EPiC Series in Computing

Volume 92, 2023, Pages 11–24

Proceedings of International Conference on Bioinformatics and Computational Biology (BICOB-2023)

Biomarker discovery in multi-omics datasets using tensor decompositions; A comprehensive review

Farnoosh Koleini[1,1] and Paul J. Gemperline[2,1] and Nasseh Tabrizi[3,1*]

[1] East Carolina University, Greenville, NC, United States

Koleinif20@students.ecu.edu, gemperlinep@ecu.edu, tabrizim@ecu.edu

Abstract

A multi-omics dataset combining clinical features with the discovery of biomarkers could contribute significantly to the timely identification of mortality risk and the development of personalized therapies for a wide range of diseases, including cancer and stroke. As well, new advances in ‘‘omics’’ technologies can open up a lot of possibilities for researchers to find disease biomarkers through system-level analysis. Machine learning methods, especially based on tensor decomposition methods (TD-based), are becoming more popular because the integrative analysis of multi-omics data is challenging due to biological complexity. Therefore, it is important to identify future research directions and opportunities on the topic of biomarker discovery using tensor decompositions in multi-omics datasets by integrating literature reviews. This article systematically reviews the research trends from 2015 to 2022. Several themes are discussed, including challenges and problems of developing and applying tensor decompactions, application areas for biomarker discovery in ‘‘omics’’ datasets, proposed methodologies, key evaluation criteria used in deciding whether the new methods are effective, and the limitations and shortcomings of this field, which call for further research and development. This review helps researchers who are interested in this field understand what research has already been done and where potential areas for future research might lie.

I. INTRODUCTION

Biomarkers are biological molecules that are indicative of normal or abnormal processes, such as disease states or responses to treatments. These biological molecules may be found in any type of organism by sampling tissue and body fluids followed by biochemical analysis. The development of high throughput methods has facilitated an explosion of research in this field. When combined with clinical data, the resulting information can be used for earlier detection of diseases and the development of personalized therapies. Moreover, new developments in ‘‘omics’’ technology provide researchers the chance to look for disease biomarkers at the system level [1]. A Tremendous amount of work has gone into discovering disease-associated biomolecules by analyzing data obtained from different ‘‘omics’’ experiments (genomics, transcriptomics, metabolomics). However, due to the complexity of biological systems and the poor integration of various forms of ‘‘omics’’ data, integrative analysis of multi-omics data is a difficult undertaking. Various feature selection procedures have been shown to

provide different sets of biomarkers [2]. A classic approach to biomarker selection comprises statistical approaches such as the Student's t-test and ANOVA, which find and choose biomolecules with a significant change in expression level between separate biological groups (normal vs. disease; untreated vs. treated). One clear disadvantage of these methods is that they ignore the fact that biomolecules in a biological system are densely interconnected and interact with one another. Integrated analysis using tensor decompositions of data from many sources has recently demonstrated the ability to improve knowledge discovery. In metabolomics, for example, biological fluids such as blood or urine are examined using various analytical techniques to find molecules associated with specific diseases or diets [3]. A joint factorization problem has been developed for the topic of data fusion [3]. Data from many sources can be represented as several matrices, which can then be evaluated jointly using tensor decomposition methods. The tensor factorization has also been found to be effective in other domains, including social network analysis [4-8], signal processing [9,10], and bioinformatics [11-13]. Also, coupled tensor decomposition methods have been developed and employed in chemometrics [14], bioinformatics [11,12], signal processing [9,15,16], and data mining [17,18]. With the introduction of high throughput technology capable of extensive analysis of genes, transcripts, proteins, and other significant biological molecules, the identification of molecular markers of disease processes has become a reality on a scale never before seen. It has, however, made it more difficult to extract relevant molecular markers of biological processes from these complex datasets. The process of biomarker discovery and characterization allows for more sophisticated approaches to integrating purely statistical and expert knowledge-based approaches, and tensor decompositions provide a great opportunity to aid in the interpretation of such interactions and the identification of reliable biomarkers [19]. There are several review papers on biomarker discovery using tensor decompositions published in the last few years [20], [21].

This paper reviews research in this area from 2015 to 2022 to provide useful insights into the recent advances in biomarker discovery using tensor decompositions and suggests future research directions. The challenges, drawbacks, and new opportunities that have arisen due to the availability of more multi-omics data and information have called for studies on developing tensor decomposition methods to detect biomarkers in recent years. Figure 1 shows the number of publications that use tensor decompositions for biomarker detection or deal with biomarker discovery challenges published between 2015 and early 2022.

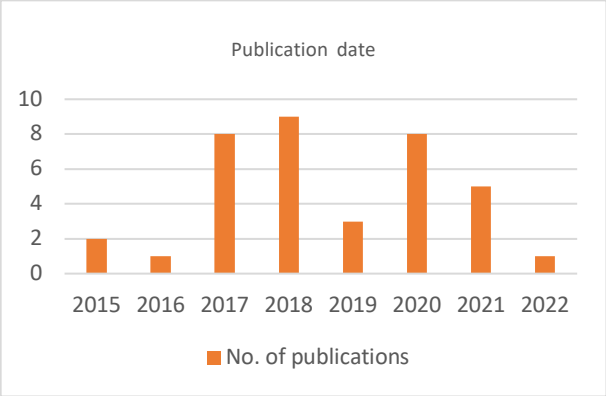

Fig.1. Number of papers from 2015 to 2022

II. SYSTEMATIC REVIEW

The first step in this systematic review was to define the goals of the survey. These goals are described as follows.

- Identifying the problems and challenges regarding the biomarker discovery in multi-omics datasets by tensor decompositions.
- Identifying algorithms and methodologies employed to solve these problems and their challenges.
- Identifying areas of application for biomarker discovery in multi-omics datasets.
- Identifying evaluation criteria used to evaluate developed TD-based methods.

In this systematic review, we first searched the literature for publications using scientific search engines and collected databases of publications. The search query used was (“biomarker” AND “discovery”) AND (“multi-omics”) AND (“tensor” AND “decompositions”). This search query was used on several databases including IEEE Xplore, ACM Digital Library, Lynda.com, ScienceDirect, and SpringerLink. Then, the selected publications were studied, and the information was used to answer the main questions of this systematic review.

III. INTEGRATING MULTI-OMICS DATASETS: OPPORTUNITIES AND CHALLENGES

Developing computational models to discover potential biomarker-disease connections in multi-omics data, which could provide insight into disease pathophysiology and improve illness diagnostic and prognostic accuracy, is gaining popularity. The recent introduction of effective and low-cost screening technologies has resulted in massive amounts of biological data, paving the door for a new era of treatments and customized medicine [22,23]. Clinical information and ‘‘omics’’ data can be acquired directly from databases or collected through screening technologies for disease [24], class prediction [25], biomarker identification [26], disease subtyping [24], better system biology understanding [27], drug repurposing, and other applications. Each ‘‘omics’’ data type is specific to a single "layer" of biological information, such as genomics, epigenomics, transcriptomics, proteomics, or metabolomics, and provides a complementary medical perspective of a biological system or an individual [22].

1) *Integration analysis of multi-omics datasets:*

Single-omics investigations were previously conducted to discover the causes of diseases to help design or pick a suitable treatment. Most diseases, on the other hand, involve complicated molecular pathways in which distinct biological layers interact with one another. Therefore, there is a greater demand for biomarker discovery in multi-omics investigations that can incorporate several layers and provide a fuller picture of a particular phenotype [28]. Faint patterns in gene expression data can be enhanced by several ‘‘omics’’ [29]. For example, complementary information can be exploited to better explain classification results [30], improve prediction performances [31,32], or comprehend complex molecular pathways [33]. Multi-omics studies, on the other hand, comprise data of varying type, scale, and distribution, with thousands of variables and only a few samples. Furthermore, biological datasets are complicated and noisy, with the possibility of errors due to measurement errors or unique biological variances. Finding relevant information and incorporating the ‘‘omics’’ into a useful model is difficult, and several methods and tactics have been developed in recent years to address this difficulty [24,34]. As a result, researchers are seeking approaches that, by adding additional ‘‘omics’’ data, result in an increase in performance rather than simply increasing the complexity and processing time of the task.

2) *Challenges of multi-omics datasets:*

When integrating multi-omics datasets, several obstacles occur. Some of these, such as the existence of missing values or class imbalance, are general to machine learning analysis. When working on rare events, such as an uncommon attribute in a population, class imbalance occurs when the distribution of

classes in the learning data is biased. This problem can be solved using a variety of strategies, including sampling and cost-sensitive learning. Sampling tries to balance the dataset before the integration process, where either the majority class is randomly under-sampled, or the minority class is oversampled by creating new artificial observations, or a combination of both methods. Cost-sensitive learning is directly integrated into the algorithm and balances the learning process by giving more weight to misclassified minority observations [35,36]. Some are more specific and include the noisiness and complexity of ‘‘omics’’ datasets, which naturally occur in biological data. Relevant patterns can occasionally be obscure and involve a large number of molecules from various ‘‘omics’’ layers. Therefore, identifying those patterns across numerous datasets is a challenging endeavor. Furthermore, due to financial constraints, the rarity of the desired phenotype, a lack of willing volunteers, etc., the collection of substantial volumes of biomedical data is frequently only possible on a small sample of patients when conducting ‘‘omics’’ or multi-omics investigations. This results in datasets with several variables greatly exceeding the number of samples. Machine learning algorithms have a propensity to overfit these high-dimensional datasets, which reduces their generalizability to new data. This problem is known as the ‘‘curse of dimensionality’’ [36]. Another difficulty is their heterogeneity, which must be handled properly because various ‘‘omics’’ methodologies may provide data with varying distributions of types (e.g., numerical, categorical, continuous, discrete, etc.). Furthermore, ‘‘omics’’ datasets can vary greatly in size (number of features), with a typical gene expression dataset having tens of thousands of variables and a metabolomics dataset having only a few thousand. Disparities between ‘‘omics’’ datasets might impede integration and create an imbalance in the learning process [37]. Scalability is an additional technical issue regarding multi-omics datasets. The scope of genomics research has been broadened from a narrow single-layer examination to a comprehensive multi-dimensional interpretation of biological data as a result of the accessibility of these massive multidimensional and heterogeneous datasets. To create rich, multi-scale characterizations of biological systems, the emphasis is on combining various forms of omics data from many layers of biological regulation. However, it necessitates systems that can scale across heterogeneous datasets while also centralizing data processing analysis, and interpretation inside a unified inference framework [38-40]. Therefore, developing a quick and effective method that can compute tensor decompositions of larger quantities of data would lead to more effective biomarker discovery in multi-omics datasets. Figure 2 shows the number of papers that deal with specific problems in biomarker discovery in multi-omics datasets: the curse of dimensionality, scalability, and noisiness problems. Typically, these papers describe the development of procedures that perform better. While these issues are still being researched to improve biomarker identification in multi-omics datasets, data heterogeneity necessitates a more thorough investigation.

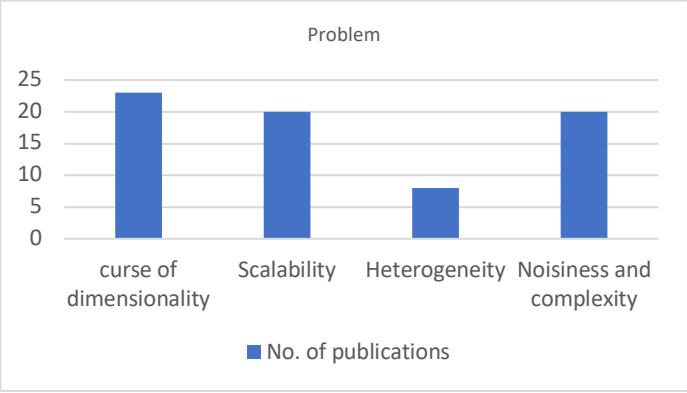

Fig.2. Number of studied dealing with a specific problem in biomarker discovery in multi-omics dataset

IV. METHODOLOGIES

In this section, we review the approaches and methodologies that are used in the literature on biomarker discovery in multi-omics datasets by tensor decompositions.

A. Parallel Factor Analysis

Parallel Factor Analysis (PARAFAC) is a popular tensor decomposition method that is widely used in biomarker discovery in multi-omics datasets. It is a method for decomposing multidimensional arrays to focus on the aspects of interest and provide a clear illustration of the results. PARAFAC is based on a mathematical model that depicts the interactions of the dimensions to be evaluated in the input data. The analysis dimensions must be defined before performing PARAFAC. Each input value can then be related to an index for each of the dimensions. Assuming N=3 dimensions, for example, x_{ijk} identifies the measured value for index i in the first dimension, j in the second dimension, and k in the third dimension. Equation (1) represents the PARAFAC model, where f denotes the number of so-called components and defined so-called loading matrices **A**, **B**, and **C** of dimensions $I \times F$, $J \times F$, and $K \times F$ and with elements a_{if}, b_{jf}, and c_{kf}, respectively, and the modeling error ε_{ijk}.

$$x_{ijk} = \sum a_{if}\, b_{jf} c_{kf} + \varepsilon_{ijk} \qquad (1)$$

Reference [41] then provides the generic model that PARAFAC uses to represent the input data. A graphical illustration of this model is given in Figure 3. The data is decomposed into triads or trilinear components, where each component comprises one score vector and two loading vectors rather than one score vector and one loading vector as in bilinear PCA. It is the standard three-way procedure to consider scores and loadings numerically similarly, without making any distinction between the two. A well-established advantage of the PARAFAC model is the mathematical uniqueness of the solution. Unique solutions can be expected if the loading vectors are linearly independent in two of the modes and the third mode, and that no two loading vectors are linearly dependent in the third mode.

PARAFAC applications:

Zhang et al defined ‘‘a temporal and spatial feature similarity measure to calculate the rate of change and velocity of each biomarker in MRI to form a vector that represents the morphological change of the biomarker, then calculating the similarity of the changing trend between two biomarkers to encode the data in a third-order tensor to extract interpretable biomarker latent factors from the original data using PARAFAC decomposition.’’ [42].

Jung et al proposed ‘‘a multi-omics analysis method called MONTI (Multi-omics Non-negative Tensor decomposition for Integrative analysis), that selects multi-omics features that can represent trait-specific characteristics.’’ They provided the usefulness of multi-omics integrated analysis for cancer subtyping. The multi-omics data were first merged in a biologically meaningful way to generate a three-dimensional tensor, which was then decomposed using the PARAFAC method. MONTI was then utilized to identify highly informative subtype-specific multi-omics features [43].

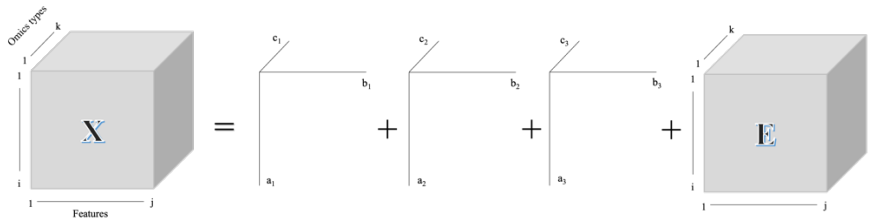

Figure3. A graphical illustration of the PARAFAC model.

B. Tucker3

Tucker3 is another tensor decomposition method used in multi-omics datasets to detect biomarkers. The Tucker3 model name is taken from psychometrician Ledyard R. Tucker who in 1966 proposed the model. He also presented a method for calculating the model's parameters, and several changes to the algorithmic approach have subsequently been suggested. The model has remained a powerful tool for analyzing three-way (and higher way) data arrays. The Tucker3 model is frequently used for decomposition, compression, and interpretation in many applications because of its generality and the way it treats the PARAFAC model as a particular instance. The Tucker3 model can be seen as an extension of the PARAFAC-CANDECOMP model along the line of outer products. Kroonenberg provided ‘‘a full mathematical description of this model as well as advanced topics such as data preparation/scaling and core rotation. Different numbers of factors in each of the modes can be extracted using the Tucker3 model [44].’’ Figure 4 is used to provide a simple explanation of the model.

Tucker3 applications:

Taguchi has focused on post-traumatic stress disorder (PTSD), a mental condition that can cause extra symptoms that do not appear to be immediately related to the central nervous system, which is thought to be directly affected by PTSD. PTSD-mediated heart disease is one such secondary disorder [45]. The spatial separation between the heart and the brain hindered researchers from clarifying the mechanisms that link the two disorders, despite the strong associations between PTSD and heart diseases. Their goal was to discover the genes that link cardiac problems with PTSD. To execute gene selection, they employed Tucker3 factorization as the tensor decomposition method to examine the gene expression profiles in diverse tissues, such as the heart and brain. The gene expression profiles were regarded as tensors. Gene expression profiles in diverse tissues were studied under various conditions such as stressful or unstressful, with varying periods of stress and rest time following the application of a stressor. Approximately 400 genes were identified as potential genes that may mediate heart problems related to PTSD based on the obtained features. Additionally, before being applied to gene expression profiles, Tucker3 was applied to a synthetic data set to illustrate the utility of their technique [45,46].

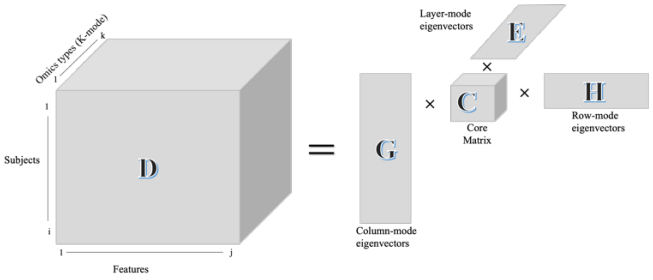

Fig.4. A graphical illustration of the Tucker3 model.

C. *Hybrid and other techniques*

Feature extraction methods are a class of techniques that try to turn a set of input biomarkers into another set of variables that are linear or non-linear combinations of the original biomarkers. The goal is to extract features in such a way that the resulting new variables retain useful information while being less noisy and less redundant. Learning from a smaller set of features or biomarkers reduces complexity while increasing computational efficiency. The interpretability of a model may be compromised by feature extraction methods since the derived features are no longer biological measurements. Feature extraction methods are frequently employed experimentally to visualize data and uncover significant features.

Principal Component Analysis (PCA) is the most extensively used feature extraction approach. [47] PCA creates new variables called principal components, which are uncorrelated linear combinations of the original features and optimize the description of variance in the dataset; however, it is sensitive to outliers and is unable to describe non-linear trends in the data. To address these issues, several

extensions have been developed, such as Kernel PCA [48] and Bayesian PCA [49]. Other similar methods such as Principal Coordinates Analysis (PCoA) [50], Correspondence Analysis (CA) [51], and Independent Component Analysis (ICA) [52] may improve PCA in certain ways. The majority of feature extraction techniques have also been developed with sparsity constraints. Sparse feature extraction methods can be used for feature selection with methods such as Sparse PCA (sPCA) [53], Sparse Canonical Correlation Analysis (CCA) [54], Sparse Non-negative Matrix Factorization (Sparse NMF) [55], and Sparse CA [56]. These approaches, however, fail to examine multi-omics datasets since applying them to concatenated ‘‘omics’’ typically yields unsatisfactory results. As a result, feature extraction methods are frequently used on each ‘‘omics’’ dataset for either block scaling or after concatenation of the extracted features or clustering, or other downstream analysis [38].

D. AI for biomarker discovery in multi-omics datasets

Gene regulatory networks, which are critical for understanding complicated disease mechanisms, have become one of the most popular topics for biomarker identification in multi-omics datasets in recent years. Several large-scale projects have been done and significant amounts of ‘‘omics’’ data have been released to identify heterogeneous genetic networks that underlie complex human diseases. The gene networks scale is increasing, and methodologies for analyzing large-scale gene networks have been proposed. Park et al. proposed a novel AI technique for analyzing gene regulation networks in depth. The multilayer networks were decomposed using an AI technique based on deep learning to identify all-encompassing gene regulatory systems distinguished by patient clinical features. They extracted global and unique mechanisms of gene regulatory systems from the vast multiple networks using an AI technique based on tensor decomposition. They developed a novel technique to do integrative analysis for multilayer gene networks, which is an essential tool for precision medicine. In their method, gene regulatory networks were built under varied sample conditions, and the multilayer networks were thoroughly examined using an AI algorithm. To construct a low-dimensional subspace of the multiway interaction between genes, a deep learning algorithm for tensor decomposition was applied to the gene network for a target sample. They were able to understand the constructed large-scale gene networks since prediction and interpretation were carried out on the constructed low-dimensional subspace. Their technique is divided into two stages: building sample-specific gene regulatory networks and globally analyzing large-scale multiple gene networks using AI technology [57,58].

V. APPLICATIONS

Discovering biomarkers has various uses in the healthcare system, such as early disease detection, disease prevention, identifying an individual's risk, monitoring disease, and drug development in the pharmaceutical sector. Therefore, biomarker discovery, specifically in multi-omics datasets by tensor decompositions could help a lot to develop biomarker applications. In this section, we will cover some of the important applications of biomarkers in the literature.

A. Early disease detection, prevention, and monitoring

Measures for the early detection of various diseases such as different cancers, and stroke, offer the opportunity to help control rising healthcare costs. We can already see that alternative disease prevention strategies will be used in the future because these strategies can and should be tailored to each patient based on their unique risk profiles. Fortunately, biomarkers make it possible to detect diseases such as Alzheimer’s, and certain cancers at a disease stage even when the patient shows no symptoms. The recent failures of potential medications that are tailored for various conditions may be an indication that the clinical trial participants are too far along to benefit clinically. Therefore, the development of new therapeutics will be greatly influenced by validated biomarkers for the early detection and precise diagnosis of diseases in their preclinical phases. When biomarkers are used synthetically, they may someday be able to identify patients in the initial stages of the disease, when therapeutic modification is most likely possible. Because whether medicine is likely to work can frequently be a genetic issue, biomarkers are also important in the development of individualized

treatment. As a result, determining or excluding specific genetic variations can make a significant contribution to therapeutic management, not only reducing costs and side effects but also improving treatment quality. Biomarkers can also be used to track treatment response [42,59,60].

B. Risk assessment

Biomarkers can be classified into susceptibility, effect, and exposure indicators. It is commonly expected that current developments in genomics, proteomics, and metabolomics will eventually translate into a constellation of advantages for human health. However, only a few biomarkers have been reported in the past ten years for risk assessment using "omics" technologies. But there is a wide range of potential applications for "omics" technology. The lack of integrated bioinformatics techniques, statistical analysis, and predictive models frequently severely restricts the use of biomarker-based monitoring systems as a tool for environmental risk assessment. Therefore, identifying pertinent and reliable biomarkers that contribute to the assessment of environmental and health risks may be necessary [61,62].

C. Drug discovery and development

Biomarkers that are robust and verified are required to improve diagnosis, monitor drug activity, and therapeutic response, and lead the development of safer and more tailored therapeutics for a variety of diseases. The development of specialized biomarkers for complicated chronic diseases can now be discovered and developed more quickly thanks to recent developments in multi-omics techniques, bioinformatics, and biostatistics. Even though there are still many obstacles to overcome, current initiatives for the discovery and development of disease-related biomarkers will help with the best decision-making during the medication development process and further our comprehension of the disease processes. To the benefit of patients, healthcare professionals, and the biopharmaceutical industry, good preclinical biomarker translation into the clinic will pave the path for the effective execution of personalized therapies across a range of complex disease areas [63,64]. Figure 5 illustrates the distribution of the all studies focusing on each area of application.

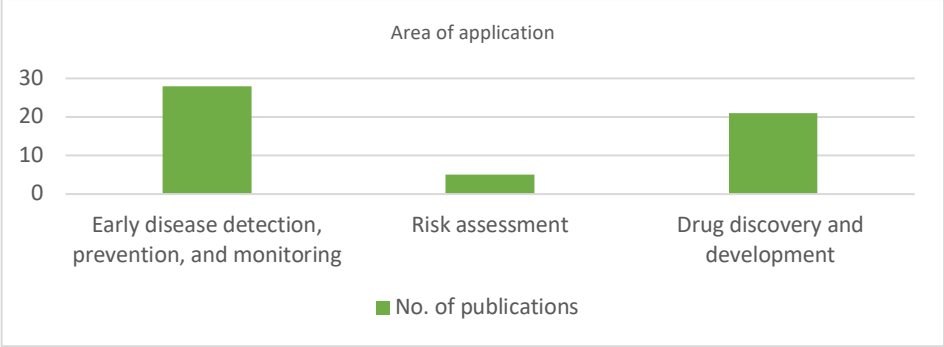

Fig.5. Areas of application for biomarker discovery in multi-omics datasets

VI. EVALUATION CRITERIA

In terms of evaluation, there are several metrics available. These metrics include and are not limited to the residual sum of squares, precision, and *f*-scores. The residual sum of squares is the sum of the squares of residuals (deviation of predicted from actual empirical values of data). It serves as a gauge for the disparity between data and the estimation model. One measure of precision is the proportion of correctly selected biomarkers over the whole set of biomarkers. The recall is calculated as the ratio of the number of correctly selected biomarkers to the total number of test biomarkers. The *f*-score is obtained using a harmonic mean between recall and precision. Other metrics, which will be introduced based on the nature of the problem and the proposed model, can be established, and used to analyze the

success of biomarker identification approaches employing tensor decompositions in multi-omics datasets. We describe the criteria used in the surveyed papers as follows.

- Root Mean Square Error (RMSE) can be formulated as shown in Equation (2), where t_i is the test rating value and p_i is the predicted rating value [38,65].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - p_i)^2} \quad (2)$$

- Residuals Sum of Squares (RSS) is the measure of the discrepancy between the data and the estimated model. The important point in tensor decompositions is that the trilinear model is found to minimize the RSS. Equation 3 shows this metric, where y_i is the i value of the biomarker to be predicted, and $f(x_i)$ is the predicted value of y_i [43].

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (3)$$

- Precision (p) is the proportion of the relevant features among all retrieved feature sets and assesses the predictive power of a method. Precision can be formulated as follows where $t\mathrm{p}$ is the true positive, and $f\mathrm{p}$ is the false-positive selected cases [43,66].

$$p = \frac{t\mathrm{p}}{t\mathrm{p}+f\mathrm{p}} \quad (4)$$

- Recall (r) calculates the proportion of the selected features as part of the optimal feature set relative to all features and assesses the effectiveness of an algorithm in identifying the true positive features. Recall can be formulated as follows where $t\mathrm{p}$ is the true positive, and $f\mathrm{n}$ is the false negative in selected cases [66].

$$r = \frac{t\mathrm{p}}{t\mathrm{p}+f\mathrm{n}} \quad (5)$$

- The *f*-score which utilizes precision (p) and recall (r) can be defined as follows. Recall and precision are balanced in the *f*-score when the β constant parameter is set to 1 and is in favor of precision when $\beta > 1$ [43,66].

$$f = \frac{(\beta^2+1)pr}{(\beta^2 p)+r} \quad (6)$$

- *P*-values are a commonly used criterion used for ranking biomarker candidates and determining the top set of markers considered for further development and validation. Thus, statistical *P*-values can play a fundamental role in the evaluation of biomarker discovery studies. In the case of control studies, the *P*-value associated with a statistic is defined as follows: [63,67]

$$P - value = Probability\,(statistic \geq observed\ data\ statistic \mid cases\ same\ as\ controls)$$

- Sensitivity and Specificity are two other measures that evaluate the diagnostic performance of a biomarker. Sensitivity is the ability to detect a disease in patients in whom the disease is truly present (i.e., a true positive), and specificity is the ability to rule out the diseases in patients in whom the disease is truly absent (i.e., a true negative) [66,68,69].

- Computation time and cost for biomarker detection in multi-omics datasets is an important evaluation criterion, especially where the problem requires a real-time application or there is a large amount of data for the computation [70].

Figure 6 shows the distribution of evaluation criteria used in the reviewed papers. The top 2 criteria are RSS and Precision. However, the majority of the papers used a combination of criteria to enhance their performance evaluation.

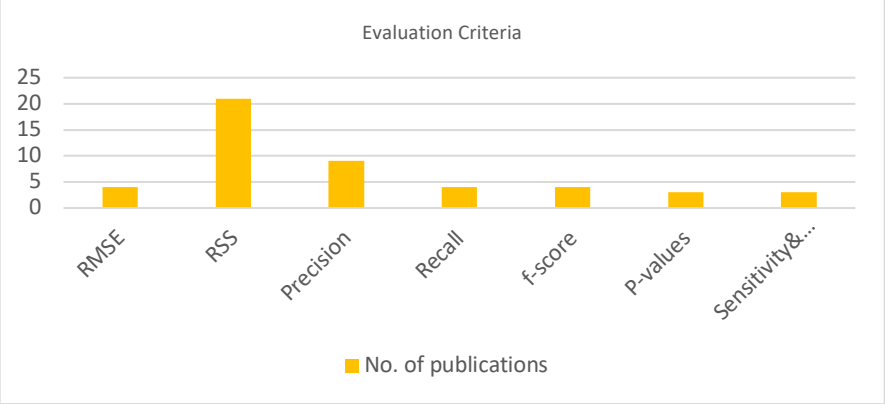

Fig.6. Evaluation Criteria

VII. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

There are many directions for future research. Dealing with long computation times and the associated costs is one of the most significant issues. Depending on the application, a large quantity of data may need to be evaluated in order to design a TD-based strategy for biomarker discovery. The majority of research in the literature focus on the development of solutions for difficulties such as interpretability and scalability; however, they rarely focus on the efficiency of the models. Proposing and developing TD-based strategies for dealing with massive amounts of data from various ‘‘omics’’ types is a research area that has not been extensively investigated. Another issue is that the recommended solutions are often developed for a specific application area. Future research directions might find it interesting to offer a framework that encompasses several application areas. Recently, some researches have used neural networks and machine learning to find biomarkers in multi-omics datasets. On this subject, several machine learning and deep learning models may be developed, and their performance can be compared to that of conventional approaches [57,58]. Additionally, different machine learning and deep learning models in terms of chemometrics can be developed in biomarker discovery in multi-omics datasets by tensor decompositions.

REFERENCES

[1] Z. Fan, Y. Zhou, and H. W. Ressom, "MOTA: Multi-omic integrative analysis for biomarker discovery," in Jul 2019, Available: https://ieeexplore.ieee.org/document/8857049. DOI: 10.1109/EMBC.2019.8857049.

[2] E. G. Armitage and C. Barbas, "Metabolomics in cancer biomarker discovery: current trends and future perspectives," *J. Pharm. Biomed. Anal.,* vol. 87, pp. 1-11, 2014.
Available: http://europepmc.org/abstract/MED/24091079 https://doi.org/10.1016/j.jpba.2013.08.041. DOI: 10.1016/j.jpba.2013.08.041.

[3] E. Acar, R. Br,o and A. K. Smilde, "Data Fusion in Metabolomics Using Coupled Matrix and Tensor Factorizations," *Jproc,* vol. 103, *(9),* pp. 1602-1620, 2015. Available: https://ieeexplore.ieee.org/document/7202834. DOI: 10.1109/JPROC.2015.2438719.
[4] W. Ma *et al*, "Local probabilistic matrix factorization for a personal recommendation," in Dec 2017, Available: https://ieeexplore.ieee.org/document/8288451. DOI: 10.1109/CIS.2017.00029.
[5] Y. Lin *et al*, "Community Discovery via Metagraph Factorization," *ACM Transactions on Knowledge Discovery from Data,* vol. 5, *(3),* pp. 1-44, 2011. Available: http://dl.acm.org/citation.cfm?id=1993081. DOI: 10.1145/1993077.1993081.
[6] H. Shuang *et al*, "Like like alike -Joint Friendship and Interest Propagation in Social Networks Hongyuan Zha".
[7] H. Mohammadi and V. Marojevic, "Artificial neuronal networks for empowering radio transceivers: Opportunities and challenges," in 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall). IEEE, 2021, pp. 1–5.
[8] B. Ermis, A. T. Cemgil, and E. Acar, "Generalized coupled symmetric tensor factorization for link prediction," in Apr 2013, Available: https://ieeexplore.ieee.org/document/6531411. DOI: 10.1109/SIU.2013.6531411.
[9] Jiho Yoo *et al*, "Nonnegative matrix partial co-factorization for drum source separation," in Mar 2010, Available: https://ieeexplore.ieee.org/document/5495305. DOI: 10.1109/ICASSP.2010.5495305.
[10] B. Ermiş, E. Acar and A. T. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Min Knowl Disc,* vol. 29, *(1),* pp. 203-236, 2013. Available: https://link.springer.com/article/10.1007/s10618-013-0341-y. DOI: 10.1007/s10618-013-0341-y.
[11] O. Alter, P. O. Brown and D. Botstein, "Generalized Singular Value Decomposition for Comparative Analysis of Genome-Scale Expression Data Sets of Two Different Organisms," *Proceedings of the National Academy of Sciences - PNAS,* vol. 100, *(6),* pp. 3351-3356, 2003. Available: https://www.jstor.org/stable/3139363. DOI: 10.1073/pnas.0530258100.
[12] L. Badea, "Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* pp. 267, 2008. Available: https://www.ncbi.nlm.nih.gov/pubmed/18229692.
[13] E. Acar, G. E. Plopper and B. Yener, "Coupled analysis of in vitro and histology tissue samples to quantify structure-function relationship," *PLoS ONE,* vol. 7, *(3),* pp. e32227, 2012. Available: https://www.ncbi.nlm.nih.gov/pubmed/22479315. DOI: 10.1371/journal.pone.0032227.
[14] A. K. Smilde, J. A. Westerhuis, and S. de Jong, "A framework for sequential multiblock component methods," *Journal of Chemometrics,* vol. 17, *(6),* pp. 323-337, 2003. Available: https://api.istex.fr/ark:/67375/WNG-VKWV2690-G/fulltext.pdf. DOI: 10.1002/cem.811.
[15] A. Yeredor, "Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation," *Tsp,* vol. 50, *(7),* pp. 1545-1553, 2002. Available: https://ieeexplore.ieee.org/document/1011195. DOI: 10.1109/TSP.2002.1011195.
[16] A. Ziehe *et al*, "A Fast Algorithm for Joint Diagonalization with Non-orthogonal Transformations and its Application to Blind Source Separation Klaus-Robert Müller," *Journal of Machine Learning Research,* vol. 5, pp. 777, 2004.
[17] A. P. Singh and G. J. Gordon, "Relational Learning via Collective Matrix Factorization," *Kdd'08.* DOI: 10.21236/ada486804.
[18] B. Long *et al*, "Spectral clustering for multi-type relational data," on Jun 25, 2006, Available: http://dl.acm.org/citation.cfm?id=1143918. DOI: 10.1145/1143844.1143918.
[19] J. E. McDermott *et al*, "Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data," *Expert Opinion on Medical Diagnostics, 7(1):37-51,* vol. 7, *(1),* pp. 37-51, 2013. Available: https://www.tandfonline.com/doi/abs/10.1517/17530059.2012.718329. DOI: 10.1517/17530059.2012.718329.
[20] C. M. Ghantous *et al*, "Advances in Cardiovascular Biomarker Discovery," *Biomedicines,* vol. 8, *(12),* 2020. . DOI: 10.3390/biomedicines8120552.
[21] D. Ledesma, S. Symes, and S. Richards, "Advancements within Modern Machine Learning Methodology: Impacts and Prospects in Biomarker Discovery," *Curr. Med. Chem.,* vol. 28, 2021. . DOI: 10.2174/0929867328666210208111821.
[22] B. B. Misra *et al*, "Integrated omics: tools, advances, and future approaches," *Journal of Molecular Endocrinology,* vol. 62, *(1),* pp. R21, 2019. . DOI: 10.1530/jme-18-0055.
[23] Z. Ahmed, "Practicing precision medicine with intelligently integrative clinical and multi-omics data analysis," *Hum Genomics,* vol. 14, *(1),* 2020. . DOI: 10.1186/s40246-020-00287-z.
[24] O. Menyhárt and B. Győrffy, "Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis," *Computational and Structural Biotechnology Journal,* vol. 19, pp. 949-960, 2021. Available: https://dx.doi.org/10.1016/j.csbj.2021.01.009. DOI: 10.1016/j.csbj.2021.01.009.
[25] Y. Hasin, M. Seldin and A. Lusis, "Multi-omics approaches to disease," *Genome Biol,* vol. 18, *(1),* 2017. . DOI: 10.1186/s13059-017-1215-1.

[26] Y. V. Sun and Y. Hu, "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases," *Advances in Genetics,* vol. 93, pp. 147-190, 2016. Available: https://www.ncbi.nlm.nih.gov/pubmed/26915271. DOI: 10.1016/bs.adgen.2015.11.004.
[27] S. Dahal *et al*, "Synthesizing Systems Biology Knowledge from Omics Using Genome-Scale Models," *Proteomics,* vol. 20, *(17-18),* 2020. . DOI: 10.1002/pmic.201900282.
[28] J. Yan *et al*, "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data," *Briefings in Bioinformatics,* vol. 19, *(6),* pp. 1370-1381, 2018.
Available: https://www.ncbi.nlm.nih.gov/pubmed/28679163. DOI: 10.1093/bib/bbx066.
[29] M. Kang *et al*, "Integration of multi-omics data for integrative gene regulatory network inference," *International Journal of Data Mining and Bioinformatics,* vol. 18, *(3),* pp. 223-239, 2017.
Available: https://search.proquest.com/docview/1989915459. DOI: 10.1504/IJDMB.2017.10008266.
[30] N. Rappoport *et al*, "MONET: Multi-omic module discovery by omic selection," *PLoS Computational Biology,* vol. 16, *(9),* pp. e1008182, 2020. Available: https://search.proquest.com/docview/2451546964. DOI: 10.1371/journal.pcbi.1008182.
[31] H. Sharifi-Noghabi *et al*, "MOLI: multi-omics late integration with deep neural networks for drug response prediction," *Bioinformatics,* vol. 35, *(14),* pp. i501-i509, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/31510700. DOI: 10.1093/bioinformatics/btz318.
[32] G. Tini *et al*, "Multi-omics integration-a comparison of unsupervised clustering methodologies," *Briefings in Bioinformatics,* vol. 20, *(4),* pp. 1269-1279, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/29272335. DOI: 10.1093/bib/bbx167.
[33] M. Akhmedov *et al*, "OmicsNet: Integration of Multi-Omics Data using Path Analysis in Multilayer Networks," 2017.
Available: https://explore.openaire.eu/search/publication?articleId=sharebioRxiv::43bbef39d5491455c65759257b27a0a7. DOI: 10.1101/238766.
[34] S. Canzler *et al*, "Prospects and challenges of multi-omics data integration in toxicology," *Arch Toxicol,* vol. 94, *(2),* pp. 371-388, 2020. Available: https://link.springer.com/article/10.1007/s00204-020-02656-y. DOI: 10.1007/s00204-020-02656-y.
[35] M. Song *et al*, "A Review of Integrative Imputation for Multi-Omics Datasets," *Frontiers in Genetics,* vol. 11, pp. 570255, 2020. Available: https://search.proquest.com/docview/2461001785. DOI: 10.3389/fgene.2020.570255.
[36] B. Mirza *et al*, "Machine Learning and Integrative Analysis of Biomedical Big Data," *Genes,* vol. 10, *(2),* 2019. . DOI: 10.3390/genes10020087.
[37] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM,* pp. 78-87, 2012.
Available: http://dl.acm.org/citation.cfm?id=2347755. DOI: 10.1145/2347736.2347755.
[38] M. Picard *et al*, "Integration strategies of multi-omics data for machine learning analysis," *Computational and Structural Biotechnology Journal,* vol. 19, pp. 3735-3746, 2021. Available: https://dx.doi.org/10.1016/j.csbj.2021.06.030. DOI: 10.1016/j.csbj.2021.06.030.
[39] D. Choi, J. Jang and U. Kang, "S3CMTF: Fast, accurate, and scalable method for incomplete coupled matrix-tensor factorization," *PLoS ONE,* vol. 14, *(6),* pp. e0217316, 2019. Available: https://search.proquest.com/docview/2249031337. DOI: 10.1371/journal.pone.0217316.
[40] Kijung Shin, Lee Sael and U. Kang, "Fully Scalable Methods for Distributed Tensor Factorization," *Tkde,* vol. 29, *(1),* pp. 100-113, 2017. Available: https://ieeexplore.ieee.org/document/7569093. DOI: 10.1109/TKDE.2016.2610420.
[41] R. Bro, "PARAFAC. Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems,* vol. 38, *(2),* pp. 149, 1997. . DOI: 10.1016/s0169-7439(97)00032-4.
[42] Y. Zhang, P. Yang and V. Lanfranchi, "Tensor multi-task learning for predicting alzheimer's disease progression using MRI data with spatio-temporal similarity measurement," in Jul 21, 2021,
Available: https://ieeexplore.ieee.org/document/9557584. DOI: 10.1109/INDIN45523.2021.9557584.
[43] I. Jung *et al*, "MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis," *Frontiers in Genetics,* vol. 12, *(1),* pp. 682841, 2021. Available: https://search.proquest.com/docview/2576911840. DOI: 10.3389/fgene.2021.682841.
[44] P. M. Kroonenberg, *Three-Mode Principal Component Analysis.* 1983Available: http://www.econis.eu/PPNSET?PPN=014065479.
[45] Y. Taguchi, "Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases," *BMC Medical Genomics,* vol. 10, *(Suppl 4),* pp. 67, 2017.
Available: https://www.ncbi.nlm.nih.gov/pubmed/29322921. DOI: 10.1186/s12920-017-0302-1.
[46] Y. Taguchi, "Tensor decomposition-based unsupervised feature extraction applied to matrix products for multi-view data processing," *PLoS ONE,* vol. 12, *(8),* pp. e0183933, 2017. Available: https://www.ncbi.nlm.nih.gov/pubmed/28841719. DOI: 10.1371/journal.pone.0183933.

[47] M. RINGNER, "What is principal component analysis?" *Nature Biotechnology,* vol. 26, *(3),* pp. 303-304, 2008. Available: http://dx.doi.org/10.1038/nbt0308-303. DOI: 10.1038/nbt0308-303.
[48] B. Schölkopf, A. Smola and K. Müller, "Nonlinear Component Analysis as a Kernel Eigenvalue Problem," *Neural Computation,* vol. 10, *(5),* pp. 1299-1319, 1998. Available: https://direct.mit.edu/neco/article/doi/10.1162/089976698300017467. DOI: 10.1162/089976698300017467.
[49] M. N. Nounou *et al*, "Bayesian principal component analysis," *Journal of Chemometrics,* vol. 16, *(11),* pp. 576-595, 2002. Available: https://api.istex.fr/ark:/67375/WNG-9XMC3V01-W/fulltext.pdf. DOI: 10.1002/cem.759.
[50] Y. Xie *et al*, "Robust principal component analysis by projection pursuit," *Journal of Chemometrics,* vol. 7, *(6),* pp. 527-541, 1993. Available: https://api.istex.fr/ark:/67375/WNG-6CX42F08-2/fulltext.pdf. DOI: 10.1002/cem.1180070606.
[51] E. J. Beh, "Simple Correspondence Analysis: A Bibliographic Review," *International Statistical Review,* vol. 72, *(2),* pp. 257-284, 2004. Available: https://api.istex.fr/ark:/67375/WNG-6GWHW1JD-G/fulltext.pdf. DOI: 10.1111/j.1751-5823.2004.tb00236.x.
[52] N. Sompairac *et al*, "Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets," *Ijms,* vol. 20, *(18),* 2019. . DOI: 10.3390/ijms20184414.
[53] H. Zou, T. Hastie and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics,* vol. 15, *(2),* pp. 265-286, 2006. Available: https://www.tandfonline.com/doi/abs/10.1198/106186006X113430. DOI: 10.1198/106186006X113430.
[54] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Mach Learn,* vol. 83, *(3),* pp. 331-353, 2010. Available: https://link.springer.com/article/10.1007/s10994-010-5222-7. DOI: 10.1007/s10994-010-5222-7.
[55] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with ℓ0-constraints," *Neurocomputing,* vol. 80, *(1),* pp. 38-46, 2012. Available: https://dx.doi.org/10.1016/j.neucom.2011.09.024. DOI: 10.1016/j.neucom.2011.09.024.
[56] L. Liu and V. W. Berger, *Two by Two Contingency Tables.* 2014Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06160. DOI: 10.1002/9781118445112.stat06160.
[57] H. Park *et al*, "Global gene network exploration based on explainable artificial intelligence approach," *PLoS ONE,* vol. 15, *(11),* 2020. . DOI: 10.1371/journal.pone.0241508.
[58] X. Chen *et al*, "S.I. : HEALTHCARE ANALYTICS Global research on artificial intelligence-enhanced human electroencephalogram analysis," 2009. . DOI: 10.1007/s00521-020-05588-x(.
[59] R. de Kock *et al*, "Circulating biomarkers for monitoring therapy response and detection of disease progression in lung cancer patients," *Cancer Treatment and Research Communications,* vol. 28, pp. 100410, 2021. Available: https://dx.doi.org/10.1016/j.ctarc.2021.100410. DOI: 10.1016/j.ctarc.2021.100410.
[60] S. E. Counts *et al*, "Biomarkers for the Early Detection and Progression of Alzheimer's Disease," *Neurotherapeutics,* vol. 14, *(1),* pp. 35-53, 2016. Available: https://link.springer.com/article/10.1007/s13311-016-0481-z. DOI: 10.1007/s13311-016-0481-z.
[61] H. Fang *et al*, "Chapter 11 - omics biomarkers in risk assessment: A bioinformatics perspective," in *Computational Toxicology*, B. A. Fowler, Ed. 2013, Available: https://www.sciencedirect.com/science/article/pii/B9780123964618000130. DOI: https://doi.org/10.1016/B978-0-12-396461-8.00013-0.
[62] H. O. World and International Programme on, Chemical Safety, "Biomarkers in risk assessment: validity and validation," 2001. Available: https://apps.who.int/iris/handle/10665/42363.
[63] Y. Taguchi, "Drug candidate identification based on gene expression of treated cells using tensor decomposition-based unsupervised feature extraction for large-scale data," *BMC Bioinformatics,* vol. 19, *(Suppl 13),* pp. 388, 2019. Available: https://www.ncbi.nlm.nih.gov/pubmed/30717646. DOI: 10.1186/s12859-018-2395-8.
[64] M. W. Yeung *et al*, "Machine learning in cardiovascular genomics, proteomics, and drug discovery," *Machine Learning in Cardiovascular Medicine,* pp. 325, 2021. . DOI: 10.1016/b978-0-12-820273-9.00014-2.
[65] J. Lee, S. Oh and L. Sael, "GIFT: Guided and Interpretable Factorization for Tensors with an application to large-scale multi-platform cancer analysis," *Bioinformatics,* vol. 34, *(24),* pp. 4151-4158, 2018. Available: https://www.ncbi.nlm.nih.gov/pubmed/29931238. DOI: 10.1093/bioinformatics/bty490.
[66] C. Christin *et al*, "A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics," *Molecular & Cellular Proteomics,* vol. 12, *(1),* pp. 263-276, 2013. Available: https://dx.doi.org/10.1074/mcp.M112.022566. DOI: 10.1074/mcp.M112.022566.
[67] M. F. Buas *et al*, "Recommendation to use exact P-values in biomarker discovery research in place of approximate P-values," *Cancer Epidemiology,* vol. 56, pp. 83-89, 2018. Available: https://dx.doi.org/10.1016/j.canep.2018.07.014. DOI: 10.1016/j.canep.2018.07.014.

[68] B. Cao, X. Kong and P. S. Yu, "A review of heterogeneous data mining for brain disorder identification," *Brain Inf,* vol. 2, *(4),* pp. 253-264, 2015. Available: https://link.springer.com/article/10.1007/s40708-015-0021-3. DOI: 10.1007/s40708-015-0021-3.
[69] D. S. Warner *et al*, "Statistical Evaluation of a Biomarker," *Anesthesiology,* vol. 112, pp. 1023, 2010.
[70] K. Ng and Y. Taguchi, "Identification of miRNA signatures for kidney renal clear cell carcinoma using the tensor-decomposition method," *Sci Rep,* vol. 10, *(1),* 123456789. . DOI: 10.1038/s41598-020-71997-6.