Predicting the Reprogrammability of Human Cells Based on Transcriptome Data and SGD Classifier with Elastic-Net Regularization

Gorkem Saygili1,2,* , Mine Turktas3, Cansu Gurcan2,4, Lucia Gemma Delogu5, and Acelya Yilmazer Aktuna2,4,*

1Tilburg University, Cognitive Science and Artificial Intelligence Department, Tilburg, The Netherlands
2Ankara University, Biomedical Engineering Department, Ankara, Turkey
3Gazi University, Department of Biology, Ankara, Turkey
4Ankara University, Stem Cell Institute, Ankara, Turkey
5University of Padua, Department of Biomedical Sciences, Padua, Italy
*gsaygili@tilburguniversity.edu, ayilmazer@ankara.edu.tr

ABSTRACT

Cell reprogramming has shown considerable importance in recent years; however, the programmability of cells and efficiency of reprogramming varies across different cell types. Considering several weeks of cell programming process and costly programming agents used through the process, every failure in reprogramming comes with a significant burden. Better planning for reprogramming experiments could be possible if there is a way of predicting the outcome of reprogramming before the experiments using transcriptome data. In this study, we have accessed the transcriptome data of successful or unsuccessful programming studies published in literature and constructed a Stochastic Gradient Descent (SGD) classifier with Elastic-Net regularization for predicting whether the cell lines are reprogrammable. We tested our classifier using 10-fold cross validation over cell lines and on each cell separately. Our results showed that it is possible to predict the outcome of cell reprogramming with accuracies up to 98% and Area Under the Curve (AUC) scores up to 0.98%. Considering the success of our experimental outcomes we conclude that an outcome of a cell reprogramming experiment can be predicted with high accuracy using machine learning on transcriptome data.

Introduction

There is an increasing trend for using artificial intelligence (AI) in healthcare and biomedical sciences. As machine learning (ML) algorithms and computing infrastructure develop, these technologies become part of diagnosis and treatment of various human pathologies. Considering the importance of applying personalized medicine and the growing trend on the application of ML techniques in medicine field, numerous studies made use of these technologies to achieve cell reprogramming, stem-cellbased therapies and cancer cell prediction1–5. There are also studies in literature aimed for precision oncology: cancer sub-typing based on various different genomic, molecular and pathological parameters6–9.

It is very well-known that cells can be reprogrammed back to a pluripotent stage by overexpression of reprogramming factors. Cell reprogramming through forced expression of reprogramming factors (including Oct3/4, Klf4, Sox2, c-Myc) has been reported in several studies in literature10–12. Different cell types have been successfully reprogrammed until now, including fibroblasts, hematopoietic stem cells, hepatocytes and even various cancer cells. For example, several groups have tried to reprogram cancer cells via different protocols including the use of embryonic microenvironment, nuclear transfer or transfection with the reprogramming factors resulting in altered tumorigenicity13–18. Even though cell reprogramming has shown considerable importance in recent years, the programmability of cells and outcome of reprogramming varies across different cell types. Furthermore, cell reprogramming requires several weeks of tedious effort and considerable cost regarding to utilized reprogramming agents and materials. Therefore, it is clear that there is a need for predicting the outcome of reprogramming in this field to help scientists to better plan for the reprogramming experiments. ML algorithms have been used in single-cell RNA sequence analysis specifically for cell identification and clustering19–21. Considering the vast variety of different ML algorithms, it is important to choose the best performing algorithm on the applied domain. Abdelaal et al.22 compared the performances of 22 classification methods including the most popular ones with Random Forest (RF), Support Vector Machines (SVM), and Neural Networks on 27 publicly available single cell RNA sequencing datasets and revealed that linear SVM had the overall best overall performance. In a different study, Michielsen et al.23 compared linear SVM
and one-class SVM and found that linear SVM outperformed one-class SVM on cell classification using RNA-seq data. The datasets that were used for both studies contain cells up to hundreds of thousands. Having a large number of samples is crucial since, depending on the their complexity, ML algorithms may require considerable amount of samples for training. Large amount of data are at the disposal of the researchers to train their algorithm to do cell classification however such data is not available for cell reprogramming in particular for the failure of the programming scenarios. When the data is limited, designing an accurate ML algorithm becomes a challenging task because of the well-known overfitting problem. To cope with it, regularization parameters such as L1 and L2 regularization have been used frequently. Elastic-Net is a combination of L1 and L2 regularizations to achieve the best possible performance against overfitting. It has been preferred recently for predicting immune cell tags because of its favorable performance on high dimensional data.

There are two main contributions of this study:

- For the first time, we show that programming outcome of a cell can be predicted before the programming process using ML and transcriptome data.

- Elastic-Net regularization with an SGD classifier provides better performance with limited number of samples than the overall best performer of previous studies, linear SVM.

Our second contribution is particularly important when there is limited number of cells such as rare cancer cell populations and the dimensionality of the data is large which is in general true for transcriptome data. Because in such instances, using deep neural networks could not be an option because of their large sample size demand for accurate training.

Results

**Performance comparison of 10 Machine Learning Algorithms Over the Replicates Indicates Predicting the Outcome of Reprogramming is Achievable with High Accuracy**

In order to achieve a higher sample number to establish ML-based algorithms, we accessed transcriptome data of various reprogramming studies from the literature that resulted in either success or failure. Transcriptome reads that have been downloaded included that of melanoma cells from Castro-Perez et al., human CD34+ cells from Friedli et al., human fibroblast cells from different studies, human acute myeloid leukemia cells from Chao et al., human thyroid-carcinoma cell-lines from Kong et al. Out of these 18 cell lines (consisting of 38 replicate (cells)), 12 cell lines (26 replicates) have been reported to be successfully reprogrammed whereas 6 (12 replicates) of them failed to generate iPSC colonies. On this data we have attempted to establish ML-based algorithm to predict the programmability of cells based on their transcriptome data of prior to reprogramming.

We predicted whether each a replicate is reprogrammable or not with leave-one-out cross validation (LOOCV) experiment. Through the experiment, a replicate was left out and the remaining replicates were used as the training set. Since our dataset was reasonably small, we aimed to have the maximum number of training samples with this approach. We repeat LOOCV experiment 10 times to minimize the stochastic results as much as possible. The hyperparameter optimization step was done using 3-fold CV on the training set and the prediction was made for the replicate in the test set with the best-performing hyperparameters on the training set. The Receiver Operating Characteristic (ROC) and Precision-Recall curves are presented in Figures 1a and 1b, respectively. From both results, Linear Support Vector Machine (SVM), Generalized Linear Model (GLM), Random Forest (RF), Naïve Bayes (NB), Stochastic Gradient Descent with Elastic-Net regularization (SGD), and XGBoost provided top performances with over 0.9 AUC, accuracy, and AP scores. In contrast to these top performers, Two classes of reprogrammable and non-reprogrammable cells were again separable with reasonably high accuracy which can be perceived from the density plots and the ROC curve. The AUC score of 1 indicates almost perfect classification performance. To have a detailed performance comparison, Table 1 presents the accuracy, sensitivity, specificity, precision, F1 score and Matthew’s Correlation Coefficient (MCC) scores of each classifier for LOOCV experiment. In particular, MCC scores are more reliable compared to other measures, thanks to its balanced use of all confusion matrix components. Based on MCC score, SGD with Elastic-Net regularization (SGD_Elastic), NB, GLM, and SVM performed on par and outperformed the other classifiers. All of these four classifiers also achieved high accuracy, sensitivity, specificity, precision and F1-score close, if not equal to 100%. The overall results indicate that the reprogramming outcome of a replicate can be predicted using transcriptome data and ML.

**Performance comparison of 10 Machine Learning Algorithms Over the Celllines Predicting the Outcome of Reprogramming is Achievable with High Accuracy**

Since ML algorithms may exploit transcriptional similarities between the replicates, we conducted another experiment, so called cell line experiment, in which we separated our test set to contain all the replicates from a cell line and hence there is no replicate from that cell line in the training set. Since there is no replicates of the same cell line in both training set and the test set, the ML algorithms can only exploit transcriptional differences between successfully programmed cells and the unsuccessful
Table 1. Accuracy, sensitivity, specificity, precision, F1-score and Matthew’s correlation coefficient (MCC) scores of 9 classifiers from LOOCV experiments. The top performance for each score is identified as bold.

| Classifier   | Accuracy | Sensitivity | Specificity | Precision | F1 Score | MCC    |
|--------------|----------|-------------|-------------|-----------|----------|--------|
| Adaboost     | 0.85     | 0.87        | 0.80        | 0.90      | 0.89     | 0.66   |
| KNN          | 0.87     | 0.85        | 0.92        | 0.96      | 0.90     | 0.73   |
| LDA          | 0.55     | 0.50        | 0.67        | 0.76      | 0.60     | 0.16   |
| SVM          | 0.97     | 1.00        | 0.92        | 0.96      | 0.98     | 0.94   |
| GLM          | 0.97     | 0.96        | 1.00        | 1.00      | 0.98     | 0.94   |
| NB           | 0.97     | 1.00        | 0.92        | 0.96      | 0.98     | 0.94   |
| XGBoost      | 0.92     | 0.92        | 0.96        | 0.94      | 0.94     | 0.82   |
| RF           | 0.95     | 0.92        | 1.00        | 1.00      | 0.96     | 0.89   |
| SGD_Elastic  | 0.98     | 0.98        | 0.98        | 0.99      | 0.98     | 0.95   |

Figure 1. (a) ROC curve and (b) Precision-Recall curve of utilized classifiers from LOOCV experiments. Area Under ROC Curve (AUC), accuracy, and Average Precision score (AP) scores are provided for each classifier.

We repeated cell line experiment 10 times to minimize the stochastic results as much as possible. Same as the LOOCV experiment, the hyperparameter optimization step was done using 3-fold CV on the training set. The prediction was made all replicates in a cell line in the test set with the best-performing hyperparameters on the training set. The Receiver Operating Characteristic (ROC) and Precision-Recall curves are presented in Figures 2a and 2b, respectively. Different from the previous results, classifiers generally performed less accurately in this experiment. Two ensemble learners, Adaboost and XGBoost, and LDA performed less accurately compared to other classifiers. Table 2 presents the accuracy, sensitivity, specificity, precision, F1 score and MCC scores of each classifier for the cell line experiment. The decrease in the performance of classifiers is clearly observed from the MCC scores. SGD classifier with Elastic-Net regularization had on par performance with LOOCV experiment however all other classifiers performed considerably worse than SGD. Furthermore, all the remaining scores of SGD is also similar to its LOOCV performance, proving that SGD classifier with Elastic-Net regularization achieved the best overall performance in the prediction of reprogramming outcome.

SGD classifier revealed important genes in the prediction of reprogramming outcome

The weights assigned to features can be used to assess the importance of that feature for the prediction of the target variable. SGD classifier with Elastic-Net regularization provides non-negative feature weights and considerable amount of them can be zero on a high dimensional dataset because of L1 regularization. Hence, SGD classifier intrinsically has feature selection capability and eventually features with larger weights indicate important features for the classification task. Figure 3a and 3b show the five genes that have the highest weights in the classification task. Both figures indicate the same set of genes namely, RN7SK, 7SK, RN7SL2, RN7SL1, and MT-RNR2. The others indicate the sum of the weights of the remaining genes. Since both experiments indicate the same set of genes, we advocate that these genes play a crucial rule for the outcome of reprogramming.
Dimensionality reduction methods reveal the distinction between successfully programmed and failed to be reprogrammed replicates. Dimensionality reduction such as Principal Component Analysis (PCA)\textsuperscript{19}, t SNE, Neighborhood Preserving Embedding (t-SNE)\textsuperscript{31}, and Unifold Manifold Approximation and Projection (UMAP)\textsuperscript{42} are commonly used with transcriptome data to visualize the clusters of the samples. Two popular visualization algorithms, t-SNE and UMAP, provide faithful representations of high dimensional data, preserving close neighborhoods in the embeddings. All of the embedding results are presented in Fig. 4 which clearly show that there are two clusters consists of successfully reprogrammed and failed to be reprogrammed replicates. In particular, the two clusters are clearly visible on t-SNE (Fig. 4b) and UMAP (Fig. 4c) embeddings. Two clusters in the lower dimensional space clearly indicates the existing similarities in terms of transcriptome profiles of the in between successfully reprogrammed replicates and in between failed to be programmed replicates.

### Discussion

Continued research and development in cell reprogramming technology will be a valuable tool for more refined in vitro disease models, identification of disease biomarkers and ultimately for the development of novel treatment protocols for patients in the future\textsuperscript{10,14-17}. For example, cancer cell reprogramming studies can provide additional opportunities including: modifying cancer stem cell populations, which would be an important tool to better understand their biological properties in resistant tumors; engineering the resulting reprogrammed cancer cells with altered tumorigenicity in order to develop novel cancer vaccines; using cancer-derived iPSC for pharmacological screenings; or creating a powerful in vitro model to understand the epigenetic and genetic mechanisms taking place during tumor development and progression\textsuperscript{19}. Published studies indicate that the starting cell status is highly variable and, in turn, may correspond to partial or complete pluripotent states\textsuperscript{14}. Therefore, there is a need for establishing ML-based prediction prior to starting the reprogramming studies. Looking at the recent years, there is

| Classifier      | Accuracy | Sensitivity | Specificity | Precision | F1 Score | MCC  |
|-----------------|----------|-------------|-------------|-----------|----------|------|
| Adaboost        | 0.88     | 0.91        | 0.83        | 0.92      | 0.91     | 0.73 |
| KNN             | 0.89     | 0.88        | 0.92        | 0.96      | 0.92     | 0.77 |
| LDA             | 0.68     | 0.58        | 0.92        | 0.94      | 0.71     | 0.46 |
| SVM             | 0.92     | **1.00**    | 0.75        | 0.90      | 0.95     | 0.82 |
| GLM             | 0.89     | 0.85        | **1.00**    | **1.00**  | 0.92     | 0.80 |
| NB              | 0.92     | 0.92        | 0.92        | 0.96      | 0.94     | 0.82 |
| XGBoost         | 0.63     | 0.73        | 0.42        | 0.73      | 0.73     | 0.15 |
| RF              | 0.89     | 0.85        | **1.00**    | **1.00**  | 0.92     | 0.80 |
| SGD_Elastic     | **0.97** | 0.98        | 0.96        | 0.98      | **0.98** | **0.93** |

Table 2. Accuracy, sensitivity, specificity, precision, F1-score and Matthew’s correlation coefficient (MCC) scores of 9 classifiers from Cellline experiments. The top performance for each score is identified as bold.
a growing trend for using ML-based approaches in biomedical sciences. This study aims to predict the reprogrammability of cells based on a ML-based algorithm which was trained on already published data. It is of crucial importance to validate the predictive performance of these models; therefore, this study makes an important contribution in the field by both establishing a ML-based prediction.

In this study, we aim to predict the reprogrammability of various cell types using 9 different classifiers including the best classifier for single cell classification based on transcriptome data, linear SVM, as mentioned in\textsuperscript{22,23}. Both LOOCV and Cell line experiment results showed that accurate predictions up to 97% accuracy are possible using transcriptome data before the reprogramming. Even more importantly, more reliable measures such as F1 score and MCC are also both over 0.93 for both experiments with SGD Classifier with Elastic-Net regularization. These results show that it is possible to predict the outcome of a reprogramming experiment based on transcriptome data and ML. Although we have not tested, we believe that stochastic gradient descent which is the main optimizer of deep neural networks, can be trained and tested faster than SVM.

An important bottleneck of this study is the size of its dataset. Our dataset is limited in terms of number of samples because the transcriptome data with unsuccessful reprogramming outcome is limited. Considering the high dimensionality of transcriptome data and limited number of samples, overfitting problem becomes a crucial concern. The difference between SGD classifier and linear SVM in our case is the Elastic-Net regularization in which Ridge (L2) and Lasso (L1) regularizations are combined. In comparison to Linear SVM which incorporates only Ridge regularization, Elastic-Net provides considerable robustness against overfitting and increases the overall performance as depicted in our results. Although plain SGD classifier with Ridge loss is similar to linear SVM, SGD classifier with Elastic-Net regularization clearly outperformed linear SVM. The power of Elastic-Net regularization against overfitting has been also showed in\textsuperscript{22}. We believe the addition of Lasso regularization which can cause some of the weights to be exact zero in the classifier causes an effect of adaptive feature selection through the training and leads to a considerable increase in performance. This advantage can be exploited in different classification tasks with transcriptome data when the available data is limited in terms of samples.

Since our overall classification performance is high, we explored the genes that played crucial role in the prediction of the reprogramming experiment. Our results show that RN7SK (ENSG00000283293), 7SK (ENSG00000202198), RN7SL2 (ENSG00000274012), RN7SL1 (ENSG00000276168), and MT-RNR2 (ENSG00000210082) were the top five important genes that contributed to the classification. RN7SK has shown to play a role in cellular differentiation\textsuperscript{43} and senescence\textsuperscript{44}. Studniarek et al.\textsuperscript{45} have found that 7SK play an important role in proper stress response and cell survival. Perhaps more importantly, Eilebrecht et al.\textsuperscript{46} has shown that 7SK directly affects HMGA1 function which induces the pluripotency gene expression. In their work, Nabet et al.\textsuperscript{47} showed that RN7SL1 activates PRR RIG-I which eventually enhances tumor growth. Following up these studies, our analysis is also suggesting that non-coding RNAs are important regulatory elements in cellular reprogramming and differentiation processes. On the other hand, Palombo et al.\textsuperscript{48} revealed the role of mitochondrial DNA (MtDNA) variants’ affect on the phenotype of iPSCs and their healthiness. Therefore, based on our analysis, it would be important to perform variant analysis of key MtDNA genes such as MT-RNR2 prior to start of cellular reprogramming experiments in order to ensure good quality
iPSC generation. We believe that analyzing the feature importances of accurate classifiers further could reveal many more essential genes that play a crucial role in explaining various factors that affect the outcome cell reprogramming.

![Figure 4](image)

**Figure 4.** (a) PCA, (b) t-SNE, and (c) UMAP embeddings of the successfully reprogrammed (success) and failed to be reprogrammed (failure) replicates.

### Methods

#### Data retrieval

The transcriptome data of Friedli et al.\(^{30}\), Tanaka et al.\(^{31}\), Castro-Perez et al.\(^{29}\) was used for the analyses and data was retrieved from NCBI Gene Expression Omnibus database. The accession numbers were given in the Supplementary Table 1. Our data consisted of 38 replicates samples from 18 cell lines and 57250 expression levels. Each cell line corresponded to one of our samples whereas each expression level constituted our features. In total, 26 of those cells were reprogrammed successfully (labeled as 1), whereas the remaining 12 were failed to be reprogrammed (labeled as 0).

#### RNA-Seq data analysis

The RNA-seq data analysis was performed using the new version of the Tuxedo protocol with HISAT2, StringTie and Ballgow\(^{49}\). Homo sapiens GRCh38 reference genome was used for genome indexing. The transcriptome reads were aligned to the indexed genome using Hisat2\(^{50}\). Mapped reads of each sample were assembled and the transcript counts in FPKM were obtained using StringTie\(^{49}\).

#### Application of machine learning algorithms

The majority of 9 classifiers were chosen based on the classifiers that were used in\(^{22}\). Multi layer perceptron (MLP) classifier has not been used since depending on the number of hidden layers, MLP classifiers need substantial amount of data to be trained successfully. Rather than MLP, different from\(^{22}\), we incorporated boosting ensemble learners such as Adaboost and XGBoost. Additionally, we included SGD classifier with Elastic-Net regularization which achieved the highest overall performance.

An important step in machine learning experiments is the hyperparameter optimization step. The hyperparameters differ among different ML algorithms, for example number of estimators and maximum depth are important hyperparameters of RF classifiers whereas the type of the kernel (RBF, linear, etc.), gamma and C values are important hyperparameters of SVM classifiers. To optimize these parameters for our ML algorithm, we applied 3-fold cross validation and grid search on the training set for each of our experiments on training set. Hence, we optimize each ML algorithm on the training set and tested this optimized ML algorithm on the test set to obtain the best possible predictions.

There are several scores to quantify the performance of a classifier. Among those, accuracy is the most common score. Yet, accuracy does not always present the whole performance especially when class labels are imbalanced. Another performance measure is the Receiver Operator Characteristic (ROC) curve and Area Under the Curve (AUC) score measured from the ROC. An AUC score close to 0.5 indicates that the classifier produces random predictions whereas an AUC score close to 1 signifies high performance. In our experiments, we used several classifier performance measures such as F1 Score, AUC, and Matthew’s Correlation Coefficient (MCC) to interpret the performance of our classifier.

We used Python programming language and scikit-learn toolkit (version 0.22) in our experiments on a standard computer with 32 GB of memory.

### References

1. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer
prognosis and prediction. *Comput. structural biotechnology journal* **13**, 8–17 (2015).

2. Diamant, A., Chatterjee, A., Vallières, M., Shenouda, G. & Seuntjens, J. Deep learning in head & neck cancer outcome prediction. *Sci. reports* **9**, 1–10 (2019).

3. Zhu, W., Xie, L., Han, J. & Guo, X. The application of deep learning in cancer prognosis prediction. *Cancers* **12**, 603 (2020).

4. Fan, K. *et al.* A machine learning assisted, label-free, non-invasive approach for somatic reprogramming in induced pluripotent stem cell colony formation detection and prediction. *Sci. reports* **7**, 1–9 (2017).

5. Coronnello, C. & Francipane, M. G. Moving towards induced pluripotent stem cell-based therapies with artificial intelligence and machine learning. *Stem cell reviews reports* **1–11** (2021).

6. Azuaje, F. Artificial intelligence for precision oncology: beyond patient stratification. *NPJ precision oncology* **3**, 1–5 (2019).

7. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S. & Sidow, A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nat. communications* **9**, 14 (2018).

8. Huang, S., Yang, J., Fong, S. & Zhao, Q. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer letters* **471**, 61–71 (2020).

9. Klughammer, J. *et al.* The dna methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat. medicine* **24**, 1611–1624 (2018).

10. Yilmazer, A., de Lázaro, I. & Taheri, H. Reprogramming cancer cells: a novel approach for cancer therapy or a tool for disease-modeling? *Cancer Lett.* **369**, 1–8 (2015).

11. Raab, S., Klingenstein, M., Liebau, S. & Linta, L. A comparative view on human somatic cell sources for ipsc generation. *Stem cells international* **2014** (2014).

12. Scudellari, M. *et al.* How ips cells changed the world. *Nature* **534**, 310–312 (2016).

13. Postovit, L.-M. *et al.* Human embryonic stem cell microenvironment suppresses the tumorigenic phenotype of aggressive cancer cells. *Proc. Natl. Acad. Sci.* **105**, 4329–4334 (2008).

14. Câmara, D. A. D., Mambelli, L. I., Porcacchia, A. S. & Kerkis, I. Advances and challenges on cancer cells reprogramming using induced pluripotent stem cells technologies. *J. Cancer* **7**, 2296 (2016).

15. Lim, K. L. *et al.* Reprogramming cancer cells: overview & current progress. *Expert. opinion on biological therapy* **16**, 941–951 (2016).

16. Pan, X.-Y. *et al.* Application of cancer cell reprogramming technology to human cancer research. *Anticancer. Res.* **37**, 3367–3377 (2017).

17. Wuputra, K. *et al.* Cancer cell reprogramming to identify the genes competent for generating liver cancer stem cells. *Inflamm. Regen.* **37**, 1–7 (2017).

18. Suvà, M. L., Riggi, N. & Bernstein, B. E. Epigenetic reprogramming in cancer. *Science* **339**, 1567–1570 (2013).

19. Aevermann, B. *et al.* A machine learning method for the discovery of minimum marker gene combinations for cell type identification from single-cell rna sequencing. *Genome research* **31**, 1767–1780 (2021).

20. Tian, T., Wan, J., Song, Q. & Wei, Z. Clustering single-cell rna-seq data with a model-based deep learning approach. *Nat. Mach. Intell.* **1**, 191–198 (2019).

21. Yan, R., Fan, C., Yin, Z., Wang, T. & Chen, X. Potential applications of deep learning in single-cell rna sequencing analysis for cell therapy and regenerative medicine. *Stem Cells* **39**, 511–521 (2021).

22. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology* **20**, 1–19 (2019).

23. Michielsen, L., Reinders, M. J. & Mahfouz, A. Hierarchical progressive learning of cell identities in single-cell data. *Nat. communications* **12**, 1–12 (2021).

24. Hawkins, D. M. The problem of overfitting. *J. chemical information computer sciences* **44**, 1–12 (2004).

25. Culos, A. *et al.* Integration of mechanistic immunological knowledge into a machine learning pipeline improves predictions. *Nat. machine intelligence* **2**, 619–628 (2020).

26. Kimmel, J. C. & Kelley, D. R. Semisupervised adversarial neural networks for single-cell classification. *Genome research* **31**, 1781–1793 (2021).
27. Dürr, O. & Sick, B. Single-cell phenotype classification using deep convolutional neural networks. *J. biomolecular screening* **21**, 998–1003 (2016).

28. Peng, J., Wang, X. & Shang, X. Combining gene ontology with deep neural networks to enhance the clustering of single cell rna-seq data. *BMC bioinformatics* **20**, 1–12 (2019).

29. Castro-Pérez, E. et al. Melanoma progression inhibits pluripotency and differentiation of melanoma-derived ipscs produces cells with neural-like mixed dysplastic phenotype. *Stem cell reports* **13**, 177–192 (2019).

30. Friedli, M. et al. Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome research* **24**, 1251–1259 (2014).

31. Tanaka, Y. et al. Transcriptome signature and regulation in human somatic cell reprogramming. *Stem cell reports* **4**, 1125–1139 (2015).

32. Yoshimura, Y. & Okuzaki, D. Error-free and error-prone dna repair gene expression data through reprogramming and passage in human ips cells. *Data brief* **29**, 105228 (2020).

33. Shimada, M., Tsukada, K., Kagawa, N. & Matsumoto, Y. Reprogramming and differentiation-dependent transcriptional alteration of dna damage response and apoptosis genes in human induced pluripotent stem cells. *J. Radiat. Res.* **60**, 719–728 (2019).

34. Hu, K. A piano (proper, insufficient, aberrant, and no reprogramming) response to the yamanaka factors in the initial stages of human ipsc reprogramming. *Int. journal molecular sciences* **21**, 3229 (2020).

35. Cevallos, R. R., Edwards, Y. J., Parant, J. M., Yoder, B. K. & Hu, K. Human transcription factors responsible to initial reprogramming predominantly undergo legitimate reprogramming during fibroblast conversion to ipscs. *Sci. reports* **10**, 1–11 (2020).

36. Rand, T. A. et al. Myc releases early reprogrammed human cells from proliferation pause via retinoblastoma protein inhibition. *Cell reports* **23**, 361–375 (2018).

37. Chao, M. P. et al. Human aml-ipscs reacquire leukemic properties after differentiation and model clonal variation of disease. *Cell stem cell* **20**, 329–344 (2017).

38. Kong, Y. et al. “reprogram enablement” as an assay for identifying early oncogenic pathways by their ability to allow neoplastic cells to reacquire an epiblast state. *Stem cell reports* **15**, 761–775 (2020).

39. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* **21**, 1–13 (2020).

40. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. educational psychology* **24**, 417 (1933).

41. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. machine learning research* **9** (2008).

42. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

43. Bazi, Z. et al. Rn7sk small nuclear rna is involved in neuronal differentiation. *J. Cell. Biochem.* **119**, 3174–3182 (2018).

44. Musavi, M. et al. Rn7sk small nuclear rna is involved in cellular senescence. *J. Cell. Physiol.* **234**, 14234–14245 (2019).

45. Studniarek, C. et al. The 7sk/p-tefb snrnp controls ultraviolet radiation-induced transcriptional reprogramming. *Cell Reports* **35**, 108965 (2021).

46. Eilebrecht, S. et al. 7sk small nuclear rna directly affects hmgal function in transcription regulation. *Nucleic Acids Res.* **39**, 2057–2072 (2011).

47. Nabet, B. Y. et al. Exosome rna unshielding couples stromal activation to pattern recognition receptor signaling in cancer. *Cell* **170**, 352–366 (2017).

48. Palombo, F. et al. The relevance of mitochondrial dna variants fluctuation during reprogramming and neuronal differentiation of human ipscs. *Stem cell reports* **16**, 1953–1967 (2021).

49. Pertea, M. et al. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nat. biotechnology* **33**, 290–295 (2015).

50. Kim, D., Langmead, B. & Salzberg, S. L. Hisat: a fast spliced aligner with low memory requirements. *Nat. methods* **12**, 357–360 (2015).

LaTeX formats citations and references automatically using the bibliography records in your .bib file, which you can edit via the project menu. Use the cite command for an inline citation, e.g.\cite{Dürr2016}.

For data citations of datasets uploaded to e.g. figshare, please use the howpublishedoption in the bib entry to specify the
platform and the link, as in the Hao:gidmaps:2014example in the sample bibliography file.

**Acknowledgements**

A.Y. would like to acknowledge the Turkish Academy of Sciences for financial support under the young investigator programme (GEBIP2018).
Author contributions statement
G.S. conceived the machine learning experiment(s) and analyzed the results, M.T.E. conducted the integration/registration of different datasets, C.G., L.G.D. and A.Y.A find and collected the relevant datasets. All authors reviewed the manuscript.

Additional information
The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.