Peer Review File

Article information: https://dx.doi.org/10.21037/atm-23-1937

Reviewer A Comments:

The paper titled "A Short Review: Machine Learning based Stem Cell (SC) Investigation" is interesting. The majority of studies focused on the disease identification process and implemented the CNN and SVM techniques. The prime limitations of the investigated studies are related to the focused area, investigated stem cells, the 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. However, there are several minor issues that if addressed would significantly improve the manuscript.

Comment 1: It is recommended to provide a detailed introduction to the key paradigms of machine learning, with a focus on equipping stem cell biologists with the necessary understanding abilities.

Reply 1: Thank you for your suggestion. We have modified the introduction section; you can find it in the updated version of the paper (lines: 67-88 and 106-121). The added information is:

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, 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 machine learning 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 machine learning 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]. Machine learning 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. Machine learning 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 machine learning techniques [6]. Additionally, machine learning models can be used for toxicity prediction, enhancing the safety assessment of pharmaceutical compounds during the early stages of development [7].

A wide array of advancements has been found in the combination of machine learning and stem cell research. For example, Machine learning can accelerate drug discovery by predicting potential drug candidates and assessing their effects on stem cells. Machine learning, particularly computer vision algorithms, can assist in the analysis of imaging data generated in stem cell research. This includes tracking cell behavior, morphological analysis, and identifying patterns in microscopy images, aiding in the characterization of stem cell differentiation and function [14]. Machine learning models can be developed to predict and model stem cell 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 stem cell biology [15]. Transfer learning, leveraging pre-trained models on large datasets, can be beneficial in stem cell research. Pre-trained models can be fine-tuned on specific stem cell datasets, facilitating improved performance with limited data and resources. Enhancing the interpretability of machine learning models is crucial in the context of stem cell research. Developing models that provide insights into the underlying biological mechanisms and decisions can aid researchers in generating hypotheses and designing targeted experiments.

Comment 2: Can machine learning be combined with bioinformatics analysis to construct new classifiers? What is the significance of stem cell research? Suggest adding relevant content.

Reply 2: Thank you for your question. In this review work, papers have reviewed all of them are ML-based applications that used ML classifiers to analyze stem cells to facilitate the human observation process that is directly related to bioinformatics. As our focus of this work is to identify the significance of the ML classifiers for stem cell research, we added more information related to ML classifiers and how they facilitated the traditional process of stem cell investigation such as human observation. Therefore, from our analysis, it is clear that the ML classifier has a significant contribution to facilitating such a process that also refers to the significance of stem cell research. We hope for the acceptance of our answer.

Comment 3: What disciplines and directions are expected to accelerate the progress of machine learning methods applied to stem cell biology? Suggest adding relevant content.

Reply 3: Thank you for your question. In the investigation result section (lines: 348-361), we have added some information related to the limitations and challenges that we found in the reviewed papers that related to the machine learning methods specifically for stem cell classification. The addressed aspects can accelerate the progress in future development. The added information is:

From the investigation of the selected papers, we can conclude that the advancement of the machine learning technique is promising in most of the domains including the stem cell

investigation process because of its several advancements such as is easier to identify specific cellular features that help to predict cell function. Mostly, many of the advancements of machine learning 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 machine learning-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 stem cell biology.

Comment 4: There are many databases. Why did the author only select five scientific databases in this study for searching? Please explain the reason.

Reply 4: Thank you for your question. This short review is our primary research on this particular field and according to the findings of this research work, we intend to conduct a detailed review work in the future. Therefore, we consider only a few databases that will be extended in the future for broad review. We hope for the acceptance of the answer.

Comment 5: Please discuss the potential applications in stem cell biology, including the development of novel cell types, and improving model maturation.

Reply 5: Thank you for your question. We added some information in the introduction section about the recent contribution of stem cell biology that you can find in the updated version of the paper in lines: 67-88 and 106-121. Also, some relevant information has been added in the discussion section in lines: 378-405.

Comment 6: The introduction part of this paper is not comprehensive enough, and the similar papers have not been cited, such as "Construction of diagnostic and prognostic models based on gene signatures of nasopharyngeal carcinoma by machine learning methods, Transl Cancer Res, PMID: 37304552". It is recommended to quote the article.

Reply 6: Thank you. We have modified the introduction and added the recommended article. You can find the added information in the updated version of the paper.

Comment 7: What are the biggest strengths and weaknesses of this research model? What is the biggest problem faced? Suggest adding relevant content.

Reply 7: Thank you for your raised question. We have added some limitations and strengths of the reviewed paper in the updated version of the paper in lines: 327-347. The added information is:

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 machine learning 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.

As the limitation of this review work, we added information to the updated version of the paper in the conclusion section (lines: 476-480). The added information is:

However, this research work is not free from limitations. As this is our initial contribution to this particular field such as Stem Cell 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.

Reviewer B Comments:

Comment 1: First, the title could be more detailed such as a narrative review and " clinical perspectives and cell culture processes".

Reply 1: Thank you for your suggestion. According to the aim of the paper we entitled that title for the paper which seems suitable from our perspective. However, clinical perspectives and cell culture processes are a part of the investigated papers but in this review, our aim is not to identify or classify anything related to these keywords. We hope for your understanding.

Comment 2: Second, the abstract needs to indicate the clinical needs for this review focus and the questions to be answered in this review. The methods need to describe the inclusion criteria for eligible studies and how the data were qualitatively analyzed. The conclusion needs to summarize the limitations and knowledge gaps of the studies reviewed.

Reply 2: Thank you for your valuable suggestions. In the abstract, we modified and added some information focusing on the need for the review from a clinical perspective. You can find the added information in the updated version of the paper (lines: 19-28). We hope for your acceptance. The added information is:

Also, most of the review work didn't consider the investigation of the effectiveness of the machine learning technique in cell biology. Thus, the objective of this review work is to investigate the effectiveness of ML techniques, especially for stem cell observation and classification. As there are a wide array of clinical solutions have been developed for stem cell investigation by adopting ML algorithms, thus understanding the advancement of ML techniques associated with the challenges and vulnerabilities of the developed solutions could improve the effectiveness of the developed clinical solutions. Thus, we aim to investigate existing literature related to the development of clinical solutions considering ML techniques, on the area of stem cell and cell culture processes. Additionally, this review highlighted current challenges and future directions that could be helpful for further improvement and development of similar types of clinical solutions

According to your suggestion, we added information about the inclusion criteria that you can find in the updated version of the paper in lines: 174-182. Besides, in the conclusion, the limitation of the reviewed studies has already been added. The added information is:

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

The inclusion criteria: i) papers must be written in English, ii) the paper must be related to two prime terms (machine learning and stem cell), iii) the paper should be journal or 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.

Comment 3: Third, in the introduction, it is helpful to analyze the limitations of existing reviews by using examples. Please also have a brief review of ML algorithm, rationale, implications, and

limitations.

Reply 3: Thank you for your suggestions. We have added new information about the limitations of existing reviews with examples in the introduction section. You can find the added information in the updated version of the paper in lines: 128-141. Additionally, for ML algorithms, their implications and limitations have been added in the reply to your next question. We hope for your acceptance. The added information is:

For example, Gupta et al. [16] reviewed several ML-based solutions for hematopoietic stem cell 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 stem cell (hematopoietic stem cell) which doesn't provide any comparative result about the effectiveness of the ML algorithms on other kinds of stem cell classification. Similarly, Coronnello and Francipane [18] reviewed several clinical solutions that are built with artificial intelligence and machine learning techniques for Pluripotent Stem Cell (iPSC) classification. Pluripotent Stem Cells (iPSC) are significant for several therapeutic solutions, however, considering single stem cells for review is not much helpful way to identify the effectiveness of artificial intelligence and machine learning techniques in cell culture or cell biology. However, in another study, Kusumoto and Yuasa [17] 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 (ML) competence in cell biology.

Comment 4: Fourth, in the methodology, the authors need to update the literature search to identify the most recent studies, and specify the inclusion criteria. In the main text of the review, in addition to review the findings, please have more details on the research design, algorithms used, classification accuracy parameters, and analyzing their methodology limitations. In the conclusion, please have comments on the major clinical issues in stem cell investigation and analyze whether ML has and how ML will address these issues.

Reply 4: Thank you for the suggestion, but unfortunately our literature search was conducted between the years of 2018 to 2021, and based on this result we have a plan to do an extensive literature review in further concerning more updated studies. We have added the inclusion criteria in the updated version of the paper.

For the other suggestion, we have added more information in Table 1 and also in the description (lines: 327-361) focusing on your suggestions that you can find in the updated version of the paper. The added information is:

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 machine learning 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 machine learning technique is promising in most of the domains including the stem cell investigation process because of its several advancements such as is easier to identify specific cellular features that help to predict cell function. Mostly, many of the advancements of machine learning 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 machine learning-based applications for the cell classification process have grown day by day, some limitations and challenges are 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 stem cell biology.

In conclusion, we added some information (lines: 462-469). The added information is:

As stem cell analysis or classification is important for many areas such as medicine screening, cell culture process, disease classification, etc. thus large number of stem cell 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, machine learning 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 of ML techniques in the stem cell investigation process. In our reviewed studies.

Comment 5: Finally, please cite several related papers:

1. Lin Y, Tang M, Liu Y, Jiang M, He S, Zeng D, Cui MY. A narrative review on machine learning in diagnosis and prognosis prediction for tongue squamous cell carcinoma. Transl Cancer Res 2022;11(12):4409-4415. doi: 10.21037/tcr-22-1669.

2. Wang S, Wang Q, Fan B, Gong J, Sun L, Hu B, Wang D. 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(3):699-711. doi: 10.21037/jtd-22-206.

3. Ji XL, Ma L, Zhou WH, Xiong M. Narrative review of stem cell therapy for ischemic brain injury. Transl Pediatr 2021;10(2):435-445. doi: 10.21037/tp-20-262.

Reply 5: Thank you. We have added the suggested literature in the updated version of the paper.