Applied Science and Biotechnology Journal for Advanced Research e-ISSN: 2583-553X
Peer Reviewed Journal
Volume-3, Issue-3 (May 2024), Page: 8-13 https://doi.org/10.5281/zenodo.11211521

A Study on Deep Learning Architectures and Dimensionality Reduction Techniques on Gene Expression Data

Remyamol K M[1] and Philip Samuel[2]
[1]Department of Information Technology, School of Engineering, CUSAT, Kerala, India
[2]Department of Computer Science, CUSAT, Kerala, India

[1]Corresponding Author: rems84@gmail.com

Received: 05-04-2024 Revised: 24-04-2024 Accepted: 13-05-2024

ABSTRACT
Genomics, driven by the evolution of high-throughput sequencing and microarray technologies, has become one of the key inventions of cracking the secrets of complex biological systems. The deep learning architecture not only provides with a powerful tool to derive the hidden insights from the huge amount of genomic data, but also enables to mine meaningful information. In this study, we will examine the application of deep learning methods in the analysis of genomics data, specifically on dimensionality reduction and predictive modeling for binary phenotypes. We focus on the problems with the existing strategies, spot the avenues for the further research, and provide you with a glimpse of the dramatic influence of deep learning on genomics. In this study, we delve into the application of deep learning methods in the analysis of genomic data, with a specific focus on two crucial aspects: dimensionality reduction and predictive modeling for binary phenotypes. Dimensionality reduction techniques are essential for tackling the high-dimensional nature of genomic data, where thousands or even millions of features (e.g., gene expressions, genetic variants) are measured for each sample. Deep learning models can effectively capture the complex relationships and patterns within this high-dimensional space, enabling the extraction of lower-dimensional representations that preserve the most salient information. Throughout this study, we critically examine the existing strategies and approaches in the field of genomics, identifying their limitations and highlighting the avenues for further research. We explore how deep learning can address these challenges and provide a glimpse into the dramatic influence this technology is poised to have on the field of genomics.

Keywords: biomarkers, cancer prediction, deep learning, dimensionality, epigenetics, feature learning, gene expression

I. INTRODUCTION

Genomics, the discipline that deals with the analysis of an organism's entire genetic makeup, puts forward the perspectives of understanding fundamental biological processes, reconstructing disease-related mechanisms, and boosting personalized medicine [1]. Nowadays, the field of genomics is experiencing an unprecedented expansion, which has been driven by the technological advances in high-throughput sequencing and microarray technologies [2]. We face some particular problems of data analysis and representation since we deal with genome data that is complex within itself and even more so when combined with other datasets such as transcriptomic data that capture gene expression levels across different biological conditions [11]. The standard methods that are mainly analytical based usually are being struggling to deal with all the dimensionality and complexity of these datasets thus the outputs they are giving are not that best [12].

The study comprehensively investigated various dimensionality reduction approaches for genomic data analysis. Key techniques examined include Non-negative Matrix Factorization for identifying biological processes and gene modules, Deep Variational Bayes for estimating probabilistic dependencies, and Manifold Learning methods like Isomap and LLE for uncovering nonlinear relationships. Additionally, sparse coding, dictionary learning, and graph-based approaches were assessed for recognizing meaningful features and preserving data geometry. Deep generative models like GANs and VAEs were explored for jointly learning data distributions and latent representations. Techniques like Independent Component Analysis, kernel methods, and Self-Organizing Maps were also evaluated for their ability to extract independent components, map nonlinear spaces, and visualize high-dimensional datasets respectively. This thorough exploration aimed to identify optimal dimensionality reduction strategies for genomic data analysis.

In this context, deep learning architectures are several advantages, such as the ability to learn non-linear representations, capture intricate relationships between genes, and handle large-scale datasets effectively [13, 14]. Through utilization of deep learning methods, researchers may discover hidden connections, discover biomarkers, and develop predictive models for biological systems and phenomena [4]. The potential strong correlation between the dimension of the

gene expression data and the reason behind the curse of dimensionality is, however, one of the most crucial limitations for the proper analysis and interpretation. Therefore, dimensionality reduction methods are to some extent the answers to this problem; they decrease the amount of noise, make the matrix less densely set and allow the principal features to expose.

The deep learning architectures have become the most effective tools for extracting insights from the complex genomic datasets [3]. The structure and functions of the human brain inspire these networks that can acquire complex models from very high dimensional data. Among the currently implemented models in genomics data analysis are Artificial neural networks (ANN), convolutional neural networks (CNN), recurrent neural networks (RNN) and autoencoders (AE) [5,6]. The Artificial Neural Networks (ANN) is a computer model of a brain that imitates the structure of neurons in the brain and are applied in feature selection, classification and dimensionality reduction tasks [6]. Convolution Neural Networks (CNN) are very effective in studying visual images and also are successfully capable of seizing on the spatial character in genome sequences [7]. That should make Recurrent Neural Networks (RNN) fit for the job because of their temporal dynamic behavior relationship; that means they are suitable for sequences data analysis tasks like time-series gene expression data [8]. The autoencoders (AE) have the capacity to learn the compact representations of the high-dimensional data, which makes it possible to reduce dimensions and extract features [9].

II. MATERIALS AND METHODS

In this paper, we move the spotlight on the utilization of deep learning architectures in the context of genomics data analysis, explaining first dimensionality reduction before we proceed to talk about the prediction models aimed at binary phenotypes. In the article, we talk about the problems and limitations of the current approaches, identify the areas for further study, and emphasize the possibility of deep learning to revolutionize our understanding of genomics.

Genomics data analysis methods that employ deep learning networks reveal multiple advantages over those based on traditional techniques [10]. Such features as the non-linear representations, the ability to catch the intricate relations between genes, and the possibility to work with large datasets [11] are the main advantages of the algorithm known as gene network reconstruction. Due to the use of deep learning technology, scientists are capable of finding concealed trait, spotting biomarkers and developing predictive models for numerous biological processes.

The enormous amount of genomic data produces a great analytical complexity related to data analysis and interpretation [8]. Transcriptomic is a type of gene expression data that has been captured in various biological phenotypes and conditions; this type of data is richest and the most complex at the same time since it is a reflection of underlying biological mechanisms [11]. The traditional analytical methods are not able to handle the high dimensionality and complexity of transcriptomic data, this usually leads to the suboptimal results [12]. Deep learning models are allowed to demonstrate several advantages over the other methods, which include the ability to learn non-linear representations, the power to capture the complicated interrelations between genes, and the capability of handling large scale datasets in an effective way [16]. Through the application of deep learning techniques researchers may be in a position where they could find hidden patterns, biomarkers, and even make predictive models for biologic occurrences [13,14].

Furthermore, dimensionality reduction methods have become essential tools in gene expression data analysis because they eliminate the curse of dimensionality, remove noise and help in the identification of features that can be used for downstream tasks of clustering, classification and biomarker discovery. Opposite to the traditional dimensional reduction methods, namely Principal Component Analysis (PCA) [11] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [12], which have been extensively used in gene expression studies, high-dimensional data visualization techniques transforms data with high dimensionality into simplified forms compatible with visual representation. Although these methods were helpful, they often did not take into account the complex, non-linear relationships in the data and they had difficulties distinguishing between the different sources of variation that lead to gene expression patterns.

The implementation of recent deep learning innovations for complex data dimensionality reduction techniques has been made possible art a high-dimensional level. High feature-learning techniques, which use the strong representative capacity of deep neural networks, have demonstrated remarkable results while extracting low-dimensional embeddings that give the data the necessary structure as well as the existing patterns [13]. Moreover, disentangled representation learning is a technique that has been developed to separate the latent factors that are responsible for data variability and thus, the biological relevant signals can be isolated from the technical or unwanted sources of variation [14].

2.1 Deep Learning Architectures Used

The deep learning architectures have developed into powerful instruments that enable us to generate useful information from the rigorous genetic datasets [3]. The most widely used data mining models in genomics include artificial neural networks (ANNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and auto-encoders (AEs) [5]. Artificial Neural Networks (ANNs): As ANNs are organic in nature, the underlying technology tries to emulate the

functionality of brain's synapses. They are built with the neurons connected together, which are arranged in layers. + Data input, which gets its values from original data, is the first layer, and the final output is the result of the output layer. Interlayers between these hypernetworks are opted for at least one or more hidden layers that perform calculations for features of input data.

In the field of genomics data analysis, ANNs can be used for different tasks such as feature selection, classification and dimensionality reduction. It is especially suitable for methods of calculating the non-linear relationships between the feature vectors which is known as an input feature. Genetic and epigenetic programming (parameterization), as well as environmental input (target variables), has significant effect on the outcome of a disease. g. The system complexity being abstractly owned by the field of computer science involves abstraction, disease status or phenotypes[7,8]. Weights between neurons is adjusted to learn these convolution during the training in the way to make true outputs as close as possible to the predictions.

Convolutional Neural Networks (CNNs) are a specialized type of neural network architecture that performs best in the processing and analysis of grid-like data, such as images and genomic sequences. They do this very easily by modeling spatial dependencies and the small intricate patterns in the data. In genomics, CNN is one of the algorithms that is able to solve different task such as sequencing for DNA or protein sequence, regulatory motif detection, and structural prediction of biomolecules. The main ingredients of CNNs are the convolutional layers which apply the filters to the input data and pooling layers for the feature mapping while keeping the most important information [9]. hrough stacks they can have the capacities to learn the hierarchical representations of the input data step-by-step, progress to the feature identification which varies from low to high level. g. This stage starts by grounding on the basics (also known morphemes) and later transitioning to more elaborate, advanced concepts (higher-order features).

Recurrent Neural Networks (RNNs) are arranged for processing sequential data, like time-series or sequences of genomic data. Unlike the feedforward neural networks that operate with no internal memory, the RNN owes its memorization property of an internal state that helps store and recall information from previous time steps in the sequence. In genomics, the group of tasks RNNs has taken into consideration includes predicting gene expression levels stably over time, modeling of regulatory elements effects on that gene expression, and the entire genome sequence annotation with functional information. RNNs are based on the recurrent layer, which serves as the key component of the model. The recurrent layer takes the current input and the previous hidden state as inputs, hence the network can learn long-term dependencies and patterns in the sequential data.

In particular, LSTM and GRU are highly popular RNN variants showing overall better performance when learning long-range dependencies of sequential data. Autoencoders (AEs) are a type of an unsupervised neural network that has learned to reproduce its given inputs through a compressed representation which is known as latent or encoded space. They consist of two main components: an encoder network that reduces the high-dimensional input data to a lower-dimensional latent space, and a decoder network that reconstructs the original input. he autoencoders have been widely used for a range of tasks, including, dimensionality reduction, denoising, and feature extraction, among others, in genomics data analysis. Through learning to rebuild the input data from its compressed portrayal, autoencoders capture the foremost characteristic features as well as patterns detected in the data, thereby resulting in dimension reduction without the issue of loss of essential data. Autoencoders have several variants, for example, variational autoencoders (VAEs), and adversarial autoencoders (AAEs), which have been developed to learn more robust and disentangled representations, so that they can separate the different factors of variation within the data [11].

These neural network architectures have demonstrated incredible power in different domains, and the application to genomics has produced wonderful results as well. Nevertheless, as it was mentioned previously, the hurdles like hermeneutics, breakdown of data and generalization as well as relevance of biological data must be surmounted in order to use fully potential of deep learning in genomics research and personalized medicine.

2.2 Dimensionality Reduction Techniques

The study explored a various of dimensionality reduction techniques, categorized into three main approaches: The study explored a range of dimensionality reduction techniques, including Non-negative Matrix Factorization for deciphering biological processes and gene modules, Deep Variational Bayes for estimating probabilistic dependencies in high-dimensional data, and Manifold Learning methods like Locally Linear Embedding for identifying nonlinear relationships and local components. Sparse coding algorithms were assessed for recognizing meaningful features and reducing dimensionality, while dictionary learning methods aimed to represent data as sums of basis elements, capturing sophisticated patterns [2]. Graph-based approaches, such as Laplacian Eigenmaps and Graph Laplacian Regularization, employed data geometry for dimensionality reduction while preserving biological relationships. Deep generative models like Generative Adversarial Networks and Variational Autoencoders jointly learned data distributions and latent representations. Independent Component Analysis revealed uncorrelated components, while kernel methods like Kernel PCA and Kernel CCA extended linear

techniques to nonlinear spaces. The Organizing Maps mapped high-dimensional data to low-dimensional grids, enabling visualization and interpretation of large datasets [3].

2.3 Deep Learning Architectures

The study employed various deep learning architectures for dimensionality reduction and predictive modeling tasks:

1. Autoencoders (AE): Unsupervised neural networks that learn to reconstruct the input data from a compressed representation, enabling dimensionality reduction and feature extraction.
2. Adversarial Variational Autoencoders (AVAE): A variant of variational autoencoders that incorporates adversarial training to learn more robust and disentangled representations.
3. Deep Clustering Models: Neural network architectures designed to learn low-dimensional embeddings that capture the inherent structure and patterns within the data, facilitating cluster discovery and visualization.
4. Variational Inference for Disentangled Representation Learning: Variational inference techniques, such as β-VAE (Beta Variational Autoencoder) and FactorVAE, aim to disentangle underlying factors of variation in high-dimensional data by explicitly modeling the latent space. These methods introduce additional regularization terms during training to encourage the learned representations to be sensitive to specific factors while being invariant to others. In genomics, variational inference for disentangled representation learning can help separate biological signals from technical or confounding factors, leading to more interpretable and biologically relevant latent representations of genomic data [4,5,6].

2.4 Predictive Modeling and Evaluation

Logistic regression models were utilized for predicting phenotypes and to assess the predictive performance of the models, a rigorous cross-validation strategy was adopted, where the data was partitioned into training and testing subsets iteratively. The Matthews Correlation Coefficient (MCC) [5] was chosen as the basic evaluation metric due to its suitability for binary classification tasks and robustness against class imbalances. Moreover, a permutation testing structure [4,5] was used as a mean to empirically determine the statistical significance of the observed predictability performance. Through random sequences of phenotype labels, repetition of the cross-validation procedure, and application of the score generation of the null distribution to access the significance of association between input features and phenotype labels was performed.

III. LIMITATIONS OF DEEP LEARNING IN GENOMICS

Deep learning (DL) models suffer from the lack of interpretability which is one of the reasons behind the reluctance for their wide-scale use. Though DL models demonstrate amazing capabilities to extract from the data the most complex patterns, they sometimes fail when it comes to the mechanisms explanation which is a major barrier for understanding the processes in biology. The black box nature of DL models makes it challenging to explain their predictions and to get insights into the biological relevance of the learned representations. Ensuring interpretability is paramount in genomics research because it grants an opportunity for scientists to reveal the reliability of model conclusions, put forward hypotheses for further scrutiny, and thus, discover new biological information. However, the interpretation of the black-box predictions is still a great challenge, while DL methods with better interpretability are being developed such as attention mechanisms and saliency maps which enable to see which features are important for these predictions [8]. Thus, genes, genomic regions, etc. are the most important parts of the model. Moreover, the domain knowledge and the biological constraints in an DL architecture can improve interpretability where the result of the model matches with the current biological knowledge.

DL models generally need a lot of annotated data to process these patterns of complexity. Nevertheless, getting genomics datasets with enough information can be a problem, especially when you are dealing with rare diseases or some specialized biological conditions. The scarcity of data we strive to get, is just the major drawback of DL models in genomics for their performance and generalization. The fine-tuning models, which are already trained on the data sets of large size and then adjusted for the small and tailor-made datasets, may overcome the barriers of data scarcity and even bring about higher accuracies of the models after-learning. The other strategy named semi-supervised and self-supervised learning paradigms which depend using unlabeled data can be considered as an efficient and fast learning model. DL models may be hampered in their ability to generalize across different biological contexts or datasets because of the differences in the experimental protocols, batch effects and data heterogeneity. Generalizing to uncharacteristic data is not an easy task, substantial effort, and it is crucial for providing thoughtful DL models in genomics research and clinical practice [9,10]. These two properties are crucial for the task of telling a story starring another entity as a core character. Providing a detailed, well-rounded background and highlighting their journey are both key aspects in this task. What is more, the adversarial training methods, which involve the augmentation of the training data with the perturbations to simulate different biological contexts, improve the generalization and the robustness of the models.

Deep Neural Networks, especially those with multiple architectures, have more prone to overfitting, the case is relevant in the high-dimensional genomic data where the number of features (i.e., variables) is relatively large. g. We have a wealth of biological literature, however, it is in the form of genes, which fluctuate proportionally with the amount of samples. Overfitting is a phenomenon when a model learns the noise or irrelevant patterns in the training data and as a consequence performs poorly on the unseen data. Regularization methods, including dropout, weight decay, and batch normalization, are highly used to restrict the forward bias of neural networks and increase the generalization level. Bayesian deep learning methods, that produce the prediction takes into account the uncertainty into predictive models, is another means that ensures overfitting is reduced as well as also the model uncertainty estimated that guides the decision-making process. DL models should be designed in a way that they can incorporate domain-specific knowledge and biological constraints into their architectures as well as their interpretation frameworks to make the predictions more biologically relevant [11]. It is imperative to combine the efforts of domain experts, computational biologists, and machine learning researchers in order to develop interpretable, data-efficient, and biologically relevant DL models that will drive the transformative discoveries and the translation of the insights into the clinical practice.

IV. FUTURE DIRECTIONS

To address the limitations of deep learning (DL) in genomics and unlock its full potential, several promising directions can be pursued. Developing techniques to translate the outputs of DL models into comprehensible and visualized interpretations will aid in the transparency of the models and bring biological explanation to the forefront. A special class of attention mechanisms, which are utilized to flag the most important features in input, and saliency maps, which reveal the input features underlying the output of the model, are valuable in deep learning model understanding [1]. Interpretability-enhanced DL models, which provide the researchers with the information about the genetic features that are responsible for the prediction of the model, can help to verify the model results and generate hypotheses for further investigation. Given that only few genomic datasets are well annotated, while others are not very adequate, transfer learning and domain adaptation techniques can come in handy by helping knowledge transfer from well-annotated datasets to related domains. Pretrained DL models, learnt on big-scale genomics data set, such as the Cancer Genome Atlas (TCGA) or the Genotype-Tissue Expression (GTEx) project, can be fine-tuned using a smaller, domain-specific data sets that can help in the improvement of the model performance and dealing with data scarcity problems [2]. The domain adaptation methods that align the feature distributions between the source and the target domains can further boost the model's generalization across different genomic contexts [3].

Due to the integration of a cross-section of omics data types involving genomics, transcriptomics, epigenomics, and proteomics is capable to yield more detailed information about biological systems operational. DL-based strategies like deep neural networks and graph-models that integrate multi-omics data at large elucidate intricate molecular interactions and disease background [4]. Through the combined analysis of various omics datasets, researchers can detect biomarkers, pathways and therapeutic targets with greater precision and reliability. The importance of the uncertainty estimates in DL models is to quantify and propagate the uncertainty in order to assess the reliability of the predictions, especially in the clinical applications. Bayesian DL techniques e.g. Bayesian neural networks and probabilistic graphical models offer not only a natural framework for expressing model uncertainties at test time but also for estimation of predictive uncertainty and confidence intervals at the confidence level. Ensembles, which combine decisions obtained from several DL models trained on different data subsets, can also provide valid uncertainty estimate taking into account data variety and model uncertainty [6].

As DL technologies are more and more used in the field of genomic research and clinical practice, the issue of ethical and social consequences should be the main concern. It is vital that data privacy, fairness in genomic technologies distribution and accurate information is communicated from the model limitations and biases stands are guaranteed for an DL responsible applied to genomics [7]. Collaborative activities of researchers, clinicians, policymakers and ethicists are necessary in designing the guidelines and identification of best practices for the ethical deployment of AI and genomics in personalized medicine.

V. CONCLUSION

\

The incorporation of deep learning (DL) architectures into genome data imputing has the potential for getting future biomedical research and medical care into personalized medicine exponentially increased. Nevertheless, the exploitation of the potential of the DL especially faces a number of challenges and limitations such as the interpreter-ability, data-efficiency, generalization, overfitting and biological relevance [7]. The goal of this framework is to improve the human understanding of gene expression data by learning compact, informative, and disentangled representations. This means that, through this framework, people will be able to gain a better understanding of the gene expression data, which will, in turn, facilitate biomarker discovery, disease subtyping, and personalized therapy development. In present context, we have to identify the

flaws and propose new routes to future that include, but not limit to interpretation improvements, better data transportation, integration of multi-omics data, quantification of uncertainty and ethical concerns, to maximize the power of DL development in genomics. Through the full engagement of these upcoming lines of development and the development of an interactive partnership between domain experts, computational biologists, and the machine learning researchers, DL can be the basis upon which sensationalized advancements are discovered and improve clinical decision making, thereby eventually improving patient outcomes in the era of precision medicine.

REFERENCES

1. S. R. Oshternian, S. Loipfinger, A. Bhattacharya, & R. S. N. Fehrmann. (2024). Exploring combinations of dimensionality reduction, transfer learning, and regularization methods for predicting binary phenotypes with transcriptomic data. *BMC Bioinformatics.*
2. Bo Chen, Wai Lam, Ivor W. Tsang, & Tak-Lam Wong. (2012). Discovering low-rank shared concept space for adapting text mining models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1284–1297.
3. Amel Ghouila. (2009). Application of Multi-SOM clustering approach to macrophage gene expression analysis. *Infection, Genetics and Evolution,* 328-336.
4. Saad Sahriar, Sanjida Akther, & Jannatul Mauya. (2024). Unlocking stroke prediction: Harnessing projection-based statistical feature extraction with ML algorithms. *Heliyon.*