Review Article | Data Science

A literature review: machine learning-based stem cell investigation

Jinat Ara[1], **Tanzila Khatun**[2]

[1]Department of Electrical Engineering and Information Systems, University of Pannonia, Veszprem, Hungary; [2]Department of Biochemistry and Biotechnology, Independent University of Bangladesh (IUB), Dhaka, Bangladesh

Contributions: (I) Conception and design: J Ara; (II) Administrative support: None; (III) Provision of study materials or patients: None; (IV) Collection and assembly of data: Both authors; (V) Data analysis and interpretation: Both authors; (VI) Manuscript writing: Both authors; (VII) Final approval of manuscript: Both authors.

Correspondence to: Jinat Ara, MSc. Department of Electrical Engineering and Information Systems, University of Pannonia, Egyetem u. 10, Veszprem 8200, Hungary. Email: jinat.ara@mik.uni-pannon.hu.

Background and Objective: Stem cell (SC) is a crucial factor of the human organ that is significantly important for clinical solutions. However, consideration of SC in the therapeutic or disease classification process is complex in terms of accurate classification and prediction. To overcome this issue, Machine learning (ML) is the most effective technique that is frequently used in cell-based clinical applications for diagnosis, treatment, and disease identification. Recently it has been implemented for SC observation which is a crucial factor for clinical solutions. Thus, the objective of this review work is to represent the effectiveness of ML techniques for SC observation from clinical perspectives with current challenges and future direction for further improvement.

Methods: In this study, we conducted a short review of ML-based applications in SCs investigation and classification for the improvement of clinical solutions. We explored studies from five scientific databases (Web of Science, Google Scholar, Scopus, ScienceDirect, and PubMed) with several keywords related to the objective of our research study. After primary and secondary screening, 15 articles were utilized for this research study and summarized the observation results in terms of ten aspects (year of publication, focused area, objective, experimented datasets, selected ML classifiers, experimental procedure, classification parameter, overall performance in terms of accuracy, advancements, and limitations) with their current limitations and future improvement directions.

Key Content and Findings: The majority of the existing literature review works are limited to focusing on specific SC-based investigation, limited evaluation attributes, and lack of challenges and future improvement suggestions. Also, most of the review work didn't consider the investigation of the effectiveness of the ML technique in SC biology. Therefore, in this paper, we investigate existing literature related to the development of clinical solutions considering ML techniques, in the area of SC and cell culture processes and highlight current challenges and future directions.

Conclusions: The majority of studies focused on the disease identification process and implemented the convolutional neural network and support vector machine techniques. The prime limitations of the investigated studies are related to the focused area, investigated SCs, the small number of experimental datasets, and validation techniques. None of the studies provided complete evidence to determine an optimal ML technique for SC to build classification or predictive models. Therefore, further concern is required to develop and improve the developed solutions including other ML techniques, large datasets, and advanced evaluation processes.

Keywords: Machine learning (ML); stem cell (SC); diagnosis; cell classification; computational cell biology

Submitted Nov 27, 2023. Accepted for publication Jan 08, 2024. Published online Mar 27, 2024.
doi: 10.21037/atm-23-1937
View this article at: https://dx.doi.org/10.21037/atm-23-1937

Introduction

Bioinformatics is a field of multidisciplinary that contains the application of computational techniques for analyzing and explaining many biological data which include genomics, proteogenomic, proteomic, and many-omic data (1). In the field of biology, machine learning (ML) algorithms can play a crucial role in handling and extracting patterns from the major and complex datasets generated in bioinformatics research (2). Different predictive models for various biological processes can be created using ML algorithms for predicting protein structure, function, or the likelihood of a genetic mutation causing disease. Among several prominent techniques of ML, pattern recognition, classification, clustering, and feature selection are prominent in drug production, and diagnosis and prognosis of disease (3). The combination of ML and bioinformatics provides a powerful approach to dealing with the complexities of biological data and increasing discoveries in the life science field. For example, integrating data from various sources, such as genomics, transcriptomics, proteomics, and epigenomics, will be essential for a comprehensive understanding of cell biology (4). ML can help in analyzing and integrating large-scale multi-omics datasets to uncover complex relationships and regulatory networks. Also, Single-cell technologies enable the study of individual cells, providing a higher resolution of cellular heterogeneity. ML algorithms can be applied to analyze single-cell data, identify cell subpopulations, and understand dynamic changes in gene expression, epigenetics, and cell fate decisions (5). Understanding the factors influencing differentiation outcomes and predicting cell fate can guide experimental design and optimization of differentiation protocols. To derive meaningful insights from biological datasets, researchers and practitioners frequently use a combination of domain expertise, statistical methods, and ML techniques (6). Additionally, ML models can be used for toxicity prediction, enhancing the safety assessment of pharmaceutical compounds during the early stages of development (7).

Therefore, ML is considered the most prominent technology in the recent era to automatically learn patterns of data and improve the performance of prediction or classification following several statistical approaches (8). ML is implemented in a wide array of domains including cell biology for its ability to automate decision-making, classify objects, and predict future perspectives (9). Referring to the cell biology and clinical perspective, Stem cells (SCs) are one of the most important factors of the human body that contribute to several aspects including disease recovery, growth, reproductive system, etc. (10). SCs are known as the undifferentiated cells that can be found in several organs in the human body at the embryonic, fetal, and adult stages of life to contribute to building tissues and organs (11). SC research holds immense potential for revolutionizing medicine and understanding biological processes. Its regenerative capacity allows researchers to replace damaged or malfunctioning cells, tissues, or organs, offering new therapeutic approaches for conditions like heart disease, diabetes, and neurodegenerative disorders (5). SCs can also be differentiated into specific cell types, allowing scientists to create models of diseases in the laboratory, study the mechanisms of various diseases, screen potential drug candidates, and understand the underlying causes of genetic disorders (12).

As SCs have a key contribution to the human body and can be found in unlimited sources with differentiation potential, it considered potential agents for several purposes such as therapeutic purposes, disease identification, and prediction, cell culture process, cell quality identification, etc. A wide array of advancements has been found in the combination of ML and SC research. For example, ML can accelerate drug discovery by predicting potential drug candidates and assessing their effects on SCs. ML, particularly computer vision algorithms, can assist in the analysis of imaging data generated in SC research. This includes tracking cell behavior, morphological analysis, and identifying patterns in microscopy images, aiding in the characterization of SC differentiation and function (13). ML models can be developed to predict and model SC differentiation trajectories and fate decisions. Deep learning techniques, such as neural networks, can be applied to analyze complex and high-dimensional biological data. These methods have shown success in image analysis, feature extraction, and pattern recognition, making them valuable for deciphering intricate aspects of SC biology (14). Transfer learning, leveraging pre-trained models on large datasets, can be beneficial in SC research. Pre-trained models can be fine-tuned on specific SC datasets, facilitating improved performance with limited data and resources. Enhancing the interpretability of ML models is crucial in the context of SC research. Developing models that provide insights into the underlying biological mechanisms and decisions can aid researchers in generating hypotheses and designing targeted experiments.

Considering the advantage of ML methods and the

importance of SCs, several past research studies have implemented ML-based SC observation to contribute to clinical solutions and cell biology and showed their contribution as effective in scientific research. Following their impressive outcome, some researchers showed their contribution to literature review work for ML-based SC research (14-16). However, the existing literature review work has limitations with a specific SC observation, focused on limited evaluation attributes, and absence of the limitation and future improvement suggestions. For example, Gupta *et al.* (15) reviewed several ML-based solutions for hematopoietic SC transplantation (HSCT). Their review found some prominent ML algorithms that perform well with significant accuracy for that particular field. However, this review is limited to considering a specific SC (hematopoietic SC) which doesn't provide any comparative result about the effectiveness of the ML algorithms on other kinds of SC classification. Similarly, Coronnello and Francipane (16) reviewed several clinical solutions that are built with artificial intelligence (AI) and ML techniques for pluripotent SC (iPSC) classification. Pluripotent SCs (iPSC) are significant for several therapeutic solutions, however, considering single SCs for review is not much helpful way to identify the effectiveness of AI and ML techniques in cell culture or cell biology. However, in another study, Kusumoto and Yuasa (15) reviewed the significance of convolutional neural network (CNN) in cell biology including the cell culture process. Though this particular review work has potential, it is limited to providing the discussion for a single ML technique which raises demand for future study considering multiple ML algorithms to represent its competence in cell biology. Therefore, in this paper, we aim to present a short review of ML-based SC investigation in terms of their publication year, focused area, objectives, experimented data, ML classifiers, experimental procedure, overall performance, classification parameter, advancement and limitation to limit the shortcomings of the past review work.

The objective of this review work is to better understand the effectiveness of ML-based applications in SC observation from clinical perspectives and cell culture processes and provide the research direction for future data analytics techniques in the field of ML and SCs. In this review work, we reviewed recently published studies that implemented ML techniques for SC observation. However, as ML is a complex data analysis method, some challenges related to appropriate ML method selection and configuration are critical. Therefore, to validate the effectiveness of the developed methods, we critically observed several challenges of the selected studies to provide a straight direction about how to overcome those challenges in future implementation.

As, in this review, we summarized the current knowledge, associated challenges, and future perspectives related to implementing ML techniques in SC observation for clinical solutions or medicine, thus this review might be beneficial for a wide group of people including technological experts, clinical practitioners, healthcare professionals, and other related authorities. This paper is structured as follows: Section II presents the research methodology to describe how research has been conducted including literature searching, literature selection, and observation. Section III lists several aspects that have been observed in the investigation and summarizes the investigation result. Section IV represents a detailed discussion of the selected papers according to the investigation results addressing challenges with future suggestions. Section V concludes the paper with a structured conclusion. We present this article in accordance with the Narrative Review reporting checklist (available at https://atm.amegroups.com/article/view/10.21037/atm-23-1937/rc).

Research methodology

To conduct this literature review, first, we searched studies in five databases using our chosen keywords. The selected five scientific databases are Web of Science, Google Scholar, Scopus, ScienceDirect, and PubMed. The literature search was conducted to identify literature that focused on several ML techniques for the classification or investigation of various types of SCs found in various organs and how those techniques can be used for various areas such as treatment process, cell culture process, diseases identification, etc. in the future.

The search period for literature selection was set from January 2018 to 2021 using the selected searching keywords such as "ML", "stem cell", "pluripotency", "teratoma formation assay", "stem cell derivation", "growth media", and "tissue transplantation". A total of 42 studies were identified initially from database searches and other sources.

Following the collection of literature, pre-processing was carried out. The pre-processing was performed through inclusion and exclusion criteria:

The inclusion criteria are: (I) papers must be written in English; (II) the paper must be related to two prime terms (ML and SC); (III) the paper should be journal or

Table 1 The search strategy summary

Items	Specification
Date of search	30 December 2021
Databases and other sources searched	Web of Science, Google Scholar, Scopus, ScienceDirect, and PubMed; ResearchGate as other sources
Search terms used	ML AND Stem cell, Pluripotency, Teratoma formation assay, Stem cell derivation, Growth media, and Tissue transplantation
Timeframe	2018–2021 (4 years of timeframe)
Inclusion criteria	Papers written in English, related to machine learning and stem cells, journal or conference proceedings, developed solutions/experimented work
Selection process	The authors primarily selected the papers based on the title and abstract. Later for final screening, the full paper was reviewed in terms of selection criteria and finally selected the papers for inducing in the review process

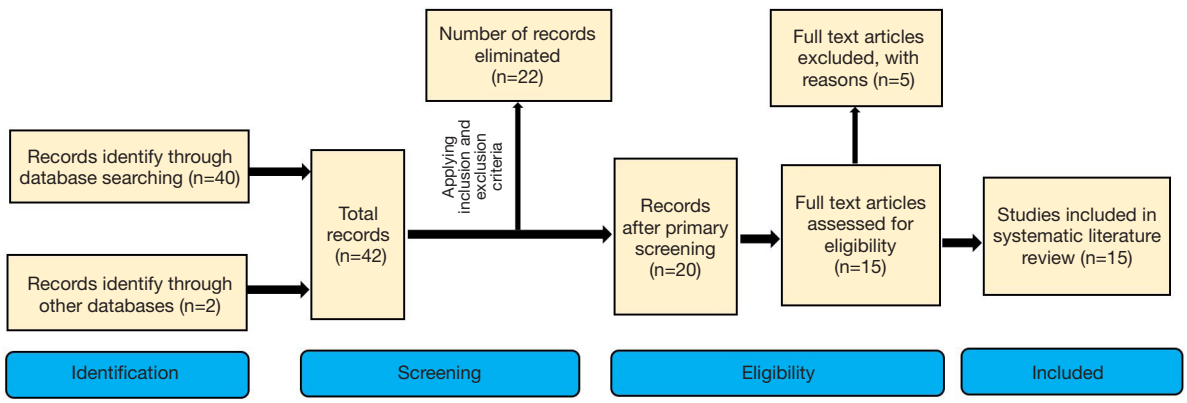

Figure 1 Overview of literature selection.

conference proceedings; (IV) the objective of the paper should be related to our research aim; and (V) papers those are developed solutions/experimented work (not any review/survey/case report).

The exclusion criteria are: (I) the paper is not written in English; (II) papers are not relevant to our research aim; (III) any book or magazine, or summary or abstract publishing; and (IV) papers not between the years of 2018 to 2021 (*Table 1*).

Along with inclusion and exclusion criteria, we applied some other criteria such as the removal of duplicate studies, papers that were not in English, and papers that were not freely accessible or downloadable. Following the initial screening, 22 papers were excluded from this study. Before making the final decision, we conducted a further screening to determine the paper's relevance to the research goal. We exempted 5 papers from the total of 20 because they were not directly related to SCs or future perspectives of SCs. The majority of them were literature reviews of SCs and others were not the experimental process or scientific procedure. Finally, for this research study, we included 15 papers. The article selection process is described in detail in *Figure 1*.

Investigation results

The selected 15 studies have investigated in terms of ten aspects: publication year, focused area, objective, experimented datasets, selected ML classifiers, experimental procedure, classification parameters, overall performance in terms of classification accuracy, their advancements, and limitations (as listed in *Table 2*). The investigation result depicts that the selected papers can be classified into five major areas: (I) cell culture quality detection; (II) diagnosis purpose; (III) cell behavior analysis; (IV) disease identification; and (V) cell classification.

Table 2 Summary of reviewed studies

References	Year	Focused area	Objective	Datasets	Algorithms/ classifiers (ML)	Methodology/procedure	Classification parameter	Classification accuracy	Advancement	Limitations
Piotrowski *et al.* (17)	2021	Cell culture quality detection for process control or cultivation	hiPSC state classification using ML methods	Life & Brain GmbH dataset (not publicly available)	U-Net deep-learning architecture	Model training; training data-set exploration; image acquisition with automated high-speed microscopy; semi-automatic segmentation chain for rapid generation of weak training data; model evaluation; data augmentation	Dice-coefficient (F1 score), IoU, and variability estimation	95.7% accuracy	• Capable of providing sufficient multi-class segmentation results for all important cell type labels occurring during hiPSC cultivation • Good segmentation performance [F1 score of 0.753 (IoU 0.777)] • Real microscope images and thus increase recognition stability • Applicable for large images • No specific thresholds are required to classify each image	• Require real-life images • Small dataset (40 images) • Single ML technique
Orita *et al.* (18)	2019	Cell culture quality detection	Classification of hiPSC-CMs	624 images were experimentally collected	CNN	Dataset collection; data splitting; perform training and testing using VGG16 architecture based on CNN; model evaluation through accuracy, precision, recall, and F1 score	Accuracy, precision, recall, and F1 score	89.7% accuracy	• Fast processing time • Has significant performance	• It requires multiple sources to train the model • A limited number of layers have been chosen for training • Not validated the model with open-source datasets • A single ML algorithm has been implemented
Waisman *et al.* (19)	2019	Cell culture quality detection	Classification of PSC	2,120 images were experimentally collected	CNN	Cells and differentiation protocols; cell imaging and image processing; cell staining and analysis; real-time PCR; CNN networks and training; colony morphological analysis	Precision, recall, and F1 score	>99% accuracy	• Can detect morphological changes in detail • Able to provide continuous, automatic, real-time detection • Comparatively fast and high accuracy	• Not enough evidence about the effectiveness of image augmentation • External validation might change the accuracy • Single ML technique is implemented
Schaub *et al.* (20)	2020	Diagnosis purpose	Predict tissue function and cellular donor using iPSC-RPE	QBAM mages	MLP, L-SVM, RF, PLSR, RR	Image processing; feature extraction; model training; model prediction; statistical analysis	R^2 values, CIs, and Kolmogorov-Smirnov, F-1, and F-2 statistics and P values	76% accuracy	• Has significance in imagining technique • Effective for multi-variant disease classification • Lower risk of losing image property during segmentation	• Small dataset (15 sample) • Multiple plugins required
Imamura *et al.* (21)	2021	Diagnosis purpose (motor neuron disease)	Classification of iPSCs	5,850 images (4,500 images for training, 1,350 images for validation)	CNN	Image acquisition; motor neuron differentiation from iPSCs; data preparation; train the model; model validation	Accuracy, AUC, and ROC curve	97% accuracy	• Has potential of predictive diagnosis of ALS • No clinical data is required	• Possibilities of overfitting to laboratory-specific technical artifacts • In the future, it is recommended to include alteration of epigenetic memory by the passage of iPSCs
Pacheco and Vidal (22)	2018	Diagnosis purpose (cardiac repair)	Classification of hESC-CMs	6,940 hESC-CMs	RNN, LSTM	Configure the proposed classification architecture; train the supervised model; clustering the result in terms of quality indexing	Prediction accuracy and euclidean distance measurement	94.73% accuracy	• Fast in computation or processing time • Able to process a large dataset • The accuracy improvement is significant compared to the state-of-the-art-literature	• Single classifiers have been used • Perform well for labeled data, that might not be suitable for unlabeled data
Teles *et al.* (23)	2021	Diagnosis purpose (cardiac disease)	Classification of healthy and diseased CMs from human iPSCs	138 videos collected from users	t-SNE, KNN, DT, NB, SVM	Cell culture and CM differentiation; profiling; ML model development; model validation	TPR, accuracy, F1 score, MCC, ROC, and AUC curve	92% accuracy (t-SNE), 91% accuracy (KNN), 88% accuracy (DT), 85% accuracy (NB), 90% accuracy (SVM)	• Multiple algorithms were tested to show the collective effectiveness of the model • Tested with different healthy controls that were not included in the training dataset	• Only two categorized health controls have been considered • The ratio between the two healthy controls was not balanced

Table 2 *(continued)*

Table 2 *(continued)*

References	Year	Focused area	Objective	Datasets	Algorithms/ classifiers (ML)	Methodology/procedure	Classification parameter	Classification accuracy	Advancement	Limitations
Kimmel *et al.* (24)	2020	Cell behavior analysis for the regeneration process	Classify MuSCs in terms of young and aged group	Collected data from laboratory experiments (not open access)	SVM	Cell isolation; cell culture; time-lapse imaging and cell behavior analysis; paired immune histochemistry and time-lapse imaging; EdU staining; DDRTree for Pseudo timing representation; velocyto parameter identification	Accuracy (through the held-out test)	95% accuracy	• Consider multimodal experiment (RNA-seq, single cell behavior experiments, and single-cell imaging experiments)	–
Zhu *et al.* (25)	2021	Disease identification (neurons of the nervous system)	NSCs differentiation	59,287 brightfield single-cell images (80% for training and 20% for testing)	CNN	Brightfield single-cell images obtained from independent experiments; image data pre-processing; apply deep learning model; model validation through ROC-AUC/PR-AUC curve	ROC-AUC, precision, and recall	0.923 or 92.3% accuracy	• Focused on multiple objects such as motor neurons or dopaminergic neurons • Multiple networks/architecture of CNN have been implemented	• Limited dataset explanation that is not fully understandable
Zhao *et al.* (26)	2020	Disease identification (LUAD cancer)	Classification of CSC in lung adenocarcinoma	TCGA	LR	Data collection; the mRNAsi calculation and analysis of WGCNA; identification of significant module and key genes; expression prognosis analysis of the key genes; interaction of co-expression analysis of the key genes	–	–	–	–
Zhang *et al.* (27)	2020	Disease identification (lung cancer)	Classification of LUAD in CSC	TCGA, Ensemble, and GEO	Limma test; Wilcox test; Kaplan-Meier plots; WGCNA analysis technique; Pearson correlation coefficient	mRNAsi data preparation; data analysis and selection; filter key genes using weighted gene co-expression network analysis; enrichment analysis of the filtered genes	–	–	–	–
Aida *et al.* (28)	2020	Disease identification (LLC)	Identification of CSC	2,851 sets of images for training, 300 sets of images for training	CNNs	Cell culture; animals and tumor tissue preparation; image processing; training and testing; model evaluation	–	–	–	–
Li *et al.* (29)	2020	Disease identification AML	Classification of LSCs expression	31,433 genes feature set (training), 11,151 genes feature set (testing)	MCFS, IFS, SVM, RIPPER	Dataset preparation (LSC$^+$, GEO); feature extraction using MCFS and IFS; model implementation for classification (SVM and RIPPER); model validation based on sensitivity, specificity, accuracy, and the MCC	Sensitivity, specificity, accuracy, and the MCC	0.921 or 92.1% accuracy (SVM), 0.815 or 81.5% accuracy (RIPPER)	• Multiple classifiers have been incorporated that show their effectiveness • Has significant effectiveness in terms of accuracy	• Multiple plugins require • Not validate with external data
Hwang *et al.* (30)	2020	Disease identification (cardiac disease)	Identification of Ca^{2+} transient abnormality in hiPSC-CMs	200 cells (training), 54 cells (testing)	SVM	Data pre-processing (Ca^{2+} transient signal data); abnormality assessment; peak detection and variable quantification; algorithmic evaluation; training and testing; model evaluation	Accuracy, sensitivity, specificity, AUC curve	83.3% accuracy (analytical algorithm), 87.0% accuracy (SVM)	• Has significant accuracy • The feature selection process facilitates the process • Incorporate external data for validation	• Specific hardware configuration is required
Kusumoto *et al.* (31)	2018	Cell classification	Identification of endothelial cells from iPSCs	800 images were experimentally collected (600 for training and 200 for validation)	CNN	iPSC culture; endothelial cell differentiation; dataset preparation; deep neural network (LeNet) configuration; performance evaluation and model validation	F1 scores and accuracy	>0.9 accuracy	• Simple network, easy to understand • Less complex network with fewer layer	• A shallow network has been used

ML, machine learning; IoU, Intersection over union; hiPSC, human induced pluripotent stem cell; CM, cardiomyocyte; CNN, convolution neural network; PSC, pluripotent stem cell; PCR, polymerase chain reaction; iPSC-RPE, pluripotent stem cells derived from retinal pigment epithelial; QBAM, quantitative bright-field absorbance microscopy; MLP, multilayer perceptron; SVM, support vector machine; RF, random forest; PLSR, partial least squares regression; RR, ridge regression; CI, confidence interval; iPSC, induced PSC; AUC, area under the curve; ROC, receiver operating characteristic; ALS, amyotrophic lateral sclerosis; hESC, human embryonic stem cell; RNN, recurrent neural network; LSTM, long short-term memory; t-SNE, t-stochastic neighbor embedding; KNN, k-nearest neighbor; DT, decision trees; NB, naïve Bayes; TPR, true positive rate; MCC, Matthews correlation coefficient; MuSC, murine muscle stem cell; NSC, neural stem cell; PR-AUC, precision recall AUC; LUAD, lung adenocarcinoma; CSC, cancer stem cell; TCGA, The Cancer Genome Atlas; LR, logistic regression; WGCNA, weighted gene co-expression network analysis; GEO, Gene Expression Omnibus; LLC, Lewis lung cancer; AML, acute myeloid leukemia; LSC, leukemia stem cell; MCFS, Monte Carlo Feature Selection; IFS, incremental feature selection; RIPPER, Repeated Incremental Pruning to Produce Error Reduction.

Cell culture quality detection

Referring to cell culture quality detection, Piotrowski *et al.* (17) claimed that by cell culture quality detection, it is possible to contribute to personalized medicine and drug screening processes. As this process requires high-quality human induced pluripotent stem cells (hiPSCs) samples, thus automatic or image processing-based solutions of hiPSCs classification might be effective for analysis and cultivating the samples, and improving the accuracy. In that regard, they proposed an automated hiPSCs cultivation model using PHANTAST and ML algorithms. The proposed U-Net deep learning architecture can differentiate the classes in terms of single cells, differentiated cells, and dead cells. The experimentation result shows that their proposed model is capable of segmenting important parameters of hiPSCs colony forming and can differentiate between colonies, single cells, and dead cells which resulted in 95.7% accuracy. In another study, Orita *et al.* (18) claimed that for developing a high clinical predictability system, pre-clinical cardiac safety is critical. Therefore, they trained convolution neural network (CNN) model on bright-field images of cultured hiPSC-derived cardiomyocytes (hiPSC-CMs) to determine whether the qualities of cell cultures are suitable for experiments. For experimentation, a total of 624 images were trained those were collected experimentally and the characteristics of the cultures were tested through 14,556 images that were labeled as 'normal' and 'abnormal'. The experimented open-source CNN framework resulted in 89.7% accuracy. Further, Waisman *et al.* (19) discussed the limitations of several technologies to the cell culture process and advancements of ML models. In that manner, they analyzed pluripotent SCs (PSC) from microscopy images using CNN to distinguish pluripotent SCs from early differentiating cells. They differentiated mouse embryonic SCs into epiblast-like cells that were collected from several time points from the initial stimulus. The trained model recognized undifferentiated cells from differentiating cells with a high accuracy of around 99%.

Diagnosis process

Referring to the diagnosis process, Schaub *et al.* (20) claimed that in clinical translation, iPSC-based therapy is challenging due to the scarcity of non-invasive, automated, fast, and robust assays. Thus, they proposed a model incorporating quantitative bright-field absorbance microscopy (QBAM), and several ML models [multilayer perceptron (MLP), linear support vector machine (L-SVM), random forest (RF), partial least squares regression (PLSR), ridge regression (RR)] to predict noninvasively tissue function. QBAM is an automated method of recording images that are reproducible across various microscopes and 5 neural networks to predict tissue function without invasive procedures. Among the selected classifiers, they found NN and L-SVM models were the most effective model than other selected models with an accuracy of 76.4%. Imamura *et al.* (21) claimed that amyotrophic lateral sclerosis (ALS) is a motor neuron disease with a low survival rate and cannot be detected until the disease has progressed. Therefore, identification of this disease at an early phase is crucial. Considering this aspect, they proposed a deep learning-based prediction model of ALS with images of motor neurons derived from patient-iPSC to support the ALS diagnosis process. The proposed model is developed considering a CNN which experimented on 5,850 images. The model is validated through the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and compared with RF classifier. The experimented result depicts that the proposed model achieved 97% accuracy and the comparison result shows that both classifiers performed the same in the experimented dataset.

Further, Pacheco and Vidal (22) argued that human embryonic SCs (hESCs) could be promising sources for cardiomyocytes (CMs)-based applications such as cell-based cardiac repair and drug screening. Thus, they proposed a semi-supervised learning framework for classifying hESC-derived cardiomyocytes (hESC-CMs). The proposed framework is implemented based on a recurrent neural network with long short-term memory (LSTM). For supervised classification, they considered the labeled data from computational models of adult CMs, while for unsupervised classification, they used the metamorphosis distance. Their findings indicate the benefit of incorporating information from both adults and SC-derived domains in the learning scheme to obtain better results around 94.73% which indicates clear computational advantages. With the same issue, Teles *et al.* (23) discussed Cardiomyocytes that are derived from human induced pluripotent stem (iPS) cells and could be potential for cardiac disease diagnosis systems. Therefore, they proposed an ML-based classification model to classify cardiomyocytes in terms of healthy and diseased cardiomyocytes from iPSCs. Five ML classifiers have been implemented namely t-distributed stochastic neighbor embedding (t-SNE), k-nearest neighbor (KNN), decision trees (DT), naïve Bayes (NB), and support vector machine

(SVM). The model is evaluated through accuracy and the AUC curve. The proposed model achieved promising accuracy for t-SNE, KNN, and SVM which is above 90%.

Cell behavior analysis

Focusing on cell behavior analysis, Kimmel *et al.* (24) observed murine-muscle SCs (MuSCs) in terms of young and old cell age. They classified efficient cell age using ML algorithms based on the data of cell behavior analysis and RNA velocity. The result found that activation kinetics are delayed in aged MuSCs, which may contribute to impaired SC function with age. They suggested that the dynamic changes in SCs may be a factor in the deterioration of SC function with age. Through the experiment, the authors discover that SC activation does not appear to be a continuous linear progression instead it resembles a random walk with frequent reversals.

Disease identification

Considering disease identification, Zhu *et al.* (25) discussed that early identification of neural SCs (NSC) is important to early disease identification though it is a difficult task. Therefore, they focused on the urgent need for an efficient, accurate, convenient, and less-wasting method. They proposed a model to evaluate single-cell images by using deep learning techniques to identify the differentiation of NSCs. The proposed model used a CNN to extract local image features to do the image classification. The experimentation was performed on 59,287 annotated single-cell images where 80% were used for training data and 20% were used for model testing. The experiment result depicts that the proposed model can identify very small morphological variations in cellular structures, for example, it can classify cell fate by discriminating differences between each type. Zhao *et al.* (26) and Zhang *et al.* (27) mentioned that cancer SCs (CSC) are involved in tumor metastasis, relapse, and drug resistance, which account for tumor heterogeneity. Therefore, they focused on cancer identification from the CSC classification located in the lung. They implemented a logistic regression (LR) ML algorithm on the mRNA expression of pluripotent SCs and their differentiated progeny to determine the mRNA stemness index (mRNAsi). The experimentation was performed using more than 500 LUAD cases from The Cancer Genome Atlas (TCGA) database. Also, weighted correlation network analysis was used to find the key genes associated with mRNAsi. The analysis process faced several challenges related to differential expressions, survival analysis, clinical stages, and gender in LUADs. The experiment result depicts that mRNA expression was significantly higher in LUAD cases than in the normal lung samples. With the same focus, Aida *et al.* (28) mentioned that most CSCs have an identical mark so they are extensively characterized by SC-like gene expression. They aimed to investigate the segmentation of CSCs in phase contrast imaging using conditional generative adversarial networks (CGAN). To do so, AI was trained for fluorescence images of the Nanog-green fluorescence protein. In the phase contrast image of the CSC cultures and tumor model, the AI model segmented the CSC region. The investigation result depicts that the CNN using CGAN might be useful in determining undescribed morphological characteristics in CSCs.

Furthermore, Li *et al.* (29) addressed that acute myeloid leukemia (AML) is a form of blood cancer that causes the rapid growth of immature white blood cells that can be detected in the bone marrow. However, the conventional process of detecting this disease is quite difficult and could be affected by various issues such as limited gene classification or identification. The author of this paper developed a ML approach considering feature selection and classification algorithms such as Monte Carlo Feature Selection (MCFS), Incremental Feature Selection (IFS), SVM, and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) to identify gene expression features specific to leukemia stem cells (LSCs). The experiment showed that 1,159 features (genes) were first identified, which could be an optimal solution for distinguishing LSC^+ and LSC^- cells. The experimental result predicted that SVM performs better (0.921) in classification than the RIPPER model (0.815). Also, Hwang *et al.* (30) claimed that abnormal Ca^{2+} transients are crucial for evaluating the cardiomyocyte function of our body but detection of its abnormality can be time-consuming. As a result, the authors of this paper created an analytical pipeline for the automatic assessment of Ca^{2+} transient abnormality using advanced ML methods and an analytical algorithm. They modify an existing analytical algorithm to identify Ca^{2+} transient peaks and determine peak abnormality using quantified peak characteristics. Then, the authors train a ML classifier namely SVM using human-expert evaluation to identify abnormality related to peak-level and cell-level predictive features. According to their observation, the cell-level SVM classifier is effective in assessing additional Ca^{2+} transient

signals. Thus, they implemented the cell-level SVM pipeline in R environment where they trained the model using 200 cells as training data and 54 independent cells as testing data. The experiment resulted in 88% training and 87% testing accuracy.

Cell classification

With the aim of cell classification, Kusumoto *et al.* (31) mentioned that deep learning techniques are advancing technology for some computational problems in cell biology. The advancement of this method is promising in regenerative medicine and disease modelling concerning SC classification. Thus, in this paper, they used deep learning techniques especially CNN to identify endothelial cells derived from iPSCs. Endothelial cells are prominent in disease classification at an early stage. To identify the endothelial cells, they considered several morphological information of the endothelial cells that could be identified from the phase-contrast image of iPSCs. The CNN-based model was trained and tested considering 800 images and validated through K-fold cross-validation. The performance of the model was validated through their accuracy which is around >90%. The experiment result depicts that the proposed model of this paper can identify endothelial cells based on morphology information.

Furthermore, according to the advancement of the reviewed studies, the majority of the proposed classifiers have significant performance, provide morphological information in details that are hard to detect through human observation, provide automatic real-time detection, the computation time is fast that reflects the difficulty for human perception, can focus multiple object/class at the same time and provide incorporation of multiple solutions to validate or improve the performance of classifiers. Additionally, the automatic process especially the ML algorithmic process allows several additional observation opportunities that are relatively impossible through traditional cell culture or human analysis processes such as the feature selection process to identify the important features that help to identify the actual cell status.

However, the reviewed studies have some limitations such as the majority of the applications require real-life images, in some cases some applications require multiple sources to execute the processing, few applications perform with their custom dataset as an alternative to a public dataset and most of the cases, the implemented dataset has fewer samples that can directly reduce the accuracy of the model. Another frequent issue observed is single ML model implementation. From the literature, it is emphasized that multiple ML model incorporation or integrated models can perform better than a single model. Also, only one model was found that adopted a completely new dataset as a validation dataset that did not use the sample from the same dataset that was used for model training and testing which showed the effectiveness of the model. Besides, regarding the architecture of the model, some proposed models are complex to understand and some models have used shallow networks with few numbers of layers that don't provide the actual efficiency of the model. Also, most of the proposed model didn't consider overfitting issues and some require their specific hardware configuration and multiple plugins.

From the investigation of the selected papers, we can conclude that the advancement of the ML technique is promising in most of the domains including the SC investigation process because of its several advancements such as easier to identify specific cellular features that help to predict cell function. Mostly, many of the advancements of ML techniques are not compatible with human observation, for example, fast computation or classification of a large number of samples, correct prediction or classification, etc. As ML-based applications for the cell classification process are growing day by day, some limitations and challenges might be observed in some developed applications. In the future, the developed solution should be more sophisticated and require careful observation related to effective network selection, multiple ML algorithm incorporation, less complex configuration, re-usable network development for multi-purpose use, open-source datasets consideration, increased sample sizes of the tested dataset and implementing external dataset for model validation. Careful observation of these addressed issues might help the model performance and increase the model effectiveness, especially for cell classification or cell culture process that could be a great progress in the application development of SC biology.

Discussion

According to the review, it can be concluded that ML models can predict the fate of SCs during differentiation processes. By analyzing large-scale omics data, these models can identify key molecular markers and signaling pathways associated with specific cell types, aiding in the directed differentiation of SCs into desired lineages. ML algorithms can be employed to optimize and tailor differentiation

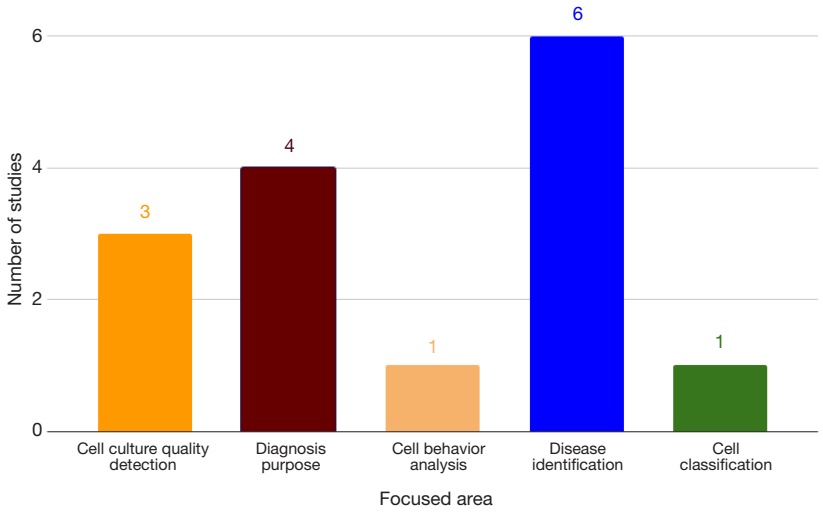

Figure 2 Distribution of studies by focused area.

protocols for generating specific cell types. By considering a variety of factors, including culture conditions, growth factors, and genetic profiles, ML models can suggest optimal conditions for efficient and reproducible differentiation.

In the context of iPSCs, ML can contribute to improving reprogramming efficiency. Models can analyze cellular reprogramming datasets to identify critical factors influencing the reprogramming process, leading to more efficient generation of iPSCs. ML techniques can assist in the characterization of SCs and ensure the quality control of cell populations. This includes the identification of pluripotency markers, assessment of cell viability, and detection of potential anomalies or abnormalities in SC cultures. Also, ML can optimize 3D culture systems, which are essential for mimicking the physiological microenvironment of cells *in vivo*. By analyzing data on cell behavior, gene expression, and extracellular matrix interactions, ML models can guide the design of 3D culture conditions for improved cell maturation and functionality. ML models can predict the maturation states of differentiated cells and their functional properties. This is particularly relevant in the context of developing cells for transplantation or drug screening assays, where the maturity of the cells is critical for their intended application.

Also, ML applications can enhance the efficiency of drug screening assays using SCs. Models can predict the potential therapeutic effects of drugs and assess their toxicity on differentiated cells, aiding in the identification of safe and effective drug candidates. ML can contribute to personalized medicine by analyzing patient-specific data, predicting responses to treatments, and modeling disease progression. This approach is valuable for studying genetic disorders, understanding patient-specific variations, and tailoring therapeutic strategies accordingly (32).

Moreover, ML algorithms can be integrated into real-time monitoring systems to provide feedback on the culture conditions and guide adjustments in response to changes in cell behavior. This dynamic feedback loop can improve the control and reproducibility of SC cultures.

According to the empirical analysis, we summarise some key aspects through *Figures 2-4*. *Figure 2* depicts the number of studies in terms of their focused areas where SCs are being used for various purposes. Among all the areas in which SCs are used, disease identification is the most frequently used purpose that is considered SCs. Diagnosis purpose and cell culture quality detection are ranked as the second and third focused areas, respectively.

Figure 3 shows the types of SCs used in the research studies that we selected for this review. According to the graph, pluripotent SCs were used the most, with six papers focused on them, and CSCs came in second with three papers focused on them. Other SCs were discovered in one paper each.

Figure 4 represents several ML classifiers implemented in the selected papers. CNN is the most frequently used classifier, considering in 6 papers, and SVM is implemented in four papers. However, the rest of the classifiers are only considered in one paper each.

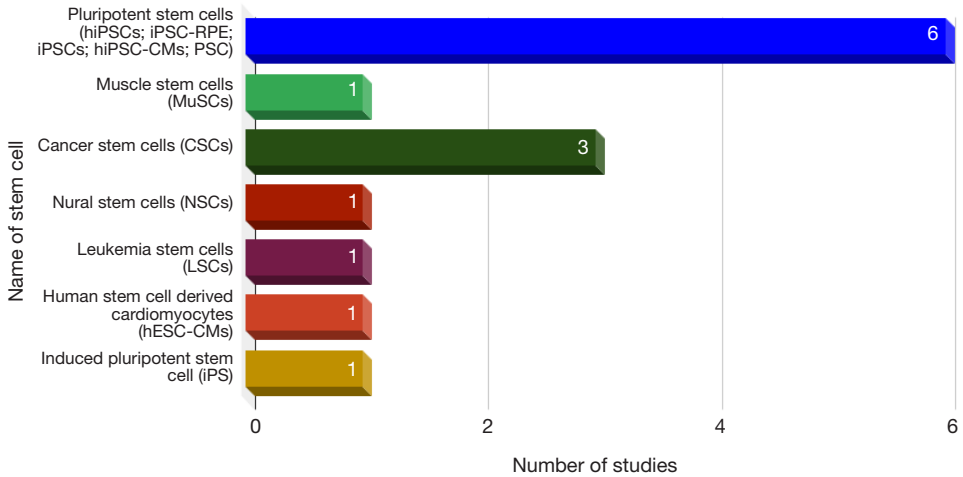

Figure 3 Distribution of studies by experimented stem cells.

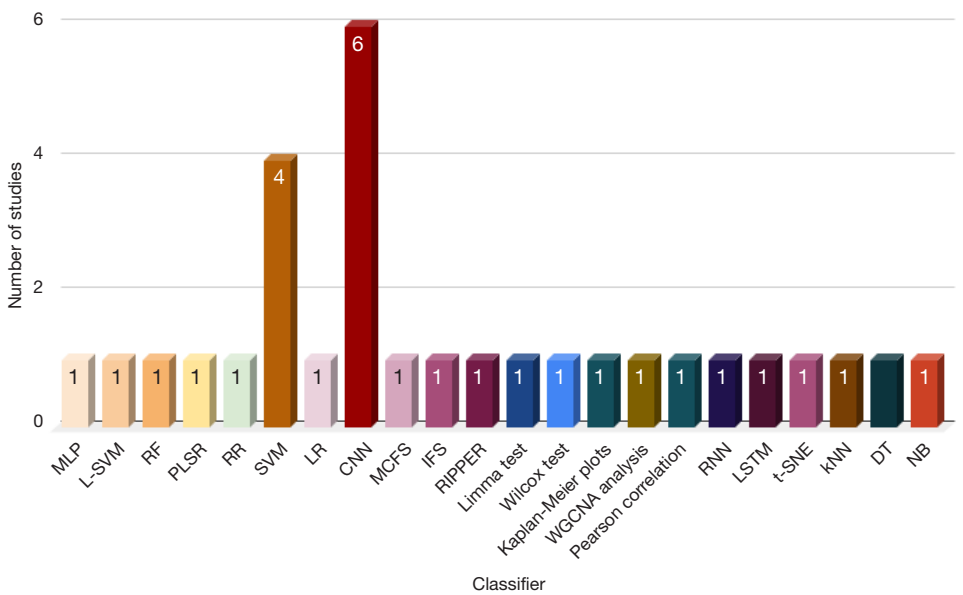

Figure 4 Distribution of studies by implemented classifiers. MLP, multilayer perceptron; SVM, support vector machine; RF, random forest; PLSR, partial least squares regression; RR, ridge regression; LR, logistic regression; CNN, convolution neural network; MCFS, monte carlo feature selection; IFS, incremental feature selection; RIPPER, Repeated Incremental Pruning to Produce Error Reduction; WGCNA, weighted gene co-expression network analysis; RNN, recurrent neural network; LSTM, long short-term memory; t-SNE, t-stochastic neighbor embedding; KNN, k-nearest neighbor; DT, decision trees; NB, naïve Bayes.

From our observation and investigation results, it could be concluded that the selected applications are not free from limitations. In terms of the investigated area, disease identification is considered frequently though other focused areas require to be considered. Considering SCs, pluripotent SC investigation has been conducted in the majority of the considered studies though in terms of identifying potential medical solutions, other SCs could also be valuable. Most of the studies did not provide robust evidence to choose that particular ML algorithm and did not argue their existing limitations and challenges. A number of studies implemented SVM and CNN though

other techniques might be beneficial to provide better accuracy. Furthermore, most of the considered studies are limited to experimenting with a small number of samples of data where a large number of data might alter the overall outcome. All the application's performances are validated through statistical observation while users and experts should be incorporated to validate the results as these applications aim to deal with the real-life complex problem for a clinical solution that is related to human life directly. However, ML could be potential in SC observation due to its interpretability. Therefore, our suggestion to the scientific community is to conduct further investigation on different areas considering various SCs in the future. Also, future studies should provide a clear explanation about the selected classifier to understand which classifier could potential for which types of problem, improve the application authenticity by validating through human incorporation, and expand the experiment data size to make the predicted model more sophisticated.

However, establishing benchmarks and standardized protocols for ML applications in SC biology will facilitate the comparison of different methods and promote the reproducibility of results. Effective collaboration between biologists and data scientists is essential for the successful application of ML in SC research. Close interaction can ensure that ML models are developed and applied in a biologically meaningful and context-aware manner. By combining expertise from ML and SC biology, researchers can leverage the strengths of both fields to propel advancements and uncover new insights in SC research. The interdisciplinary nature of these collaborations is expected to drive innovation and accelerate progress in the understanding and application of SCs. Overall, the integration of ML methods into SC biology holds tremendous potential for advancing our understanding of SC behavior, optimizing culture conditions, and facilitating the development of novel cell types for therapeutic applications and disease modeling. The synergistic collaboration between experimentalists and data scientists is key to realizing these advancements in the field.

Conclusions

In this paper, we conducted a short review of 15 studies related to ML techniques for SC observation. Day by day, the importance of ML and SC-based clinical applications is increasing in the healthcare sector. Therefore, to identify the potential of the experimented work, a detailed literature review is crucial as well as to provide a brief scenario of the effectiveness of the experimented ML techniques. As SC analysis or classification is important for many areas such as medicine screening, cell culture process, disease classification, etc. thus large number of SC investigations through human observation or the traditional process is time-consuming and there is a doubt about the complete accuracy. To overcome these challenges related to human observation or human participation in different phases of cell classification or investigation process, ML techniques are an effective and potential solution that attracted many biologists and researchers in that area. Thus, following our objective, in this paper we provide several potentials uses of ML techniques in the SC investigation process. In our reviewed studies, a promising number of studies focused on the disease identification process and implemented the CNN and SVM techniques compared to other ML techniques but were limited to a small number of experimental datasets and validation techniques. None of the studies provided complete evidence to determine an optimal ML technique for SCs to build classification or predictive models. The observation results suggested that to improve the effectiveness of the current applications, further concern is demanded including other ML techniques, large datasets, and model evaluation processes.

However, this research work is not free from limitations. As this is our initial contribution to this particular field such as SC investigation considering ML techniques, we limited our search to 4 years span and we chose only 15 studies that are related to our research objective. These two factors are aligned as the limitation of this review work that might change the addressed findings while we consider a large number of papers. Therefore, future work will be aligned with the consideration of a wide array of studies and perform a broad literature review work on this particular field to present the current scenario in a broader aspect with the comparison of existing literature review work.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://atm.amegroups.com/article/view/10.21037/atm-23-1937/rc

Peer Review File: Available at https://atm.amegroups.com/article/view/10.21037/atm-23-1937/prf

Conflicts of Interest: Both authors have completed the ICMJE uniform disclosure form (available at https://atm.amegroups.com/article/view/10.21037/atm-23-1937/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Del Sol A, Jung S. The Importance of Computational Modeling in Stem Cell Research. Trends Biotechnol 2021;39:126-36.
2. Ashraf M, Khalilitousi M, Laksman Z. Applying Machine Learning to Stem Cell Culture and Differentiation. Curr Protoc 2021;1:e261.
3. Wang S, Wang Q, Fan B, et al. Machine learning-based screening of the diagnostic genes and their relationship with immune-cell infiltration in patients with lung adenocarcinoma. J Thorac Dis 2022;14:699-711.
4. Wang Y, He Y, Duan X, et al. Construction of diagnostic and prognostic models based on gene signatures of nasopharyngeal carcinoma by machine learning methods. Transl Cancer Res 2023;12:1254-69.
5. Yan R, Fan C, Yin Z, et al. Potential applications of deep learning in single-cell RNA sequencing analysis for cell therapy and regenerative medicine. Stem Cells 2021;39:511-21.
6. Kusumoto D, Yuasa S, Fukuda K. Induced Pluripotent Stem Cell-Based Drug Screening by Use of Artificial Intelligence. Pharmaceuticals (Basel) 2022;15:562.
7. Kumar D, Baligar P, Srivastav R, et al. Stem Cell Based Preclinical Drug Development and Toxicity Prediction. Curr Pharm Des 2021;27:2237-51.
8. Shamout F, Zhu T, Clifton DA. Machine Learning for Clinical Outcome Prediction. IEEE Rev Biomed Eng 2021;14:116-26.
9. Allugunti VR. A machine learning model for skin disease classification using convolution neural network. Int J Comput Programming Database Manage 2022;3:141-7.
10. Volarevic V, Bojic S, Nurkovic J, et al. Stem cells as new agents for the treatment of infertility: current and future perspectives and challenges. Biomed Res Int 2014;2014:507234.
11. Zakrzewski W, Dobrzyński M, Szymonowicz M, et al. Stem cells: past, present, and future. Stem Cell Res Ther 2019;10:68.
12. Fidanza A, Stumpf PS, Ramachandran P, et al. Single-cell analyses and machine learning define hematopoietic progenitor and HSC-like cells derived from human PSCs. Blood 2020;136:2893-904.
13. Lin Y, Tang M, Liu Y, et al. A narrative review on machine learning in diagnosis and prognosis prediction for tongue squamous cell carcinoma. Transl Cancer Res 2022;11:4409-15.
14. Kusumoto D, Yuasa S. The application of convolutional neural network to stem cell biology. Inflamm Regen 2019;39:14.
15. Gupta V, Braun TM, Chowdhury M, et al. A Systematic Review of Machine Learning Techniques in Hematopoietic Stem Cell Transplantation (HSCT). Sensors (Basel) 2020;20:6100.
16. Coronnello C, Francipane MG. Moving Towards Induced Pluripotent Stem Cell-based Therapies with Artificial Intelligence and Machine Learning. Stem Cell Rev Rep 2022;18:559-69.
17. Piotrowski T, Rippel O, Elanzew A, et al. Deep-learning-based multi-class segmentation for automated, non-invasive routine assessment of human pluripotent stem cell culture status. Comput Biol Med 2021;129:104172.
18. Orita K, Sawada K, Koyama R, et al. Deep learning-based quality control of cultured human-induced pluripotent stem cell-derived cardiomyocytes. J Pharmacol Sci 2019;140:313-6.
19. Waisman A, La Greca A, Möbbs AM, et al. Deep Learning Neural Networks Highly Predict Very Early Onset of Pluripotent Stem Cell Differentiation. Stem Cell Reports 2019;12:845-59.
20. Schaub NJ, Hotaling NA, Manescu P, et al. Deep learning predicts function of live retinal pigment epithelium from quantitative microscopy. J Clin Invest 2020;130:1010-23.

21. Imamura K, Yada Y, Izumi Y, et al. Prediction Model of Amyotrophic Lateral Sclerosis by Deep Learning with Patient Induced Pluripotent Stem Cells. Ann Neurol 2021;89:1226-33.
22. Pacheco C, Vidal R. Recurrent neural networks for classifying human embryonic stem cell-derived cardiomyocytes. In: Frangi A, Schnabel J, Davatzikos C, et al. editors. Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, Springer; 2018:581-9.
23. Teles D, Kim Y, Ronaldson-Bouchard K, et al. Machine Learning Techniques to Classify Healthy and Diseased Cardiomyocytes by Contractility Profile. ACS Biomater Sci Eng 2021;7:3043-52.
24. Kimmel JC, Hwang AB, Scaramozza A, et al. Aging induces aberrant state transition kinetics in murine muscle stem cells. Development 2020;147:dev183855.
25. Zhu Y, Huang R, Wu Z, et al. Deep learning-based predictive identification of neural stem cell differentiation. Nat Commun 2021;12:2614.
26. Zhao M, Chen Z, Zheng Y, et al. Identification of cancer stem cell-related biomarkers in lung adenocarcinoma by stemness index and weighted correlation network analysis. J Cancer Res Clin Oncol 2020;146:1463-72.
27. Zhang Y, Tseng JT, Lien IC, et al. mRNAsi Index: Machine Learning in Mining Lung Adenocarcinoma Stem Cell Biomarkers. Genes (Basel) 2020;11:257.
28. Aida S, Okugawa J, Fujisaka S, et al. Deep Learning of Cancer Stem Cell Morphology Using Conditional Generative Adversarial Networks. Biomolecules 2020;10:931.
29. Li J, Lu L, Zhang YH, et al. Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. Cancer Gene Ther 2020;27:56-69.
30. Hwang H, Liu R, Maxwell JT, et al. Machine learning identifies abnormal Ca(2+) transients in human induced pluripotent stem cell-derived cardiomyocytes. Sci Rep 2020;10:16977.
31. Kusumoto D, Lachmann M, Kunihiro T, et al. Automated Deep Learning-Based System to Identify Endothelial Cells Derived from Induced Pluripotent Stem Cells. Stem Cell Reports 2018;10:1687-95.
32. Ji XL, Ma L, Zhou WH, et al. Narrative review of stem cell therapy for ischemic brain injury. Transl Pediatr 2021;10:435-45.

Cite this article as: Ara J, Khatun T. A literature review: machine learning-based stem cell investigation. Ann Transl Med 2024. doi: 10.21037/atm-23-1937