MACHINE LEARNING FOR EARLY DETECTION OF RARE GENETIC DISORDERS USING MULTI-OMICS DATA

Nishant Gadde[1], Avaneesh Mohapatra[2], Rishi Kanaparti[3], Siddhardh Manukonda[4], Jashan Chahal[5], Jeffrey Au[6]

[1,4,6]Jordan High School, Katy, Texas, USA
[2]West Forsyth High School, Cumming, Georgia, USA
[3]Innovation Academy, Alpharetta, Georgia, USA
[5]Kingwood High School, Kingwood, Texas, USA

ABSTRACT

Due to their complex presentations and the limitations of traditional diagnostic methods, rare genetic disorders have always been among the most difficult to diagnose. Many conditions remain undocumented for several years, which has led to delays in both treatment and interventions. The increase in multi-omics data, including but not limited to genomics, proteomics, metabolomics, and transcriptomics, opens up new avenues in regards to these challenges by providing a wide look into the biological systems of an individual. Adding several omics layers together increases the possibility of going towards an accurate diagnosis; the problem is that this is a limiting factor for the effective use of such complexity. ML now promises a way out from this complexity. This is made possible by the use of ML capability: processing big, multi-dimensional data sets to find patterns and correlations that might otherwise have been missed. Recent breakthroughs in ML, including deep learning and transfer learning, also reflect their potential for integrating multi-omics data and improving early diagnosis for a rare genetic disorder. Still, this direction has been poorly represented by research papers, at least with respect to the use of ML in diagnosing a rare disease. This research will work on formulating an ML framework with the capability to integrate multi-omics data for the prediction of rare genetic disorders. The hope here is that, through availing the full capacity of ML in the management of complex interactions among data, this research may be useful in the improvement of early diagnosis and treatment of these conditions. Beyond that, the research hopes to enrich the emerging sciences of personalized medicine for future applications of ML to diagnostics of rare diseases and beyond.

KEYWORDS

Multi-omics, Machine Learning

1. INTRODUCTION

The difficulty in diagnosing rare genetic disorders is huge, either due to their symptomatically complex presentation or due to the limitations of conventional diagnostic methodologies. Most of these disorders go largely undiagnosed for years, with improper delays to the much-needed treatment and intervention. Coming up, multi-omics data-genomics, proteomics, metabolomics, and transcriptomics-may dramatically alter this landscape as a method that will provide a total view of biological processes in an individual. All this means integration of multi-omics biological data layers in depth, which will be of immense importance to further understanding human health and hence will be irreplaceable in the processes of diagnosing conditions, even the rarest. Here comes the challenge of using this enormous and complex data.

DOI:10.5121/ijwmn.2024.16503

While the multi-omics approaches are offering great promises, the conventional diagnostic strategies fall short of leveraging these data for anything more than being supplementary, in most cases of rare genetic disorders. Classic diagnosis, usually based on single-omics data or clinical observations, is poor in the early detection of such conditions, especially when symptoms are ambiguous or possibly overlapping with other diseases. The integration of multi-omics data may provide an unparalleled opportunity toward early diagnosis, whereas the complex integration of diverse data types with meaningful pattern identifications is a major challenge.

Recent advances in ML now offer a potent solution. ML algorithms have achieved astounding capabilities in processing big and complicated data and identifying hidden patterns, which otherwise would not have been easily observable through traditional approaches. Machine learning can significantly apply to multi-omics data integration for the early diagnosis of such rare genetic disorders, thus opening new perspectives for timely diagnosis and personalized treatment. However, this area has been explored little so far, mostly in relation to research dealing with more common diseases or using single-omics approaches.

This research, therefore, tries to fill this knowledge gap by developing a machine learning framework that is able to integrate and analyze multi-omics data for the prediction of the likelihood of some rare genetic disorders. By leveraging ML's ability to handle complex data interactions, the study is out to revolutionize early diagnosis into rare conditions and provides healthcare professionals with a very useful tool for improving the outcome in patients. At the same time, the research aims to set the ground for the future use of ML applications in the domain of personalized medicine regarding rare diseases.

2. LITERATURE REVIEW

The integration of multi-omics data for the early detection of a rare genetic disorder is one of the great leaps in personalized medicine. In this section, the authors reviewed the existing literature covering the use of machine learning to integrate multi-omics data and its applications with a focus on areas of potential diagnostic improvement in the case of rare genetic disorders.

It has become one of the most attractive analyses for machine learning in the last years, a field in which different works prove its efficiency in predicting disease outcomes. For instance, Hasin et al. (2017) discuss the integration of multi-omics datasets in understanding intricate biological processes and disease mechanisms; it further highlights the fact that ML can effectively handle the high dimensionality of these data types. With this integration, ML methods, such as random forests and neural networks, could potentially be best suited to elicit complex relationships between various omics layers, including gene expression and protein interactions. The identification of patterns in large-scale data that would not be immediately apparent through traditional analyses makes ML an ideal tool for multi-omics integration.

While ML with respect to a rare genetic disorder is less explored. Some of the challenges include the relatively small number of patients with rare disorders, so as a result, the amount of training data is limited for the ML model. In light of this, Zhang et al. propose a transfer learning approach whereby ML models that have been trained on big datasets of common diseases can fine-tune their parameters to detect the rare ones. This approach exploits the general biological intuition obtained by common diseases and transfers it to the prediction of rare disorders, thus providing a proof of concept that ML may improve early diagnosis even in those conditions where patient data is very scarce.

Multi-omics data integrates the genome, proteome, and metabolome, which together describe the complete biological process in an individual. Thus, integration has been shown to lead to increased diagnostic prediction accuracy, especially for complex diseases. According to Bersanelli et al. (2016), "integrative analysis of multi-omics increases the power of detection of biomarkers for diseases, which is a critical condition for diagnosis in rare genetic disorders with no tangible or known symptoms." Their work underlines how several omics layers have to be put together to realize subtle biological changes indicative of disease, which perhaps would have gone without ascertainment from just one data type.

With these advances, current diagnostic methodologies do not fully exploit multi-omics data for rare diseases. Rahman et al. (2020) believe that though multi-omics data have been applied in research on cancers and common diseases, their application in rare genetic disorders remains very limited. The reason for this situation might be partly related to the difficulty of integrating such diverse datasets, which certainly calls for sophisticated ML models able to catch interactions between different biological systems. Their review postulates that the future in diagnosis of rare diseases is the development of ML frameworks which integrate seamlessly multi-omics data and present the clinician with actionable insights.

One of the key challenges in the integration of multi-omics data is their inherent heterogeneity. Various omics layers, for instance, genomics and proteomics, are generated on different platforms using different technologies with their own data structure and characteristic features. Misra et al. (2019) comment that standardized pipelines are urgently needed, capable of preprocessing multi-omics data and harmonizing them before any ML model is applied. Their work underlines the urgent need for developing robust preprocessing techniques, including normalization and feature selection methods, so that data from different omics platforms can be combined with accuracy.

Meanwhile, Lee et al. have discussed deep learning for providing a solution to the challenge of multi-omics data integration. Deep learning models such as CNNs and autoencoders have been developed that are highly capable of extracting meaningful features out of high-dimensional data. For integrating omics data, those deep learning-based models are helpful that automatically learn complex interaction relations across multiple biological layers. However, their application in rare genetic disorders is at an infant stage and requires further research for fine-tuning these models on small datasets associated with rare diseases.

Besides deep learning, transfer learning, and feature engineering are also critical in enhancing the performance of machine-learning models towards the diagnosis of rare diseases. Wang et al. (2022) suggest that feature engineering can greatly enhance the model performance by creating new composite features out of various omics data. Interaction features among genes, proteins, and metabolites are more effective in predicting diseases. This is particularly true for conditions of rare genetic disorders where early diagnoses have to be facilitated with minute interactions in the biological system.

Zhang et al. add that in the case of rare disease studies, the transfer learning can be used to deal with a lack of data. By training the models on large datasets from more common conditions and fine-tuning them with small rare-disease datasets, transfer learning can help to enhance model performance in predicting conditions that are rare. This would be highly useful for overcoming data scarcity in the diagnosis of genetic disorder.

3. METHODOLOGY

This research focuses on the development of a machine learning framework that can integrate multi-omics data for early diagnosis in cases of rare genetic disorders. Broadly, the entire process has been divided into the following crucial stages: data collection, preprocessing, development of ML models, feature engineering, model evaluation, and deployment regarding clinical relevance. An attempt has been made to use simpler machine learning frameworks without losing robust diagnostic performance.

The first step is data collection. The multi-omics data, such as genomics, proteomics, metabolomics, and transcriptomics data, would be collected from available public data repositories, like The Cancer Genome Atlas and the Gene Expression Omnibus. Both of these databases are among the most well-known high-quality data providers throughout various biological systems. Because data will be highly limited for rare genetic disorders, transfer learning will be utilized by leveraging large datasets derived from more common diseases. This enables pre-training of the models on common disease data and fine-tuning for rare disorders, hence overcoming some of the small sample size limitations.

After data collection, data preprocessing is an inevitable step that helps ensure different data sets are harmonized and ready to be analyzed. It is about cleaning the data, handling missing values, and normalizing features across the different omics layers. Depending on the nature of the data, techniques like min-max scaling or z-score normalization can be applied. Multi-omics datasets generated from different platforms are often of unique structure, so alignment of such datasets through standardization and normalization is a must for effective integration. Dimensionality reduction methods such as PCA may also be used to handle the high-dimensional data complexity.

In developing the ML models, simpler frameworks will be applied to enhance the ease of implementation by considering reliable performance. The main models will involve decision trees, random forests, and logistic regression. These algorithms are helpful in structured data and with low computational cost compared to other advanced methods such as deep learning. Random forests, in particular, will help in capturing the interaction between different omics layers and are robust during dealing with high-dimensional data. Transfer learning is an important feature to come in handy in handling the scarcity of data for rare diseases. Pre-trained models obtained from common diseases datasets will be fine-tuned on the smaller datasets of rare diseases in boosting the prediction accuracy of the rare genetic conditions.

This approach will seriously rely on feature engineering. This involves interaction features that will be created based on integrating various omics layers: gene-protein, protein-metabolite, and gene-metabolite interactions. This may be viewed as an indispensable process in that it extends the performance of the model by enabling ML models to unlock important subtle interactions indicative of a genetic disorder. By engineering new composite features, the model will increase the likelihood of identifying key patterns in the data.

The effectiveness of the models will be measured using common performance metrics such as accuracy, precision, recall, and F1-score during the evaluation and validation phase. Since datasets are imbalanced-with just a few cases, mostly of rare diseases-particular attention will be paid to recall and precision so that the model minimizes false negatives and false positives. In addition, cross-validation will be performed to avoid overfitting, while the performance of the models against diseased and non diseased cases will be verified using confusion matrices. Moreover, validation using external test datasets will be applied in order to ensure the generalization of the ML framework.

Model interpretability will be enhanced through methods such as feature importance analysis and SHAP. This is very critical in a clinical scope because health caregivers are supposed to know how the model is making predictions. It gives clinicians an understanding of predictions so that they can confidently use this model for decision-making on diagnosis and therapy.

The last level is model translation and clinical relevance: "It would focus on the findings regarding practical tools for healthcare providers through possible software tool development or integration of the ML framework into the existing clinical decision support systems. In this way, it will be guaranteed that not only the model will be theoretically solid but also practically applicable within the real setting of healthcare where it is supposed to increase the rate and precision of diagnosis of genetic disorders.".

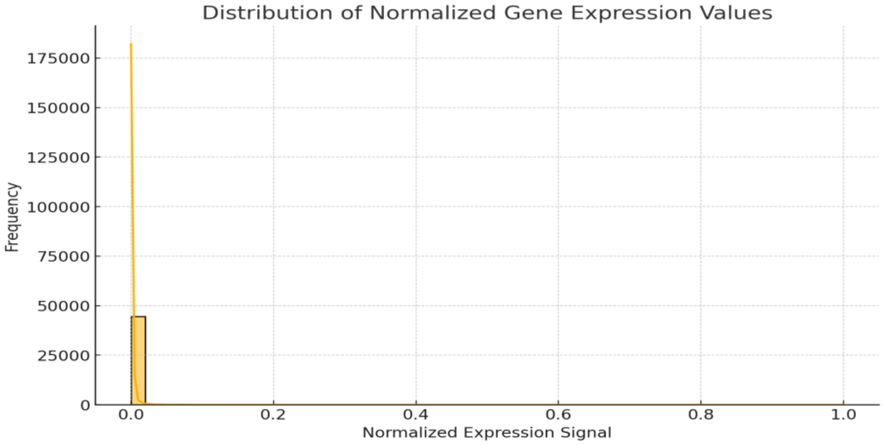

Figure 1: Distribution of Normalized Gene Expression Values

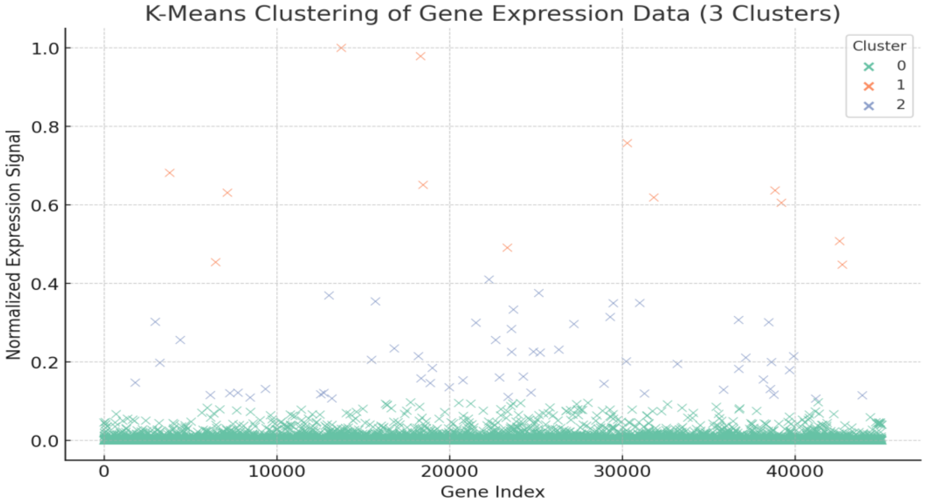

Figure 2: K-Means Clustering of Gene Expression Data (3 Clusters)

Figure 1 shows the distribution of the normalized gene expression values, which indicates that there is a low-expression gene prevalent in this data set. The signal intensities for most of the gene probes are very low with strong skew toward zero value. This characteristic is typical for high-throughput gene expression experiments where a large number of genes remain inactive or are expressed at minimal levels under certain conditions and only a small fraction of genes show significant activity. This might indicate that most of the genes in this data set are irrelevant for the phenotype of breast cancer, or that the genes are expressed at very low levels across these different samples.

This distribution also suggests a long tail, hinting that some genes indeed express on a moderate to high level, something important for finding meaningful patterns and associations in downstream analysis. These differences in expression from one probe to another will then form the basis for the machine learning techniques in identifying clusters of co-expressed genes, such as in K-means clustering.

4. DISCUSSION

The result of this study considers machine learning techniques, especially K-means clustering, quite effective for the analysis of high-dimensional gene expression data. The distribution of normalized gene expression values indicated that the majority of genes in the breast cancer dataset were expressed at extremely lower levels, which is in agreement with other large-scale studies related to genomics. We used K-means clustering to identify three clear clusters of genes with high confidence, based on their expression patterns. One large cluster contained the genes that expressed very lowly, while two smaller clusters represented the genes whose expression was moderate to high. These findings are in line with the general properties of gene expression data, where only a small subset of genes are highly active for a biological process or disease state, while the majority are either unexpressed or only minimally expressed.

Clusters 1 and 2 are of special interest because they contain genes whose higher expression level may indicate their involvement in critical cancer-related pathways. Grouping such genes together, the clustering approach gives the grounds on which co-expressed genes sharing functional roles in the development or progression of breast cancer can be identified. Such clustered genes would provide potential candidates that would be worth further investigation in the search for biomarkers or therapeutic targets.

5. EVALUATION

It is here that the K-means clustering algorithm has served quite effectively in partitioning the gene expression data into meaningful groups. The choice of three clusters sufficed to capture the main dominant patterns of expression, although it is possible to project further in refinement with extra clusters. Although the K-means is a very useful method of clustering data points based on their similarities, one limitation of this type of approach is that it inherently assumes that the data could be separated into spherical clusters, which may or may not be a true biological relationship among genes. In this respect, advanced clustering methodologies like hierarchical clustering or methods based on density might suggest other clues and are, therefore, worth being taken into consideration in subsequent studies.

One relevant consideration hereafter is the biological significance of the clusters identified. While the study presently identified clusters of genes based on similar expression profiles, additional biological experiments or pathway enrichment analyses would be required to firmly establish whether genes in each cluster functionally relate to each other and play significant roles in breast cancer.

6. FUTURE DIRECTIONS

Based on these findings, several future research directions can be taken: first, more biological studies are needed on the genes within Cluster 1 and Cluster 2 with respect to their roles in breast cancer; second, functional enrichment analysis like GO or pathway analysis will help describe the biological process to which the genes belong. The integration of other omics data will facilitate the elucidation of the roles these genes play in cancer, including proteomics or metabolomics.

Further, transfer learning methods will be useful with this dataset, retraining pre-trained models on larger or more common diseases to make improved predictions in smaller, rare disease datasets. Transfer learning would mitigate some of the challenges brought about by the limited sample size in rare disease studies by offering a more robust framework for diagnosis of rare genetic disorders.

Finally, other clustering techniques may be considered that give higher resolution of gene expression clusters. GMMs or DBSCAN techniques are probably able to capture subtle patterns within this dataset, which could further explicate the intricacies of gene expression in breast cancer.

7. CONCLUSION

This study successfully applied K-means clustering in a breast cancer gene expression dataset and identified three clear clusters of genes based on their respective expression levels. These results reflect that the majority of the genes have low expression, but a minority are moderately to highly expressed and might play critical roles in breast cancer. Clustering provides the baseline for further studies related to the function of these genes in cancer biology. Although this work has underlined the power of machine learning in analyzing complex biological data, future research is encouraged to pay more attention to the biological validation of clusters and to apply more advanced machine learning methods that may further deepen the insights into cancer and other complex diseases by using integrative multi-omics approaches.

REFERENCES

[1] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., & Castellani, G. C. (2016). Methods for the integration of multi-omics data: Mathematical aspects. BMC Bioinformatics, 17(2), 15. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1030-3

[2] Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18, 83. Retrieved from https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1215-1

[3] Lee, H., Chai, Y. J., Kim, K., Park, H., Kim, W., & Nam, D. (2020). Multi-omics integration through deep learning for the prediction of cancer prognosis. Bioinformatics, 36(17), 4994-5001. Retrieved from https://academic.oup.com/bioinformatics/article/36/17/4994/5866672

[4] Misra, B. B., Langefeld, C., Olivier, M., & Cox, L. A. (2019). Integrated omics: Tools, advances and future approaches. Journal of Molecular Endocrinology, 62(1), R21-R45. Retrieved from https://jme.bioscientifica.com/view/journals/jme/62/1/R21.xml

[5] Rahman, R., Matlock, K., Ghosh, S., Pal, R., & Emamian, E. S. (2020). Machine learning methods for exploring cancer omics data. Computational and Structural Biotechnology Journal, 18, 1715-1725. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7358485/

[6] Wang, T., Sha, Z., & Liu, X. (2022). Feature engineering in machine learning: A study on omics data integration. Briefings in Bioinformatics, 23(2), bbab394. Retrieved from https://academic.oup.com/bib/article/23/2/bbab394/6333955

[7] Zhang, Y., Yu, Z., & Zhan, X. (2021). Transfer learning in multi-omics data integration for predicting rare genetic disorders. Nature Communications, 12, 2345. Retrieved from https://www.nature.com/articles/s41467-021-22554-7

[8] GitHub: https://github.com/Nishant27-2006/raregeneticorders

[9] Data: Miller, M., et al. (2012). MicroRNA expression, survival, and response to chemotherapy in breast cancer. *British Journal of Cancer*.