International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 13, Issue 4, April-2024, Impact Factor: 8.285
Presented at "ICRETETM-2024", Organized by GSMCOE, Pune, on 22nd - 23rd April 2024

A Survey Computational Models for Gene SelectionBased on Machine Learning

Alka Kumbhar[1], Smita Sapkal[2], Neelam Jadhav[3]

[123]Assistant Professor , Department of Computer Engineering, Genba Sopanrao Moze college ofengineering Balewadi , Pune

ABSTRACT

Understanding gene expression data analysis is crucial for several reasons. Firstly, it helps researchers comprehend how genes are activated or suppressed, providing insights into biological processes, such as development, disease progression, and response to treatment. Additionally, gene expression analysis facilitates the identification of biomarkers for disease diagnosis, prognosis, and therapeutic targets. Moreover, it aids in unraveling complex regulatory networks within cells and organisms, contributing to advancements in fields like personalized medicine and drug discovery. Ultimately, gene expression data analysis is fundamental for advancing our understanding of genetics, biology, and human health.

Gene expression data analysis plays a vital role in machine learning-based classification tasks. By leveraging machine learning algorithms, researchers and practitioners can extract meaningful patterns from gene expression data to classify samples into different categories, such as disease subtypes, treatment response groups, or healthy versus diseased states.

Machine learning models can be trained on gene expression data to learn complex relationships between gene expression profiles and the labels or classes associated with each sample. These models can then be used to predict the class of new, unseen samples based on their gene expression patterns.

Keywords : Gene Expression , Machine Learning , Data Science , Supervised , Unsupervised, Gene selection.

INTRODUCTION

Gene expression is the process by which information from a gene is used to synthesize a functional gene product, typically a protein. This process plays a fundamental role in virtually all biological functions, including cell differentiation, growth, and response to environmental stimuli. Gene expression data refers to the quantitative measurement of the activity levels of genes within a cell or tissue at a given moment. These data provide insights into the molecular mechanisms underlying various biological processes and diseases.[1]

Machine learning (ML) techniques play a crucial role in analyzing gene expression data due to the complex and high-dimensional nature of the data. Here are some key roles that ML methods fulfill in this context: There are many techniques available to capture the gene expressions such as Northern blot, RNA protection assay, Reverse Transcription – Polymerase Chain Reaction (RT - PCR), Serial Analysis of Gene Expression (SAGE), Subtractive Hybridization, DNA Microarrays, Second Generation Sequencing (NGS) and many others. Among these, the most widely used these days is DNA Microarray [7]. The DNA microarray technology manages to capture gene expressions of thousands of genes simultaneously. However, the Microarray result is enormous, with a high dimension, which makes the analysis challenging. Thus, it is necessary to perform gene selection to handle the high dimensional problem by removing the redundant and irrelevant genes. There are many computation techniques used in the field of bioinformatics been carried out over the years, such as Pattern Recognition, Data Mining, and many others to manage the high dimensional issue, yet ineffective [9].Machine learning techniques have emerged as powerful tools for gene selection, offering the potential to extract meaningful insights from high-dimensional genomic data. By leveraging computational models and algorithms, machine learning enables the automated identification of informative gene subsets while effectively handling noise, redundancy, and complex interactions within the data.

This research paper presents a comprehensive survey of computational models for gene selection based on machine learning approaches. We aim to provide an overview of the diverse methodologies, techniques, and advancements in this rapidly evolving field. By synthesizing insights from a wide range of literature, we seek to elucidate the strengths, limitations, and emerging trends in computational gene selection.

LITURATURE SERVEY

Paper Title and Authors	Methodology and Contribution	Limitations
"Deep Feature Selection: Theory and Application to Identify Enhancers and Promoters" by Li et al. (2019)	Introduces Deep Feature Selection (DFS), a deep learning-based method for gene selection, particularly in identifying enhancers and promoters	Requires large amounts of labeled data for training deep learning models. Interpretability of deep models can be challenging.
"Feature Selection in High-Dimensional Omics Data Using Robust Estimators" by Alzubi et al. (2020)	Proposes a feature selection method using robust estimators to handle outliers and noisy data in high-dimensional omics data.	Performance may degrade if the data contains a large proportion of outliers. Limited to specific types of robust estimators.
"Sparse Group Lasso with Weighted Graph for Gene Selection in Cancer Classification" by Li et al. (2021)	Introduces a sparse group Lasso method with a weighted graph regularization term for gene selection in cancer classification tasks.	May require tuning of additional hyperparameters related to graph weighting, which can affect performance. Computational complexity may increase with large-scale datasets.
"Gene Selection Using Multi-Objective Evolutionary Algorithms for Cancer Classification" by Sarhan et al. (2022)	Proposes the use of multi-objective evolutionary algorithms for gene selection in cancer classification, considering both classification accuracy and gene subset size as objectives.	Scalability can be an issue with large-scale datasets due to the computational complexity of evolutionary algorithms. Optimal parameter setting for evolutionary algorithms may require domain expertise.
"Feature Selection for RNA-Seq Data: A Review" by Li et al. (2023)	Provides a comprehensive review of feature selection methods specifically tailored for RNA-Seq data, including filter, wrapper, embedded, and hybrid approaches.	Limited comparison of the effectiveness of feature selection methods across different RNA-Seq datasets. Difficulty in evaluating feature selection methods due to the lack of ground truth in biological data.

These papers highlight recent advancements in computational models for gene selection using machine learning techniques within the past five years. However, each approach comes with its own set of limitations, ranging from computational complexity and scalability issues to challenges in interpretability and data requirements.

METHODOLOGY

In recent years, there has been a notable surge in researchers' interest in genomics and gene expression, with Machine Learning, a subset of Artificial Intelligence, emerging as a focal point. Machine Learning, integral to Data Science, is designed to empower models to train and autonomously make decisions in the future. It is commonly classified into Supervised, Unsupervised, and Semi-supervised or Semi-unsupervised learning paradigms. Supervised learning utilizes labeled data, while unsupervised learning operates on unlabeled data. Semi-supervised or Semi-unsupervised learning, on the other hand, addresses both labeled and unlabeled data.

In the context of gene expression microarray data, Machine Learning follows a workflow encompassing Pre-processing and Classification or Clustering stages. Machine learning-based feature selection methods, such as gene selection approaches, play a crucial role in isolating essential genes from the vast pool of data. These techniques aid in distilling pertinent information for subsequent analysis and interpretation.

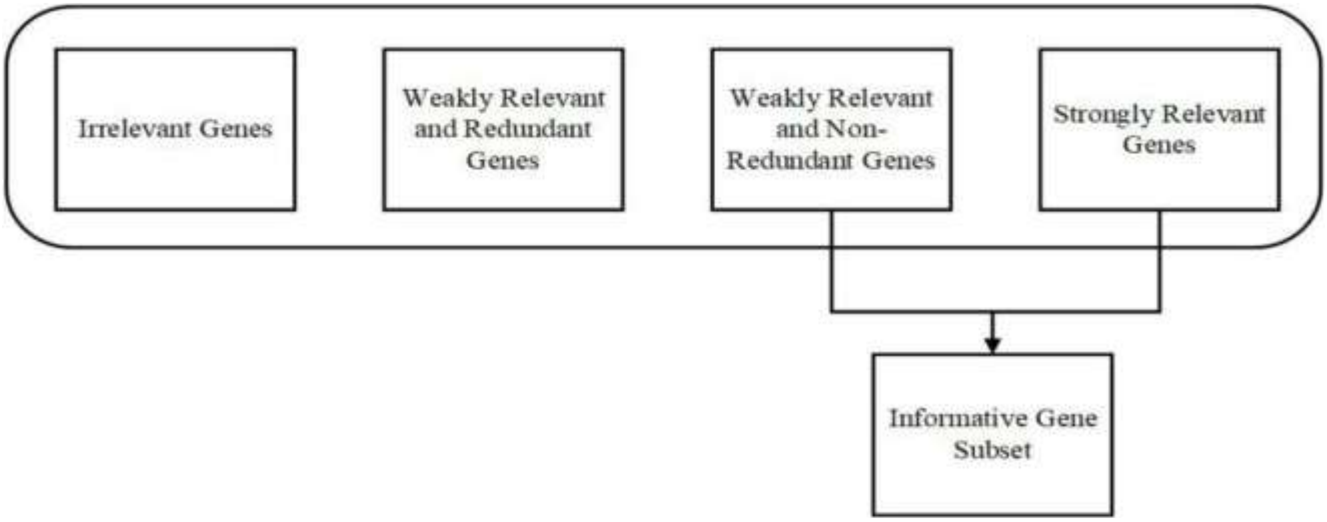
Irrelevant Genes
Weakly Relevant and Redundant Genes
Weakly Relevant and Non-Redundant Genes
Strongly Relevant Genes
Informative Gene Subset

Fig 1. Gene Selection Representation.

The Gene Selection based on machine learning can be classified into three types, Supervised, Unsupervised, and Semi-Supervised. Supervised Gene Selection utilizes the genes that are labelled already the input and output labels are knownin advance in this method. However, the data continues to grow and overwhelm the process, leading to data mislabeling, making it unreliable. The main issue in deploying Supervised Gene Selection is overfitting, which can be caused by selecting irrelevant or sometimes eliminating the most relevant gene [8].

The Gene Selection based on machine learning can be classified into three types, Supervised, Unsupervised, and Semi-Supervised. Supervised Gene Selection utilizes the genes that are labeled already [6] . The input and output labels are known in advance in this method. However, the data continues to grow and overwhelm the process, leading to data mislabeling, making it unreliable. The main issue in deploying Supervised Gene Selection is overfitting, which can be causedby selecting irrelevant or sometimes eliminating the most relevant gene [12].

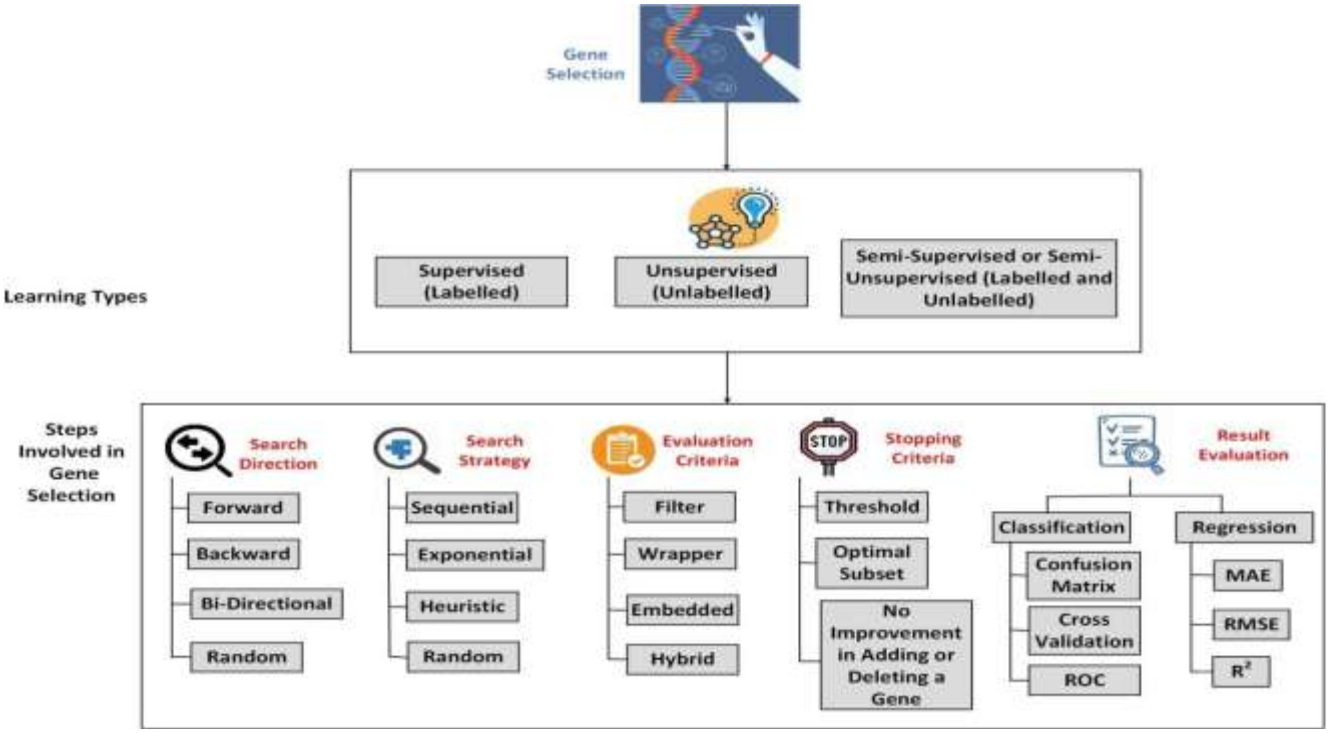

Fig 2 .Representation of different methods for gene selection

Unsupervised Gene Selection, unlike Supervised, will not have any labels to guide the selection process. The data used in Unsupervised Gene Selection is unlabelled. That makes it unbiased and serves as an effective way to find the necessary insights into the classification process The main issue in Unsupervised Gene Selection is that it does not consider the interaction among the Genes (correlation), making the resultant gene subset insignificant in the discrimination task.

Semi-supervised or Semi-unsupervised Gene Selection is like an add-on to the Supervised and Unsupervised Gene Selection. A Gene Selection is considered semi-supervised when most of the data is labeled, and a Gene Selection is said to be Semi-unsupervised when most of the data is unlabelled. The labeled data in the Semi-supervised or unsupervised is usedto increase the distance between the data points that belongs to different classes, whereas the unlabelled data will help identify the geometrical structure of the feature space .

CHALLENGES

Gene selection, also known as feature selection in the context of machine learning, involves choosing a subset of relevant genes from a larger pool for analysis or modeling. While it's a crucial step in various biological and computational applications, it comes with its own set of challenges:

- **High Dimensionality**: Gene expression data often involve a large number of genes, leading to a high-dimensional feature space. This can lead to computational inefficiency and increased risk of overfitting.

- **Noise and Redundancy**: Gene expression data may contain noise and redundancy, where some genes might not contribute significantly to the predictive power or might be highly correlated with other genes. Identifying and removing such noise and redundancy is a challenge.

- **Curse of Dimensionality**: With a high-dimensional feature space, the curse of dimensionality becomes a significant concern. As the number of features increases, the amount of data needed to effectively cover the feature space grows exponentially, making it difficult to find meaningful patterns.

- **Biological Complexity**: Gene-gene interactions and non-linear relationships among genes add to the complexity of gene selection. Traditional feature selection methods might not capture these interactions effectively.

- **Data Imbalance**: In biological datasets, there might be an imbalance in the distribution of samples across different classes or conditions. This can bias feature selection towards genes that are overrepresented in the majority class.

- **Dynamic Gene Expression**: Gene expression patterns can vary across different conditions, time points, or tissues. Static feature selection methods might not capture this dynamic nature effectively.

- **Evaluation Metrics**: Choosing appropriate evaluation metrics for gene selection is non-trivial. Metrics such as accuracy, precision, and recall might not fully capture the biological relevance of selected genes.

CONCLUSION

In conclusion, the paper presents an overview of the challenges and methodologies involved in gene selection using machine learning models. Through an extensive review of the literature, we have identified key issues such as high dimensionality, noise and redundancy in data, biological complexity, and the need for interpretable and robust methods. Our analysis underscores the importance of addressing these challenges to effectively identify relevant genes for various biological and clinical applications. We have highlighted the significance of integrating domain knowledge with advanced machine learning techniques to navigate the complexities of gene expression data. In conclusion, this paper contributes to the advancement of machine learning methods for gene selection by providing insights into current approaches, identifying challenges, and outlining future directions. By leveraging these methodologies, we can unlock the potential of genomic data to drive discoveries in biology, medicine, and personalized healthcare.

This comprehensive survey delves into the realm of computational models for gene selection, focusing on the utilization of machine learning techniques. Through a meticulous examination of recent research spanning the last five years, we have uncovered a diverse array of methodologies aimed at extracting biologically relevant information from high- dimensional genomic data.

REFERENCES

[1] https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2020.603808/full

[2] https://cancerres.aacrjournals.org/content/65/1/291

[3] https://www.ebi.ac.uk/training/online/course/network-analysis-protein-interaction-data-

[4] https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-020-00771-4

[5] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, charity W law, Wei Shi, and Gordon K Smyth. Limma powers differential expression analyses for rna-sequencing and microarray studies. Nucleic Acids Res, 43(7):e47–e47, 2015.

[6] Raut et al., 2010 ; Wang and van der Laan, 2011

[7] Filippone et al., 2006

[8] Abdulla, M., and Khasawneh, M. T. (2020). G-Forest: an ensemble method for costsensitive feature selection in gene expression microarrays.

[9] Artif. Intell. Med. 108:101941. doi: 10.1016/j.artmed.2020.101941 Abinash, M. J., and Vasudevan, V. (2018).

[10] "A Study on Wrapper-Based Feature Selection Algorithm for Leukemia Dataset," in Proceedings of the Intelligent Engineering Informatics, (New York, NY: Springer), doi: 10.1007/978- 981-10-7566-7_31

[11] Acharya, S., Saha, S., and Nikhil, N. (2017). Unsupervised gene selection using biological knowledge: application in sample clustering. BMC Bioinform. 18:513. doi: 10.1186/s12859-017-1933-0

[12] Algamal, Z. Y., and Lee, M. H. (2015). Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Exp. Syst. Appl. 42, 9326–9332. doi: 10.1016/j.eswa.2015. 08.016

[13] Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., et al. (2014). Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinform. 15:49. doi: 10.1186/1471-2105-15-8

[14] https://www.sciencedirect.com/science/article/abs/pii/S0957417422019649