Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2023;15(3):150-60

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA DOI: 10.5336/biostatic.2023-98597

Classification of RNA-Sequencing Data Via Poisson and Negative Binomial Linear Discriminant Analyses: A Methodological Study

RNA-Dizileme Verilerinin Poisson ve Negatif Binom Doğrusal Ayırma Analizleri ile Sınıflandırılması: Metodolojik Bir Çalışma

Dinçer GÖKSÜLÜK[a], Ahmet Ergün KARAAĞAOĞLU[b]

[a]Department of Biostatistics, Erciyes University Faculty of Medicine, Kayseri, Türkiye
[b]Department of Biostatistics, Lokman Hekim University Faculty of Medicine, Ankara, Türkiye

ABSTRACT Objective: Microarray and RNA sequencing (RNA-Seq) technologies are frequently employed in genetic data analysis for detecting disease-associated genes, identifying cancer subtypes, and enabling molecular diagnosis. While numerous methods have been proposed for classification problems using microarray data, there is a paucity of developed methods for classifying RNA-Seq data. This study aims to compare the performance of novel methods developed for RNA-Seq data on 3 distinct real-life datasets. **Material and Methods:** Cervical cancer, Alzheimer's disease, and kidney cancer RNA-Seq data were utilized in this study. The data were divided into training and test sets in a %70 and %30 ratio, respectively. Various preprocessing steps, such as normalization, power transformation, and variance filtering, were applied to the data. The Poisson Linear Discriminant Analysis (PLDA) and Negative Binomial Linear Discriminant Analysis (NBLDA) models were used for classification purposes, and the predictive performances of these models were compared. **Results:** Among the three datasets, the Alzheimer's data exhibited the lowest level of dispersion, while the cervical cancer data had the highest overdispersion. The NBLDA model demonstrated superior classification performance compared to the PLDA model. In cases of mild-to-moderate overdispersion, the predictive performance of the PLDA model improved when power transformation was applied, resulting in performance similar to that of the NBLDA model. **Conclusion:** PLDA and NBLDA models are two novel and promising techniques used in classifying RNA-Seq data. The performance of these models is influenced by the degree of overdispersion. In cases of high overdispersion, it is recommended to utilize the NBLDA model.

Keywords: Genomics; RNA-Sequencing; PLDA; NBLDA; classification

ÖZET Amaç: Mikrodizi ve RNA dizileme teknolojileri, genetik çalışmalarda hastalıkla ilişkili genlerin tespiti, kanser alt tiplerinin belirlenmesi, moleküler teşhis gibi amaçlar için sıklıkla kullanılan yöntemlerdir. Mikrodizi verilerinde sınıflama problemleri için literatürde birçok yöntem önerilmiştir. Bununla birlikte RNA dizileme verilerinde sınıflama problemleri için sınırlı sayıda yöntem bulunmaktadır. Bu çalışma, RNA dizileme verileri için geliştirilen yeni yöntemlerin performansını 3 farklı gerçek veri seti üzerinde karşılaştırmayı amaçlamaktadır. **Gereç ve Yöntemler:** Bu çalışmada, serviks kanseri, Alzheimer hastalığı ve böbrek kanseri RNA dizileme verileri kullanılmıştır. Veriler, sırasıyla %70 ve %30 oranında eğitim ve test kümelerine ayrılmıştır. Normalizasyon, güç dönüşümü ve varyans filtreleme gibi çeşitli ön işlemlerden sonra veriler, Poisson Doğrusal Ayırma Analizi (PDAA) ve Negatif Binom Doğrusal Ayırma Analizi (NBDAA) modelleri kullanılarak modellenmiş ve modellerin tahmin performansları karşılaştırılmıştır. **Bulgular:** Üç veri seti arasında Alzheimer verisi en düşük, serviks kanseri verisi ise en yüksek aşırı dağılıma sahipti. NBDAA modeli, PDAA modeline göre daha iyi sınıflandırma performansı göstermiştir. Hafif-orta derecede aşırı dağılım gözlendiği durumlarda, PDAA modelinin tahmin performansı güç dönüşümü uygulandığında iyileşmiş ve NBDAA ile benzer performans elde edilmiştir. **Sonuç:** PDAA ve NBDAA modelleri, RNA dizileme verilerinin sınıflandırılmasında kullanılan yeni ve umut verici tekniklerdir. Bu modellerin performansı, veri setindeki aşırı yaygınlığın derecesinden etkilenmektedir. Veride yüksek aşırı yaygınlık olması durumunda NBDAA modelinin kullanılması önerilmektedir.

Anahtar kelimeler: Genomik; RNA-dizileme; Poisson Doğrusal Ayırma Analizi; Negatif Binom Doğrusal Ayırma Analizi; sınıflama

Correspondence: Dinçer GÖKSÜLÜK
Department of Biostatistics, Erciyes University Faculty of Medicine, Kayseri, Türkiye
E-mail: dincergoksuluk@erciyes.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 22 Jun 2023 ***Received in revised form:*** 05 Sep 2023 ***Accepted:*** 12 Sep 2023 ***Available online:*** 20 Sep 2023

Gene-expression-based studies hold significant importance in molecular biology, as they allow for the examination of transcriptional activities across different tissue samples or cell populations.[1] Over the past 2 decades, extensive literature has focused on evaluating the impact of transcriptional expression patterns on specific conditions, such as biological conditions and tumor subtypes.[2] Two high-throughput technologies, microarray and next-generation sequencing (NGS), play a crucial role in quantifying gene expression. Among these, RNA sequencing (RNA-Seq) utilizes the capabilities of NGS technology to characterize and quantify gene expression.[3] Recent advancements have made it feasible to simultaneously examine the expression levels of thousands of genes, leading researchers to concentrate on multiple analysis tasks such as class discovery, class comparison, and class prediction. Although both microarray and RNA-Seq techniques provide expression levels of thousands of genes simultaneously, RNA-Seq has emerged as the state-of-the-art approach in such analysis tasks due to its major advantages.[4]

In the early stages, gene expression data from microarray technology played a pivotal role in the molecular diagnosis of diseases. Thanks to the continuous nature of microarray data, it became feasible to employ classical machine learning algorithms with minor modifications to the algorithm or preprocessing of the gene expression data, such as normalization and/or transformation.[5] However, when it comes to the utilization of RNA-Seq data in classification problems, it is necessary to take into consideration additional analysis steps because the algorithms proposed for microarray data are not directly applicable to RNA-Seq data due to the underlying discrete distribution.

In classification studies of RNA-Seq data, 2 strategies are available: proposing a novel algorithm based on discrete distributions, such as negative binomial (NB) and Poisson, and transforming the data to make it distributionally closer to microarrays, then applying microarray-based algorithms.[4-8] Preferring the latter strategy to enable numerous classical machine learning algorithms may initially appear reasonable. However, the transformation of discrete data into a continuous space can result in a substantial loss of information, thereby potentially leading to biased conclusions. Therefore, employing novel classifiers based on discrete distributions may be more suitable for conducting such studies.

Numerous algorithms have been developed or adapted for classification tasks in gene expression studies, with each performing well under specific conditions such as the underlying probability distribution (i.e., continuous or discrete), data structure, and dimensionality (i.e., low or high dimensional in terms of the number of samples and/or features). In recent years, considerable effort has been devoted to the examination of differential expression and classification analysis of RNA-Seq data.[9-11] Although differential expression and classification analyses are crucial in understanding gene expression data, advancements specific to the classification of RNA-Seq data have been relatively limited until recently. Two prominent and contemporary techniques employed in the classification of RNA-Seq data are Poisson Linear Discriminant Analysis (PLDA), proposed by Witten, and Negative Binomial Linear Discriminant Analysis (NBLDA), introduced by Dong et al.[4,7]

This study aimed to employ PLDA and NBLDA techniques for the classification of RNA-Seq data, utilizing three real-life genomic datasets. The primary objective was to compare the predictive accuracy of the aforementioned methods. Furthermore, this study sought to investigate potential differences between the 2 in the presence of overdispersion and high dimensionality, thereby illuminating their respective strengths and weaknesses in making predictions. By conducting this analysis, we aim to provide valuable insights into the performance and applicability of PLDA and NBLDA in the classification of RNA-Seq data, offering researchers a better understanding of their suitability for different scenarios.

MATERIAL AND METHODS

Let $\mathbf{X}$: $\{x_{ij};\ i: 1,2,\ldots,p\ \ j: 1,2,\ldots,n\}$ represent a matrix with dimensions p-by-n that encompasses the gene expression data obtained through RNA-sequencing technology. In this matrix, the columns correspond to

samples, while the rows represent features. The RNA-sequencing process generates mapped read counts x_{ij}, which are presumed to correlate with the gene expression levels. However, these counts depend not only on the gene expression level but also on factors such as sequencing depth, gene length, and the quality of sequences. Therefore, raw counts cannot be directly utilized as an accurate measure of gene expression level unless they undergo preprocessing for downstream analysis. Figure 1 shows the detailed workflow of the analysis steps.

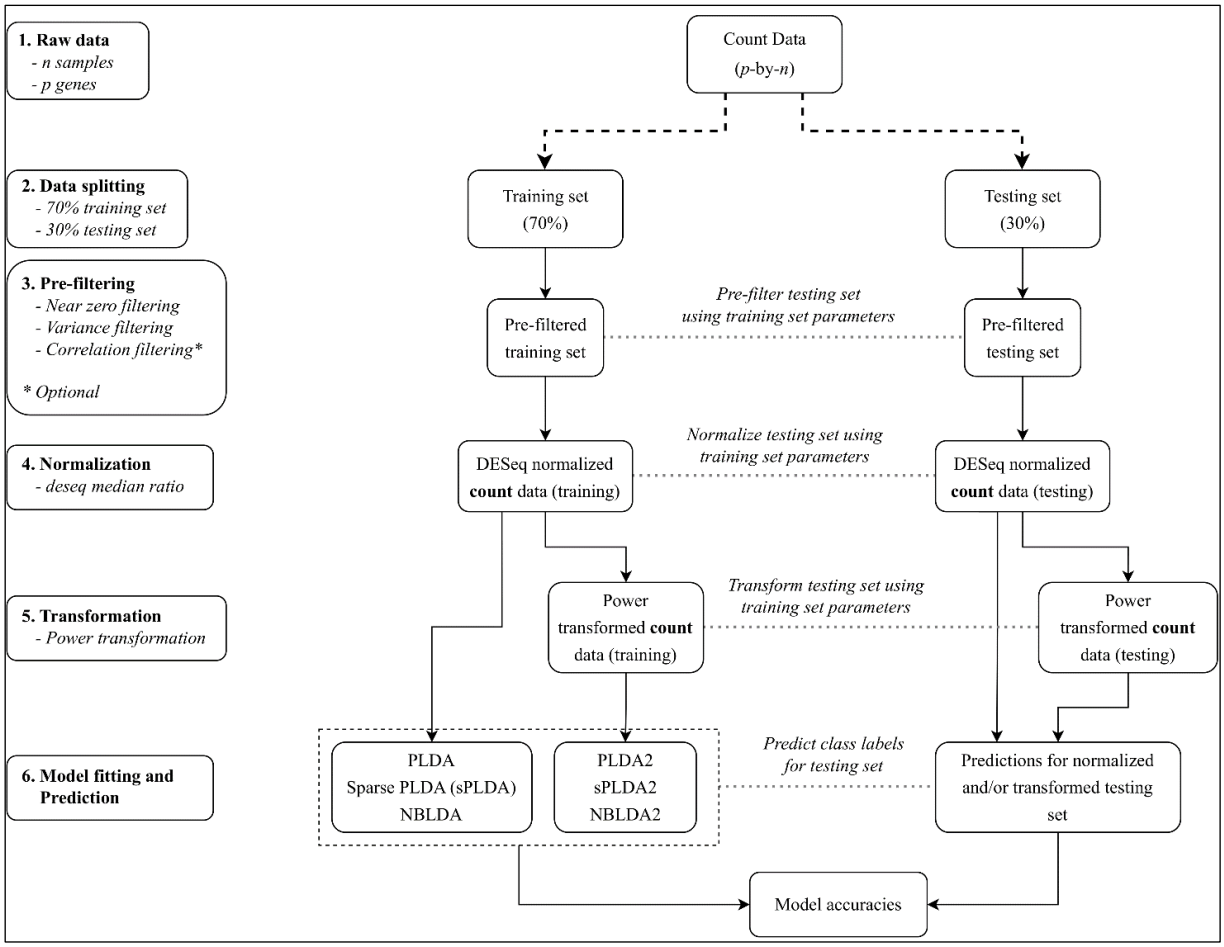

FIGURE 1: Workflow of RNA-sequencing data classification.

PLDA/sPLDA: Poisson Linear Discriminant Analysis/Sparse PLDA; NBLDA: Negative Binomial Linear Discriminant Analysis.

In the classification of RNA-sequencing data, the dataset was partitioned into training and testing sets using a split ratio of 70% and 30%, respectively. Subsequently, a series of pre-filtering procedures, such as near-zero filtering and variance filtering, was implemented on the raw counts to eliminate genes with low read counts and/or poor sequencing quality. Following this, a normalization step was applied to the pre-filtered raw counts before downstream analysis to remove the effect of sequencing depth, technical variation, and possible bias while preserving the biological variations between samples. Several normalization techniques have been proposed to address variations between samples.[12] In our study, we utilized the *DESeq median ratio normalization method* to normalize the raw counts while fitting data to PLDA and NBLDA as detailed in subsections 2.1 and 2.2.[11] This normalization method is known to be robust against outliers and ef-

fectively remove technical variation and bias in the raw data. The size factors are estimated by $s_j = m_j / \sum_{j=1}^{n} m_j$ where m_j is defined as:

$$m_j = \text{median}\left\{\frac{x_{ij}}{G_i}\right\}_{i:G_i \neq 0} \qquad G_i = \left(\prod_{j=1}^{n} x_{ij}\right)^{1/n} \qquad (1)$$

Here, G_i is the geometric mean of read counts for *i*-th feature and m_j is calculated by using the features having nonzero geometric mean. Additionally, we applied a power transformation, as proposed by Witten, to mitigate overdispersion when the data exhibit slight or moderate overdispersion.[4] Finally, the pre-filtered, normalized, and transformed data were fitted to the PLDA and NBLDA models with and without power transformation. The power-transformed models were denoted as PLDA2, sparse PLDA (sPLDA2), and NBLDA2 in the results section.

The predictive accuracy of these methods was calculated using independent testing sets. We should note that the predictive accuracy of fitted models may significantly change depending on the training and testing set samples. Therefore, the analysis steps outlined in Figure 1 were repeated 100 times for each RNA-Seq data to enhance the generalizability of our findings, measure the variation in classification accuracies across multiple experiments, and provide more reliable results while mitigating the risk of overfitting. We evaluated the fitted models using averaged accuracies and standard deviations over 100 repetitions. All analyses were executed utilizing the MLSeq package within the R/Bioconductor[1] network and the NBLDA package within the CRAN[2] network.[6]

PLDA

The Poisson distribution is commonly employed for modeling count data, such as in RNA-sequencing. Witten introduced a log-linear model, based on the Poisson distribution, for mapped read counts, expressed as follows:[4]

$$X_{ij} \mid y_i = k \sim \text{Poisson}(\mu_{ij} d_{ik}) \qquad \mu_{ij} = s_j g_i \qquad (2)$$

Here, s_j denotes the *size factor* of the *j*-th sample and g_i represents the *gene length* of the *i*-th feature. Moreover, d_{ik} serves as a class-specific *offset parameter* that enables the interpretation of the extent to which the observed counts of the *i*-th gene differ from the expected (or baseline) counts for the *k*-th class.

Let $\mathbf{x}^* = \{x_1^*, x_2^*, \ldots, x_p^*\}$ denote the observed mapped reads, and let y^* represent the unknown true class of a test sample. The class of the test sample can be predicted using the fitted model in equation 1. By applying Bayes' rule, we can calculate the posterior probabilities for each class as follows:

$$P(y^* = k \mid \mathbf{x}^*) \propto f_k(\mathbf{x}^*)\, \pi_k$$

where π_k represents the prior probability for the *k*-th class, and $f_k(.)$ is the probability density function. To obtain the discrimination function, we incorporate a Poisson distribution into $f_k(.)$, resulting in:

$$\begin{aligned} \log P(y^* = k \mid \mathbf{x}^*) &= \log \hat{f}_k(\mathbf{x}^*) + \log \hat{\pi}_k + c \\ &= \sum_{i=1}^{p} x_i^* \log \hat{d}_{ik} - \hat{s}^* \sum_{i=1}^{p} \hat{g}_i \hat{d}_{ik} + \log \hat{\pi}_k + c' \end{aligned} \qquad (3)$$

The estimated offset parameter $\hat{d}_{ik}$ possesses a straightforward and valuable interpretation: the *i*-th gene is down-regulated if $\hat{d}_{ik} < 1$ or up-regulated if $\hat{d}_{ik} > 1$ for class *k*. If the $\hat{d}_{ik}$ estimates equal 1 for all

[1] https://www.bioconductor.org/packages/release/bioc/html/MLSeq.html
[2] https://cran.rstudio.com/web/packages/NBLDA/index.html

classes, it can be assumed that the *i*-th gene is unrelated to the disease. Consequently, it can be removed from the model since the observed counts of the *i*-th gene do not contribute to the discrimination score in equation 3. This property endows PLDA with sparsity, achieved by selecting the features through $\hat{d}_{ik}$ that contribute to the discrimination score.

NBLDA

The NB distribution is an extension of the Poisson distribution that accounts for overdispersion in data, characterized by a larger variance than the mean. Likewise the Poisson model, let X_{ij} denote a random variable representing observed read counts. The fitted NB model, proposed by Dong et al. can be expressed as:[7]

$$X_{ij} \mid y_i = k \sim \mathrm{NB}(\mu_{ij} d_{ik}, \phi_i) \qquad \mu_{ij} = s_j g_i \qquad (4)$$

where ϕ_i represents the estimated overdispersion parameter obtained through the relationship between the mean and variance, given by $Var(X_{ij}) = \mu_{ij} + \mu_{ij}^2 \phi_i$. Finally, employing the NB probability density function and applying Bayes' rule, we derive the discrimination function as follows:

$$\begin{aligned} \log P(y^* = k \mid \mathbf{x}^*) &= \log \hat{f}_k(\mathbf{x}^*) + \log \hat{\pi}_k + c \\ &= \sum_{i=1}^{p} x_i^* \left[\log \hat{d}_{ik} - \log(1 + \hat{\phi}_i \hat{\omega}_{i*k})\right] \\ &\quad - \sum_{i=1}^{p} \hat{\phi}_i^{-1} \log(1 + \hat{\phi}_i \hat{\omega}_{i*k}) + \log \hat{\pi}_k + c' \end{aligned} \qquad (5)$$

where $\hat{\omega}_{i*k} = \hat{s}^* \hat{g}_i \hat{d}_{ik}$. It is evident from the discrimination score that the estimated dispersion parameter $\hat{\phi}_i$ exerts an influence on the model. In contrast to the Poisson model, a feature with a non-zero dispersion value (i.e., $\hat{\phi}_i \neq 0$) is included in the model, even if it does not exhibit differential expression across classes, that is, $\hat{d}_{ik} = 1$ for all $k = 1, 2, \ldots, K$. Consequently, the NBLDA algorithm does not function as a sparse classifier, encompassing all the features within the model. Furthermore, we observe that the NBLDA model converges to the PLDA model as $\phi_i \to 0$.

RNA-SEQ DATASETS

We conducted classification analyses on three distinct RNA-Seq datasets, which are publicly available through references: cervical cancer data, Alzheimer's disease data, and renal cell carcinoma (RCC) data obtained from The Cancer Genome Atlas (TCGA).[13-15] It is important to note that the present study is a methodological investigation, and the datasets employed are publicly accessible via databases or articles. As a result, obtaining approval from a Local Ethics Committee is not necessary for the utilization of these specific datasets in this study.

The cervical cancer dataset consisted of miRNA-Seq data from 58 human cervical tissue samples, including 29 tumor samples and 29 matched control samples. The Solexa/Illumina platform was utilized for the sequencing process, and a total of 714 miRNAs were investigated in this study.

The Alzheimer's disease dataset comprised sequencing reads of 2,801 miRNAs extracted from blood samples of 48 Alzheimer's patients and 22 age-matched control subjects.

Lastly, the RCC dataset downloaded from TCGA encompassed sequencing reads of 20,531 known human RNAs derived from 1,020 RCC patients. These patients were categorized into the three most prevalent subcategories: kidney renal papillary cell, kidney renal clear cell, and kidney chromophobe carcinomas, with sample sizes of 606, 323, and 91, respectively.

RESULTS

We have presented the pre-processing and testing set classification results for 3 real-life RNA-sequencing datasets in Table 1. The upper part of this table provides an overview of the pre-processing results applied to the datasets during the training process. Additionally, the lower part of Table 1 displays the classification results of the fitted models on the testing set. While training the models, the pre-filtering step has eliminated a substantial proportion of features in the datasets related to cervical cancer and Alzheimer's disease, resulting in the exclusion of 36.2% and 77.3% of all features on average, respectively. Due to the high dimensionality and computational complexity, we maintained the number of features in the RCC dataset at a constant value of 2000, which were selected through maximum variance filtering (Table 1). Furthermore, Figure 2 illustrates the estimates of gene-wise overdispersion, which were obtained from the normalized counts. The percentage of features in each dataset with overdispersion estimates exceeding 1, indicating high overdispersion, was calculated. The results reveal that 89.1%, 22.1%, and 41.8% of all features exhibit highly overdispersed read counts in the cervical cancer, Alzheimer's disease, and RCC datasets, respectively. Therefore, the cervical cancer data showed the highest overdispersion, while the Alzheimer's disease data showed the least.

TABLE 1: Classification results in testing set for real-life RNA-sequencing data.

	Cervical		**Alzheimer**		**Kidney**	
Number of features						
Raw data	714		2801		20531	
Pre-filtered (average)	455.2		634		2000	
Class sizes	29/29		22/48		91/323/602	
Class ratios	1:1		1:2.18		1:3.55:6.62	
Models*	**Accuracy**	**Sparsity**	**Accuracy**	**Sparsity**	**Accuracy**	**Sparsity**
PLDA	0.8759		0.4880		0.8630	
PLDA2	0.9253		0.7620		0.8778	
sPLDA	0.8729	1.000	0.4860	1.000	0.8619	1.000
sPLDA2	0.9094	0.299	0.7620	1.000	0.8778	1.000
NBLDA	0.9400		0.7930		0.8997	
NBLDA2	0.9493		0.7970		0.9045	

Lower part represents the model accuracies in the testing set.

*The suffix '2' stands for the power transformation; PLDA/sPLDA: Poisson Linear Discriminant Analysis/Sparse PLDA; NBLDA: Negative Binomial Linear Discriminant Analysis.

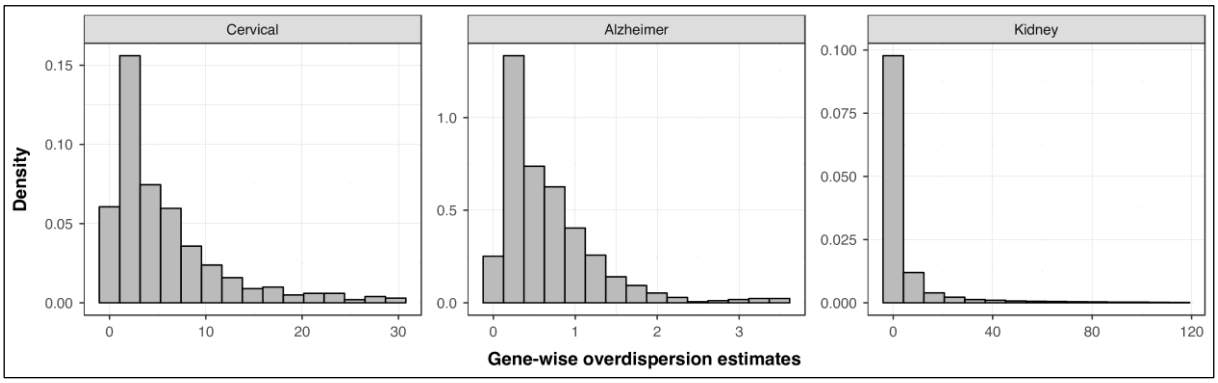

FIGURE 2: Gene-wise overdispersion estimates.

Classification results of PLDA and NBLDA models with and without transformation were graphically presented in Figures 3, Figure 4 and Figure 5. The prediction accuracies clearly demonstrate the significant impact of overdispersion on the model performances. Across all datasets, the NBLDA and NBLDA2 models consistently outperformed the others.

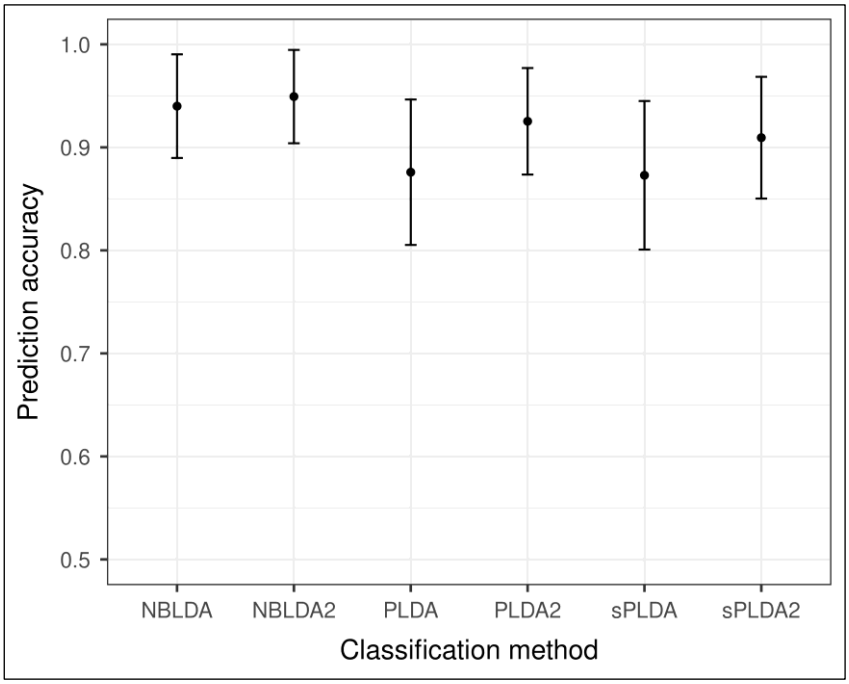

FIGURE 3: Prediction accuracy of fitted models for cervical cancer data.
NBLDA: Negative Binomial Linear Discriminant Analysis; PLDA/sPLDA: Poisson Linear Discriminant Analysis/Sparse PLDA.

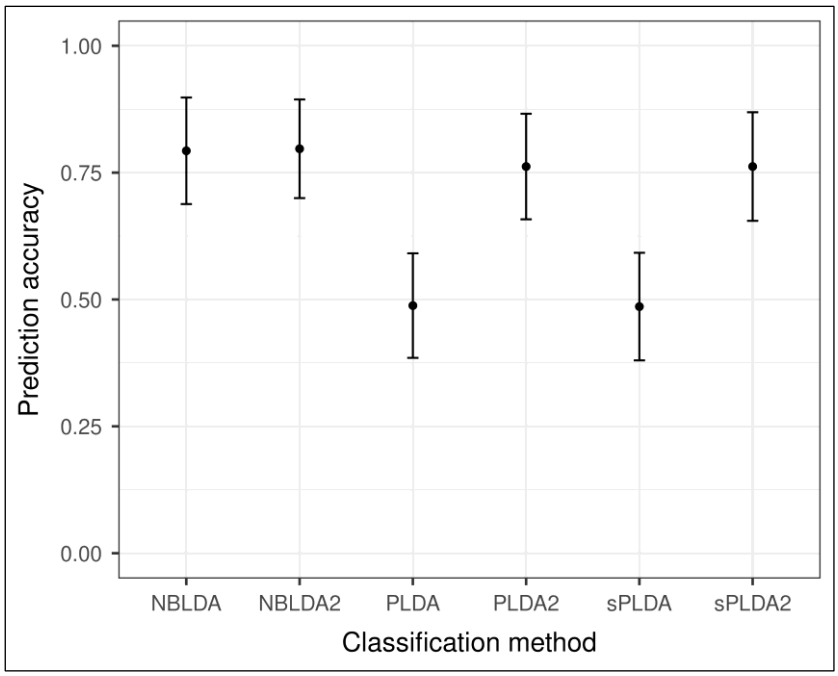

FIGURE 4: Prediction accuracy of fitted models for Alzheimer disease data.
NBLDA: Negative Binomial Linear Discriminant Analysis; PLDA/sPLDA: Poisson Linear Discriminant Analysis/Sparse PLDA

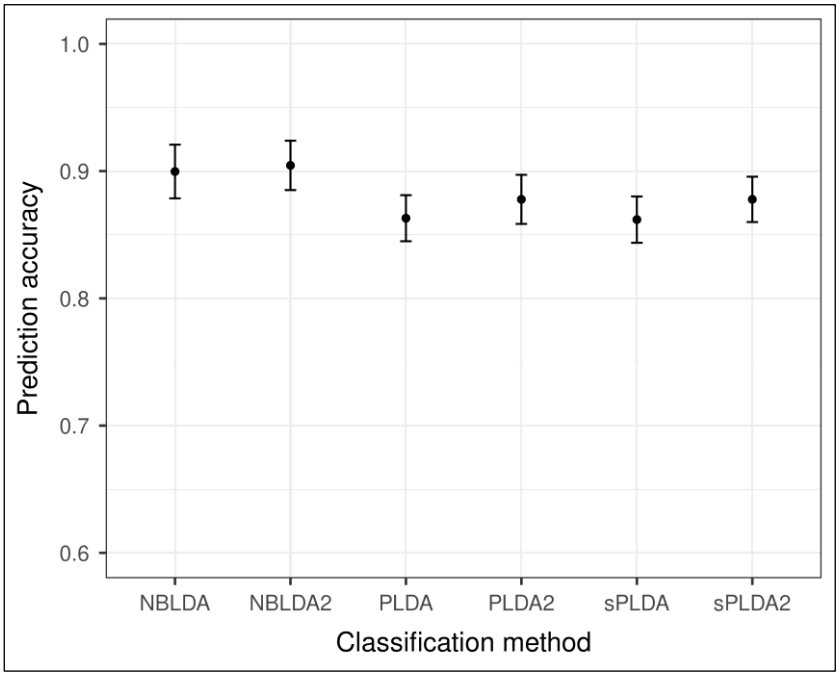

FIGURE 5: Prediction accuracy of fitted models for kidney cancer data.
NBLDA: Negative Binomial Linear Discriminant Analysis; PLDA/sPLDA: Poisson Linear Discriminant Analysis/Sparse PLDA.

In the cervical cancer dataset, characterized by the highest degree of overdispersion, the prediction accuracies were generally above 0.9 (Figure 3). Notably, the classification accuracies of both PLDA and NBLDA classifiers improved when power transformation was applied to the normalized counts. However, the influence of power transformation was more pronounced in the PLDA compared to NBLDA. Furthermore, the sPLDA model failed to select a subset of features using its built-in feature selection algorithm. However, after the application of power transformation, the sPLDA2 model achieved a prediction accuracy above 0.9, including only 30% of the available features.

The classification results for the data on Alzheimer's disease are presented in Figure 4. Despite being the least overdispersed dataset among the others, the PLDA classifier yielded the lowest classification accuracy unless a power transformation was applied. Upon performing the power transformation, both the PLDA and NBLDA classifiers demonstrated similar performance, achieving a prediction accuracy slightly above 0.75. Conversely, in the case of cervical cancer data, the sparse PLDA classifier failed to select a subset of features regardless of whether a power transformation was applied or not.

Finally, we have presented the classification accuracy for the RCC dataset in Figure 5. Similarly, as observed in the other datasets, NBLDA classifiers demonstrated better performance compared to PLDA classifiers in the RCC data. Due to the failure of sparse PLDA to select a feature subset, the prediction accuracies were nearly identical for both non-sparse and sparse classifiers. The application of a power transformation had a negligible effect on the prediction accuracies.

DISCUSSION

Microarrays and RNA sequencing, two widely utilized high-throughput technologies, play significant roles in genomic research for diverse objectives, including differential expression analysis, biomarker and gene discovery, class discovery, cell sub-type classification, and molecular diagnosis and disease

classification.[4,11,16-21] This study specifically focuses on disease classification using PLDA and NBLDA classifiers, which are specifically designed for RNA-Seq data. This study has contributed to the literature by providing a fair comparison of discrete classifiers from various RNA-sequencing datasets (i.e., miRNA and mRNA samples) under similar modeling conditions. Also, the datasets utilized in this study encompassed samples and features spanning from *small* to *large*. Consequently, our findings highlight the impact of varying the number of samples and features on classification accuracies using multiple datasets simultaneously. Furthermore, this study introduces a novel workflow for the classification of RNA-Sequencing data. In this workflow, testing samples undergo preprocessing using parameters derived from the training set rather than those of the testing set (Figure 1). This approach was initially proposed in our previous study, and has significantly influenced the way scientists approach RNA-Seq classification, leading them to implement similar workflow in future studies.[6]

Our findings demonstrate that NBLDA consistently outperforms PLDA across all utilized datasets. The results unequivocally establish the superiority of NBLDA over PLDA, particularly in the presence of high overdispersion within the data. The findings of this study are consistent with several previously published studies.[7,22,23] However, it remains unclear whether these related papers employed the same workflow utilized in this paper. Therefore, comparing the classification accuracies in numbers between these papers and our current study is not feasible, even though their conclusions may align similarly.

Overdispersion is a significant factor that profoundly impacts the predictive accuracy of fitted models. Within the three datasets examined, the Alzheimer's disease dataset exhibited a mild-to-moderate level of overdispersion, while the remaining two datasets, namely RCC and cervical cancer, demonstrated a considerably high degree of overdispersion. As expected, in Alzheimer disease dataset, the application of power transformation yielded similar prediction accuracies for PLDA and NBLDA. However, surprisingly, the performance of the PLDA classifier was notably lower when power transformation was not applied. The evident beneficial impact of power transformation on the prediction accuracies of the Alzheimer's disease dataset highlights its efficacy.

This study employed the DESeq median ratio normalization on the raw counts.[11] Despite RNA-Seq technology generating less noisy data compared to microarrays, normalization remains crucial in the classification of RNA-Seq data. Previous studies have consistently demonstrated the beneficial effects of normalization methods, particularly in differential expression analysis.[12,24] However, its impact on the model performances may be limited in the case of disease classification studies. In this study, a comparison of prediction accuracies under different normalization techniques was not conducted; however, a comprehensive comparison of other normalization techniques in disease classification can be found in the related paper.[25]

PLDA is an efficient sparse classifier that demonstrates the ability to select a subset of features associated with the response variable. This advantageous characteristic allows PLDA to eliminate redundant features from the model, making it suitable for identifying differentially expressed genes among distinct disease subsets. Contrarily, NBLDA is not inherently a sparse classifier as it incorporates all features into the model. Nevertheless, one can employ various feature selection methods to identify differentially expressed features and subsequently integrate them into the NBLDA classifier.[1,10,11] The utilization of a specific set of differentially expressed genes has the potential to enhance the prediction accuracy of the fitted model. Moreover, it is plausible to extend NBLDA into a sparse classifier by incorporating an intrinsic feature selection criterion. However, we regard this topic as a subject for future research and leave it unexplored in the current context.

RNA-Seq data generally exhibit an abundance of zeros, which can be attributed to zero-inflation. The current body of literature suggests the utilization of zero-inflated mixture distributions for the classification analysis of RNA-Seq data. Two recent techniques in this regard are the zero-inflated Poisson logistic discriminant analysis, introduced by Zhou et al., and the zero-inflated negative binomial logistic discriminant

analysis, proposed by Zhu et al.[26,27] Although the zero-inflated classifiers were not employed in our study, it is worth considering the comparison of NBLDA and PLDA with zero-inflated models.

Recently, there have been significant advancements in the field of machine learning and artificial intelligence (AI), leading many researchers to employ these techniques in genomic research.[28-30] Machine learning and AI-based approaches have been increasingly utilized for the purposes mentioned above. Unlike methods such as PLDA and NBLDA, which rely on specific underlying probability distributions, these techniques are not restricted by such assumptions. Consequently, machine learning and AI-based methods offer greater flexibility in the classification of RNA-sequencing data. Moreover, conducting a comprehensive simulation study for a fair comparison could provide valuable insights into the prediction accuracy of these methods in the context of RNA-Seq studies.

This study demonstrated the viability of PLDA and NBLDA classifiers, along with their respective extensions, for the classification of RNA-sequencing data. In conclusion, these two novel and widely-used techniques exhibit great promise and demonstrate strong performance under specific conditions.

CONCLUSION

PLDA and NBLDA models represent 2 innovative and promising techniques employed for the classification of RNA-Seq data. These methods are based on discrete distributions such as Poisson and negative binomial, enabling them to preserve the count data structure while fitting the models. Consequently, PLDA and NBLDA classifiers exhibit reduced susceptibility to information loss that may occur during the transformation of data into a continuous space. The performance of these models is influenced by the extent of overdispersion present in the data. When faced with mild-to-moderate overdispersion, the application of a power transformation generally yields a positive impact on model performance. However, in the case of high overdispersion, which is a common issue encountered in RNA-sequencing data, the effect of power transformation becomes limited and only marginally alters model performance. In conclusion, it is advisable to employ the NBLDA model when high overdispersion is observed in the data. Conversely, when slight overdispersion is present, it is reasonable to utilize the Poisson model with sparse extension on power-transformed data. This approach offers the advantage of reduced complexity and the ability to employ a smaller subset of all features.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Dinçer Göksülük, Ahmet Ergün Karaağaoğlu;* ***Design:*** *Dinçer Göksülük, Ahmet Ergün Karaağaoğlu;* ***Control/Supervision:*** *Dinçer Göksülük, Ahmet Ergün Karaağaoğlu;* ***Data Collection and/or Processing:*** *Dinçer Göksülük;* ***Analysis and/or Interpretation:*** *Dinçer Göksülük, Ahmet Ergün Karaağaoğlu;* ***Literature Review:*** *Dinçer Göksülük;* ***Writing the Article:*** *Dinçer Göksülük.*

REFERENCES

1. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. [Crossref] [PubMed] [PMC]
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-7. [Crossref] [PubMed]
3. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509-17. [Crossref] [PubMed] [PMC]
4. Witten DM. Classification and clustering of sequencing data using a Poisson model. Annals of Applied Statistics. 2011;5(4):2493-518.[Crossref]
5. Dudoit D, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association. 2002;97(457):77-87. [Crossref]
6. Goksuluk D, Zararsiz G, Korkmaz S, Eldem V, Zararsiz GE, Ozcetin E, et al. MLSeq: Machine learning interface for RNA-sequencing data. Comput Methods Programs Biomed. 2019;175:223-31. [Crossref] [PubMed]
7. Dong K, Zhao H, Tong T, Wan X. NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. BMC Bioinformatics. 2016;17(1):369. [Crossref] [PubMed] [PMC]
8. Zararsiz G, Goksuluk D, Korkmaz S, Eldem V, Zararsiz GE, Duru IP, et al. A comprehensive simulation study on classification of RNA-Seq data. PLoS One. 2017;12(8):e0182507.[Crossref] [PubMed] [PMC]
9. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29. [Crossref] [PubMed] [PMC]
10. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40. [Crossref] [PubMed] [PMC]
11. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. [Crossref] [PubMed] [PMC]
12. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94. [Crossref] [PubMed] [PMC]
13. Witten D, Tibshirani R, Gu SG, Fire A, Lui WO. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC Biol. 2010;8:58. [Crossref] [PubMed] [PMC]
14. Leidinger P, Backes C, Deutscher S, Schmitt K, Mueller SC, Frese K, et al. A blood based 12-miRNA signature of Alzheimer disease patients. Genome Biol. 2013;14(7):R78. [Crossref] [PubMed] [PMC]
15. Saleem M, Padmanabhuni SS, Ngomo AN, Almeida JS, Decker S, Deus HF. Linked cancer genome atlas database. Proceedings of the 9th International Conference on Semantic Systems, I-SEMANTICS'13. New York, NY, USA. 2013. p.129-34. [Crossref]
16. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods. 2008;5(7):613-9. [Crossref] [PubMed]
17. Osabe T, Shimizu K, Kadota K. Differential expression analysis using a model-based gene clustering algorithm for RNA-seq data. BMC Bioinformatics. 2021;22(1):511. [Crossref] [PubMed] [PMC]
18. Si Y, Liu P, Li P, Brutnell TP. Model-based clustering for RNA-seq data. Bioinformatics. 2014;30(2):197-205. [Crossref] [PubMed]
19. Le H, Peng B, Uy J, Carrillo D, Zhang Y, Aevermann BD, et al. Machine learning for cell type classification from single nucleus RNA sequencing data. PLoS One. 2022;17(9):e0275070. [Crossref] [PubMed] [PMC]
20. Hopper MA, Wenzl K, Hartert KT, Krull JE, Dropik AR, Novak JP, et al. Molecular classification and identification of an aggressive signature in low-grade B-cell lymphomas. Hematol Oncol. 2023. [PubMed]
21. Shen R, Fu D, Dong L, Zhang M, Shi Q, Shi Z, et al. Simplified algorithm for genetic subtyping in diffuse large B-cell lymphoma. Signal Transduction and Targeted Therapy. 2023;8(1):145. [Crossref] [PubMed] [PMC]
22. Rahman T, Huang H-E, Li Y, Tai A-S, Hseih W-P, McClung CA, et al. A sparse negative binomial classifier with covariate adjustment for RNA-seq data. Ann Appl Stat. 2022;16(2):1071-89. [Crossref]
23. Das S, Rai SN. Statistical methods for analysis of single-cell RNA-sequencing data. MethodsX. 2021;8:101580. [Crossref] [PubMed] [PMC]
24. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):R25. [Crossref] [PubMed] [PMC]
25. Han H, Men K. How does normalization impact RNA-seq disease diagnosis? J Biomed Inform. 2018;85:80-92. [Crossref] [PubMed]
26. Zhou Y, Wan X, Zhang B, Tong T. Classifying next-generation sequencing data using a zero-inflated Poisson model. Bioinformatics. 2018;34(8):1329-35. [Crossref] [PubMed]
27. Zhu J, Yuan Z, Shu L, Liao W, Zhao M, Zhou Y. Selecting classification methods for small samples of next-generation sequencing data. Front Genet. 2021;12:642227. [Crossref] [PubMed] [PMC]
28. Song JK, Zhang Y, Fei XY, Chen YR, Luo Y, Jiang JS, et al. Classification and biomarker gene selection of pyroptosis-related gene expression in psoriasis using a random forest algorithm. Front Genet. 2022;13:850108. [Crossref] [PubMed] [PMC]
29. Zhou Y, Peng M, Yang B, Tong T, Zhang B, Tang N. scDLC: a deep learning framework to classify large sample single-cell RNA-seq data. BMC Genomics. 2022;23(1):504. [Crossref] [PubMed] [PMC]
30. Chen X, Balko JM, Ling F, Jin Y, Gonzalez A, Zhao Z, et al. Convolutional neural network for biomarker discovery for triple negative breast cancer with RNA sequencing data. Heliyon. 2023;9(4):e14819. [Crossref] [PubMed] [PMC]