Increasing more than three layers dropped the accuracy due to the limited number of raw data.

B. PERFORMANCE EVALUATION OF scDAE

To evaluate scDAE for classifying cell subtypes, we compared our model to the state-of-the-art cell subtype classification methods. The conventional machine learning algorithms such as support vector machine (SVM) with linear kernel [54] and random forest (RF) [55] were also included for the performance comparison. Default parameter settings for each method were used for performance evaluation. Since the pancreas group dataset was used to optimize and evaluate the competing models in previous studies [22]–[24], scDAE was trained using the pancreas group dataset as well, and the performance was measured by 10-fold stratified cross-validation. As scDAE is an unsupervised feature learning classification model, the performance was compared to models that do not require manual feature extraction or marker gene selection. From the experimental results, our proposed model outperformed all other models with respect to the highest average Matthews correlation coefficient (MCC) score of 0.9859 (Table 3, Fig 2). The other models had an average MCC score as follows: ACTINN: 0.9766, SVM: 0.9845, CaSTLe: 0.9795, scPred: 0.9303, RF: 0.9206, while CHETAH showed the lowest MCC score of 0.6573. Statistical significance test for the performance comparison between the scDAE and other methods was also performed based on the Student t-test to MCC scores, where all the comparison results showed P-value < 0.01.

We also further investigated the average precision and recall results for the delta and gamma cell subtypes having relatively low number of samples. scDAE showed the precision and recall higher than 0.98 for both subtypes, while SVM, ACTINN, and CasTLe achieved higher than 0.95. RF showed a low precision of 0.79 and 0.50, high recall of 0.96 and 0.97, respectively for delta and gamma subtype, predicting the alpha and beta cells as them. scPred also showed a low precision of 0.41 and high recall

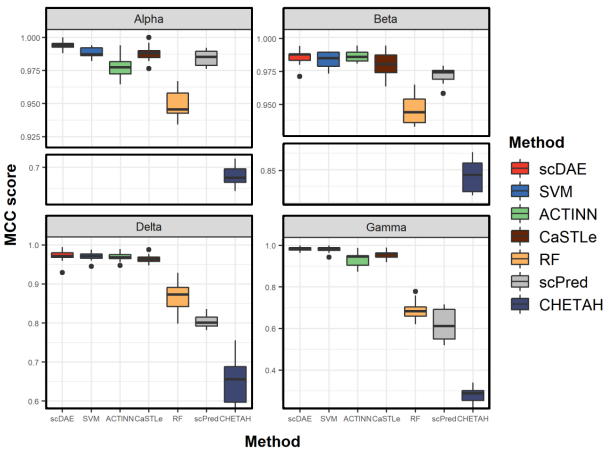

FIGURE 2. Performance comparison of four state-of-the-art cell subtype classification methods and two machine learning algorithms with scDAE conducting 10-fold cross-validation.

of 0.95 for the gamma subtype, however having a high precision of 0.90 and low recall of 0.75, misclassifying the other cells, especially predicting the delta cells as gamma subtype. CHETAH showed low precision and low recall less than 0.82 for both subtypes. Although our proposed model showed improved performance compared to other methods and the ability to learn complex hidden relationships between the high-dimensional gene expression dataset for individual cell subtypes, due to the small number of cell subtypes in the pancreas dataset leading to the relatively easier classification problem compared to the single cell subtype prediction for more than 15 subtypes, the evaluation results did not show a significant performance difference.

Since there was not much room for improving the classification MCC score on the pancreas dataset due to the relatively simple goal, we used three additional data sets. scDAE and competing algorithms were further tested on the bladder, neonatal muscle, and lung tissue dataset, where each of the dataset has 16, 27, and 31 cell subtypes respectively (Table 1). From the result, our proposed model could maintain the average MCC score above 0.9 in classifying more than 30 varying cell subtypes, while others showed significant performance drop when increasing the number of cell subtypes (Fig 3, Table 4). scDAE showed the highest average MCC score of 0.9330 for predicting 31 cell subtypes,

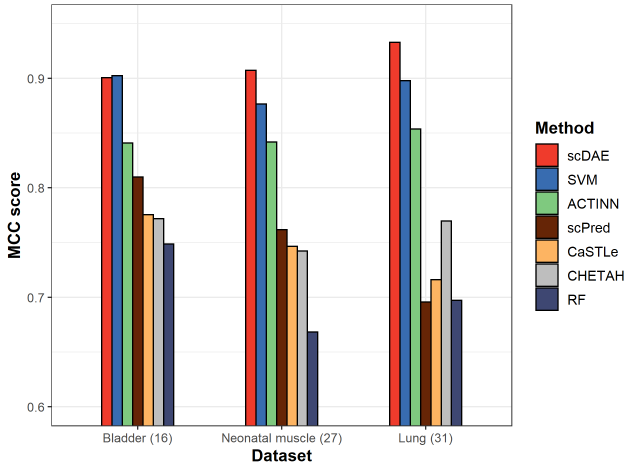

FIGURE 3. Performance validation for predicting a large number of cell subtypes using Mouse cell atlas group dataset.

TABLE 4. Performance comparison based on the average MCC scores for classifying cell subtypes for each tissue dataset in the Mouse cell atlas group datasets.

Dataset	scDAE	SVM	ACTINN	CaSTLe	scPred	RF	CHETAH
Bladder	0.9007	0.9025	0.8410	0.7756	0.8100	0.7486	0.7719
Neonatal muscle	0.9072	0.8767	0.8417	0.7466	0.7619	0.6683	0.7423
Lung	0.9330	0.8978	0.8583	0.7161	0.6958	0.6972	0.7696.

compared to ACTINN: 0.8583, scPred: 0.6958, CaSTLe: 0.7161, and CHETAH: 0.7696. scDAE also outperformed conventional machine learning methods (SVM: 0.8978, RF: 0.6972).

As the mouse cell atlas group dataset has the different number of cells for each subtype, the data imbalance problem can occur. To address this issue, we randomly sampled 50 cells for each cell subtype from the lung tissue dataset and used them for the training and testing dataset by dividing them into a 7 to 3 ratio. Samples for 11 cell subtypes were excluded due to the small number of samples (<50) and the experiment was repeated five times. Our scDAE showed the highest average MCC score of 0.9310, while SVM showed 0.8845, ACTINN: 0.7803, RF: 0.7338, CHETAH: 0.8396, CaSTLe: 0.6466, and ScPred: 0.3318. From the results, the robustness of scDAE was validated both for predicting the cell subtypes using the balanced and imbalanced datasets.

In addition, to visually assess the effect of our proposed representation learning method, we utilized the t-distributed stochastic neighbor embedding (t-SNE) method [56]. The representative features extracted from our model were further compressed into two or three-dimensional t-SNE spaces and labeled for each corresponding cell subtype (Fig 4, SupplementaryFigure S3). Although the loss of information may occur when high-dimensional features are mapped directly to t-SNE spaces, from a visual assessment of our learned features, cell subtypes were clearly separated into individual clusters.

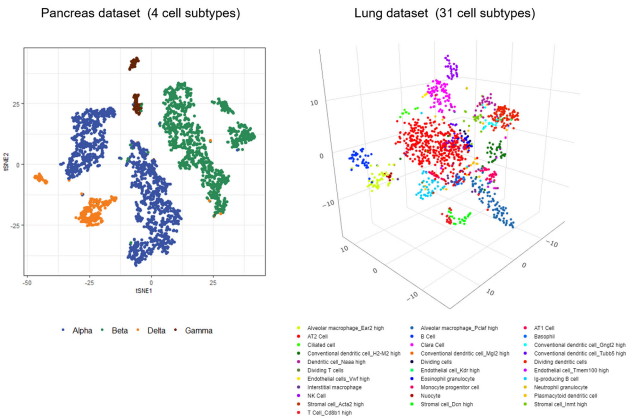

FIGURE 4. t-SNE visualization of the latent features extracted from the proposed representation learning.

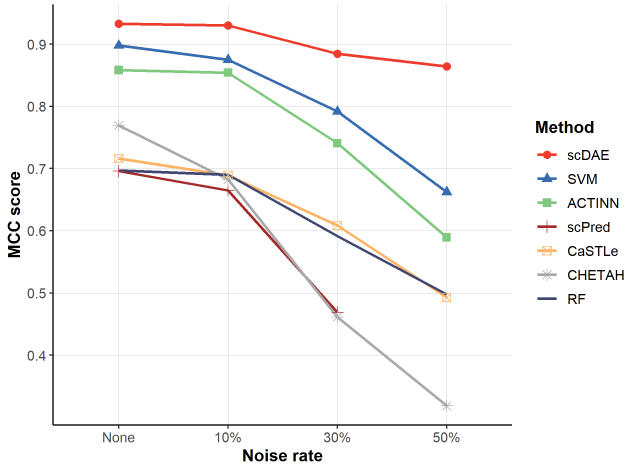

FIGURE 5. Performance change results with different noise rates for scDAE and other methods.

C. TESTING FOR THE ROBUSTNESS OF scDAE

In measuring gene expression profiles in single-cell transcriptomics, various technical factors such as amplification bias, cell cycle effects, and low RNA capture rate can introduce substantial noise in scRNA-seq experiments. These factors can lead to biological signal corruption and false interpretation of analysis results. The impact from these noises becomes more significant as the number of cell subtypes increases. In this experiment, we tested whether scDAE could maintain the stable performance for cell subtype prediction within noises, by generating two noisy datasets based on the lung dataset from the mouse cell atlas group. The first dataset was provided with label noise by randomly shuffling the cell subtype labels of 10%, 30%, and 50% of the training dataset. Label noise brings difficulties for machine learning-based models to extract discriminative features and consistent patterns for classification, leading to degenerate the robustness of learned models [57]. As the noise increases, MCC score for predicting cell subtypes dropped significantly in the competing methods, while scDAE could reasonably maintain (Fig 5, Table 5). When the label noise was introduced up

TABLE 5. Average classification MCC scores on noisy lung dataset for six competing methods.

Noise rate	scDAE	SVM	ACTINN	RF	CaSTLe	CHETAH	scPred
None	0.9330	0.8978	0.8583	0.6972	0.7161	0.7696	0.6958
10%	0.9303	0.8749	0.8541	0.6900	0.6891	0.6825	0.6650
30%	0.8845	0.7918	0.7407	0.5912	0.6084	0.4614	0.4686
50%	0.8641	0.6622	0.5893	0.4974	0.4927	0.3184	-

to 50% of the dataset, scDAE still perform well, with slight 8% performance drop, while others showed 20% to 45% decrease. scDAE could predict most of the cells maintaining the MCC score of 0.86, while the performance of other models decreased less than 0.66. scPred could not predict cells with 50% noise, as it could not identify informative principal components for conventional dendritic cell (H2-M2 high) having noisy 15 cells for the training dataset.

The second dataset for noise test was generated another noisy dataset by converting non-zero read counts to zero values. Dropout event is one of the main problems caused by the low RNA capture rate in scRNA-Seq experiment [58]. Due to the non-trivial distinction between true and false zero counts, not all zeros cannot be considered missing values, where true zero counts represent the lack of gene expression. We randomly selected 30%, 50%, and 70% of genes for each cell in the training dataset and converted the non-zero read counts to zeros. The experiment was repeated five times. In this dropout simulation test, scDAE showed a stable performance compared to other methods (Table 6). From the experiments testing the models for predicting scRNA-seq datasets with substantial noises, our proposed model could provide reliable performance for different input datasets.

Furthermore, we tested our proposed model to validate whether it can accurately identify cell subtype when using a heterogeneous scRNA-seq dataset generated across different sequencing platforms. Training dataset of the pancreas group composed of gene expression profiles from multiple platforms was used, and the prediction accuracy for datasets sequenced on different platforms was measured. From the result, scDAE trained on datasets generated from SMARTer, inDrop, CEL-Seq2 platform, and tested on a dataset based on the Smart-Seq2 protocol, achieved an average accuracy of 99.43%, which demonstrated the proposed model is robust to platform bias. Other results based on different combinations of training and testing datasets are available in SupplementaryTable S4.

TABLE 6. Average prediction MCC score result on lung dataset adding zero count noises.

Zero count noise rate	scDAE	SVM	ACTINN	RF	CaSTLe	CHETAH	scPred
None	0.9330	0.8978	0.8583	0.6972	0.7161	0.7696	0.6958
30%	0.9325	0.8870	0.8433	0.6741	0.6967	0.7469	0.6158
50%	0.9303	0.8672	0.8382	0.6614	0.6439	0.7080	0.5200
70%	0.9163	0.8338	0.7296	0.5381	0.5123	0.5369	0.3032

D. IDENTIFICATION OF NEW CELL SUBTYPES

In scRNA-seq experiment, it could be possible that the new cell subtype is only introduced in future data but not in the training data set. To show whether our proposed model can be utilized for identifying a new cell subtype, we designed an additional experiment. We output the probabilities for each cell subtype estimated through the softmax function in the classification step from our scDAE and labeled the cell 'uncertain' if the highest probability is lower than 0.95, otherwise classified as a predicted cell type. From the bladder dataset in the mouse cell atlas group, we excluded "Basal epithelial" cells from the dataset, and only added 100 samples of those cells to the testing dataset as new cell subtype (Training data: 15 cell subtypes, Testing data: 16 cell subtypes). From the results, most of 15 cell subtypes were classified correctly with an accuracy of 93.20%. Moreover, 90% of Basal epithelial cells were assigned "uncertain" showing that our proposed model is able to identify cell subtype that was not introduced in the training dataset. We added this use case to the manual provided from our github repository (https://github.com/cbi-bioinfo/scDAE).

IV. DISCUSSION

ScRNA-seq has provided the characterization of transcriptomic profiles at the single-cell resolution and has been widely applied in biological and medical research. Identification of single-cell subtype is an essential step before in-depth investigations and further analysis of their functional roles. Several supervised-based methods utilizing machine learning algorithms have been developed, but they still suffer from the high level of noises. To address those issues in cell subtype classification, we developed a DNN-based model employing a multilayer DAE to extract informative representations robust to the noises and trained the model to predict cell subtypes.

We first obtained two datasets, the pancreas group and mouse cell atlas group dataset, and evaluated our proposed model for classifying cell subtypes. The performance comparison with the state-of-the-art cell identification methods shows that scDAE provides more accurate prediction results and stable performance by maintaining the MCC score above 0.9 for classifying more than 30 cell subtypes. We also visually assessed the effect of the representation learning method utilizing t-SNE, and it is proved that our representation learning through the proposed DAE-based model has the potential to correctly extract features from a high-dimensional dataset and map them to a low-dimensional space.

Moreover, we tested our model's robustness to the noises caused by technical factors, which could lead to false biological analysis and corrupt the investigation of functional roles for each cell subtype. Two noisy datasets were generated based on the lung tissue dataset from the mouse cell atlas group by introducing the label noises and false zero counts to simulate the dropout event. From our experiments, scDAE showed robust cell subtype classification for both

noises. By employing a DAE to separate signals from noises, allowing the model to robustly reconstruct the output from partially destroyed input, scDAE could learn complex hidden relationships between the high-dimensional gene expression dataset for individual cell subtypes and predict most of the noisy cells maintaining the classification performance compared to the competing method.

Overall, these findings indicate that scDAE can help understand a heterogeneous population composed of various cells by providing the accurate annotation of cell subtypes and minimizing the bias caused by noises. It is also expected that our proposed model will facilitate in-depth biological findings by discovering new cell subtypes, which can support to study complex differentiation and developmental trajectories and explore the cell basis of human disease.

V. CONCLUSION

In this paper, we presented a DNN-based cell subtype classification model utilizing a scRNA-seq dataset, scDAE. The proposed model was designed to employ DAEs to learn informative representations from input data and learned features were further tuned through FC layers to improve the classification accuracy. The model was then evaluated with four different state-of-the-art cell subtype classification methods and two conventional machine learning methods with 10-fold cross-validation. scDAE outperformed all other models with the highest MCC score and demonstrated stable performance for three group datasets with various cell subtypes. The model also showed stable performance for predicting cell subtypes when the noise was introduced. Moreover, the effect of our representation learning method using DAE was assessed through t-SNE visualization, and it was shown to have the ability to extract significant features from input data and to capture discriminative patterns automatically by learning the relationship between features. We believe that our classifier will efficiently predict cell subtype on a well-trained representation learning model, which may help to improve the precision of single-cell analysis. In future research work, we will consider developing an interpretable neural network-based cell subtype classification model, which can help us to identify marker genes for each cell subtype.

REFERENCES

[1] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Exp. Mol. Med.*, vol. 50, no. 8, pp. 1–14, Aug. 2018.

[2] L. Kester and A. van Oudenaarden, "Single-cell transcriptomics meets lineage tracing," *Cell Stem Cell*, vol. 23, no. 2, pp. 166–179, Aug. 2018.

[3] J. Park, R. Shrestha, C. Qiu, A. Kondo, S. Huang, M. Werth, M. Li, J. Barasch, and K. Suszták, "Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease," *Science*, vol. 360, no. 6390, pp. 758–763, May 2018.

[4] A. K. Shalek *et al.*, "Single-cell RNA-seq reveals dynamic paracrine control of cellular variation," *Nature*, vol. 510, no. 7505, p. 363, 2014.

[5] H. M. Levitin, J. Yuan, and P. A. Sims, "Single-cell transcriptomic analysis of tumor heterogeneity," *Trends Cancer*, vol. 4, no. 4, pp. 264–268, Apr. 2018.

[6] M. J. T. Stubbington, O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann, "Single-cell transcriptomics to explore the immune system in health and disease," *Science*, vol. 358, no. 6359, pp. 58–63, Oct. 2017.

[7] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suva, A. Regev, and B. E. Bernstein, "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, Jun. 2014.

[8] W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, and W.-Y. Park, "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer," *Nature Commun.*, vol. 8, no. 1, p. 15081, Aug. 2017.

[9] K.-T. Kim, H. W. Lee, H.-O. Lee, H. J. Song, D. E. Jeong, S. Shin, H. Kim, Y. Shin, D.-H. Nam, B. C. Jeong, D. G. Kirsch, K. M. Joo, and W.-Y. Park, "Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma," *Genome Biol.*, vol. 17, no. 1, p. 80, Dec. 2016.

[10] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh, "Robust enumeration of cell subsets from tissue expression profiles," *Nature Methods*, vol. 12, no. 5, p. 453, 2015.

[11] B. Li, E. Severson, J.-C. Pignon, H. Zhao, T. Li, J. Novak, P. Jiang, H. Shen, J. C. Aster, S. Rodig, S. Signoretti, J. S. Liu, and X. S. Liu, "Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy," *Genome Biol.*, vol. 17, no. 1, p. 174, Dec. 2016.

[12] D. Aran, Z. Hu, and A. J. Butte, "XCell: Digitally portraying the tissue cellular heterogeneity landscape," *Genome Biol.*, vol. 18, no. 1, p. 220, Dec. 2017.

[13] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden, "Single-cell messenger RNA sequencing reveals rare intestinal cell types," *Nature*, vol. 525, no. 7568, p. 251, 2015.

[14] C. Xu and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method," *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, Jun. 2015.

[15] J. M. Zhang, J. Fan, H. C. Fan, D. Rosenfeld, and D. N. Tse, "An interpretable framework for clustering single-cell RNA-seq datasets," *BMC Bioinf.*, vol. 19, no. 1, p. 93, Dec. 2018.

[16] B. Wang, D. Ramazzotti, L. De Sano, J. Zhu, E. Pierson, and S. Batzoglou, "SIMLR: A tool for large-scale genomic analyses by multi-kernel learning," *Proteomics*, vol. 18, no. 2, Jan. 2018, Art. no. 1700232.

[17] S. Park and H. Zhao, "Spectral clustering based on learning similarity matrix," *Bioinformatics*, vol. 34, no. 12, pp. 2069–2076, Jun. 2018.

[18] M. B. Pouyan and D. Kostka, "Random forest based similarity learning for single cell RNA sequencing data," *Bioinformatics*, vol. 34, no. 13, pp. i79–i88, Jul. 2018.

[19] R. Zheng, M. Li, Z. Liang, F.-X. Wu, Y. Pan, and J. Wang, "SinNLRR: A robust subspace clustering method for cell type detection by non-negative and low-rank representation," *Bioinformatics*, vol. 35, no. 19, pp. 3642–3650, Oct. 2019.

[20] X. Zhao, S. Wu, N. Fang, X. Sun, and J. Fan, "Evaluation of single-cell classifiers for single-cell RNA sequencing data sets," *Briefings Bioinf.*, vol. 21, no. 5, pp. 1581–1595, Sep. 2020.

[21] V. Y. Kiselev, A. Yiu, and M. Hemberg, "Scmap: Projection of single-cell RNA-seq data across data sets," *Nature Methods*, vol. 15, no. 5, p. 359, 2018.

[22] Y. Lieberman, L. Rokach, and T. Shay, "Correction: CaSTLe–classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments," *PLoS ONE*, vol. 13, no. 11, Nov. 2018, Art. no. e0208349.

[23] J. Alquicira-Hernandez, A. Sathe, H. P. Ji, Q. Nguyen, and J. E. Powell, "Scpred: Accurate supervised method for cell-type classification from single-cell RNA-seq data," *Genome Biol.*, vol. 20, no. 1, pp. 1–17, 2019.

[24] J. K. de Kanter, P. Lijnzaad, T. Candelli, T. Margaritis, and F. C. Holstege, "CHETAH: A selective, hierarchical cell type identification method for single-cell RNA sequencing," *Nucleic Acids Res.*, vol. 47, no. 16, p. e95, 2019.

[25] H. A. Pliner, J. Shendure, and C. Trapnell, "Supervised classification enables rapid annotation of cell atlases," *Nature Methods*, vol. 16, no. 10, pp. 983–986, 2019.

[26] A. W. Zhang *et al.*, "Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling," *Nature Methods*, vol. 16, no. 10, pp. 1007–1015, Oct. 2019.
[27] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015.
[28] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings Bioinf.*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.
[29] J. Tan, J. H. Hammond, D. A. Hogan, and C. S. Greene, "ADAGE-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions," *mSystems*, vol. 1, no. 1, p. e00025, Feb. 2016.
[30] J. Tan, G. Doing, K. A. Lewis, C. E. Price, K. M. Chen, K. C. Cady, B. Perchuk, M. T. Laub, D. A. Hogan, and C. S. Greene, "Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks," *Cell Syst.*, vol. 5, no. 1, pp. 63–71, 2017.
[31] G. P. Way and C. S. Greene, "Evaluating deep variational autoencoders trained on pan-cancer gene expression," 2017, *arXiv:1711.04828*. [Online]. Available: http://arxiv.org/abs/1711.04828
[32] G. P. Way and C. S. Greene, "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *Proc. Pacific Symp. Biocomput.*, vol. 23, 2018, pp. 80–91.
[33] C. Lin, S. Jain, H. Kim, and Z. Bar-Joseph, "Using neural networks for reducing the dimensions of single-cell RNA-seq data," *Nucleic Acids Res.*, vol. 45, no. 17, p. e156, Sep. 2017.
[34] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, no. 12, p. 1053, 2018.
[35] F. Ma and M. Pellegrini, "ACTINN: Automated identification of cell types in single cell RNA sequencing," *Bioinformatics*, vol. 36, no. 2, pp. 533–538, 2020.
[36] T. Tian, J. Wan, Q. Song, and Z. Wei, "Clustering single-cell RNA-seq data with a model-based deep learning approach," *Nature Mach. Intell.*, vol. 1, no. 4, p. 191, 2019.
[37] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni, "Normalizing single-cell RNA sequencing data: Challenges and opportunities," *Nature Methods*, vol. 14, no. 6, p. 565, 2017.
[38] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, Dec. 2014.
[39] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, "Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1930–1943, Aug. 2013.
[40] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: http://arxiv.org/abs/1511.07289
[41] B. Karlik and A. V. Olgac, "Performance analysis of various activation functions in generalized mlp architectures of neural networks," *Int. J. Artif. Intell. Expert Syst.*, vol. 1, no. 4, pp. 111–122, 2011.
[42] G. E. Hinton and R. R. Salakhutdinov, "Replicated softmax: An undirected topic model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1607–1614.
[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[44] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: http://arxiv.org/abs/1609.04747
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[46] M. Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/
[47] R. Edgar, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–210, Jan. 2002.
[48] A. Brazma, "ArrayExpress—A public repository for microarray gene expression data at the EBI," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 68–71, Jan. 2003.
[49] Y. Xin, J. Kim, H. Okamoto, M. Ni, Y. Wei, C. Adler, A. J. Murphy, G. D. Yancopoulos, C. Lin, and J. Gromada, "RNA sequencing of single human islet cells reveals type 2 diabetes genes," *Cell Metabolism*, vol. 24, no. 4, pp. 608–615, Oct. 2016.
[50] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai, "A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure," *Cell Syst.*, vol. 3, no. 4, pp. 346–360, 2016.
[51] M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti, E. J. P. de Koning, and A. van Oudenaarden, "A single-cell transcriptome atlas of the human pancreas," *Cell Syst.*, vol. 3, no. 4, pp. 385–394, 2016.
[52] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, D. M. Smith, M. Kasper, C. Ämmälä, and R. Sandberg, "Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes," *Cell Metabolism*, vol. 24, no. 4, pp. 593–607, Oct. 2016.
[53] X. Han *et al.*, "Mapping the mouse cell atlas by microwell-seq," *Cell*, vol. 172, no. 5, pp. 1091–1107, 2018.
[54] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.
[55] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[56] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[57] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8527–8537.
[58] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature Methods*, vol. 11, no. 7, pp. 740–742, Jul. 2014.

JOUNGMIN CHOI received the master's degree in computer science from Sookmyung Women's University, South Korea, in 2020. Her research interests include machine learning and epigenetic data analysis.

JE-KEUN RHEE received the Ph.D. degree in bioinformatics from the Interdisciplinary Program in Bioinformatics, Seoul National University, in 2014. He is currently an Assistant Professor with the Department of Bioinformatics and Life Science, Soongsil University. His research interests include bioinformatics, machine learning, and cancer genome analysis.

HEEJOON CHAE received the Ph.D. degree in computer science from Indiana University at Bloomington, in 2016. He is currently an Assistant Professor with the Division of Computer Science, Sookmyung Women's University. His research interests include epigenomics and machine learning.