**OMICS Data in the Diagnosis of Diabetic Retinopathy: A Comparison between Transcriptome Data and DNA Methylation Data**

|  |  |  |
| --- | --- | --- |
| Pratheeba Jeyananthan  Department of Computer Engineering  Faculty of Engineering,  University of Jaffna.  Jaffna, Sri Lanka.  [pratheeba@eng.jfn.ac.lk](mailto:pratheeba@eng.jfn.ac.lk) | Pratheeba Jeyananthan  Department of Computer Engineering  Faculty of Engineering,  University of Jaffna.  Jaffna, Sri Lanka.  [pratheeba@eng.jfn.ac.lk](mailto:pratheeba@eng.jfn.ac.lk) | Pratheeba Jeyananthan  Department of Computer Engineering  Faculty of Engineering,  University of Jaffna.  Jaffna, Sri Lanka.  [pratheeba@eng.jfn.ac.lk](mailto:pratheeba@eng.jfn.ac.lk) |

**Abstract:**

Diabetic retinopathy is a serious issue caused by diabetic that can damage the eye retina. This can lead to many complications up to the level of blindness if untreated. Even though this can be diagnosed using dilated eye exams, current technologies left the world with many different digital data such as medical images and omics data. Medical images are widely used in the diagnosis of this disease using machine learning algorithms. However analysing the omics data has much more advantageous compared to image processing such as less complex, low time and computer power, especially biomarker identification can be done during the study of omics data. In the same way, this study uses three different omics data such as DNA methylation, total RNA and small RNA in the diagnosis of diabetic retinopathy using different machine learning algorithms. Four different feature selection algorithms are used with each data individually to select the biomarkers of the study and the best set of features are used with different machine learning algorithms to reveal the model with the highest accuracy. Comparing the accuracies between models shows that best 14 total RNA features selected using Random Forest Feature Importance along with Naïve Bayes algorithms outperforms other model with the accuracy value of 0.9625. Further to this selected set of features are biologically validated using gene ontology (GO) analysis.

**Keywords:**

Diabetic Retinopathy; Diagnosis; Machine Learning Models; DNA Methylation; Transcriptome data

**Introduction:**

**Material and Methods:**

This study uses two different omics data in the diagnosis of diabetic retinopathy with the aid of machine learning algorithm. DNA methylation, small RNA and total RNA data are downloaded from a public data repository Gene Expression Omnibus (GEO) and used in this machine learning based study. Four different feature selection methods are individually used in the selection of relevant features of this diagnosis and the selected set of features from each method are used with linear-support vector machines (SVM), logistic regression, Naïve Bayes and Random Forest. Performance differences between each model (omics data+feature selection method+ classification method) are compared and the best one is proposed for the diagnosis of diabetic retinopathy.

**Material:**

All the data used in this study are obtained from a public data repository called Gene Expression Omnibus (GEO). TotalRNAdata and smallRNA data (accession number of GSE160310) are obtained using expression profiling by high throughput sequencing and non-coding RNA profiling by throughput sequencing respectively.In each data, there are 79 samples where 39 patients and 40 control samples [1]. DNA methylation data (accession number of GSE140842) is also downloaded from the same repository where methylation profiling by high throughput sequencing is used to measure this methylation data. In their experiment, they used 35 patients with 35 age, gender, diabetic duration matched control samples[2].

**Methods:**

This study compares the performance difference between four different feature selection algorithms such as mutual information, correlation coefficient, Chi-square method and feature importance. Selected set of features are used with four different machine learning algorithms, linear-SVM, logistic regression, Naïve Bayes, Random Forest and their performances also compared. Accuracy is the measure used in measuring the accuracy of the models.

**Results:**

This study starts with data preprocessing. In DNA methylation data, there are no null values and it is already normalized data. In both datasets, any existing categorical values are changed into numerical values before the next step. Same as epigenomic data, there are no null values in transcriptome data as well. However, this data is not normalized, hence the data is normalised between 0 and 1.

After preprocessing, next step is feature selection. Out of all four feature selection methods used in this study, two of them are correlation measures. First, in DNA methylation data, feature importance value of each feature is used to determine the number of relevant features needs to be considered in further analysis. Figure 1 shows that after top 189 features, the feature importance reached zero value.

|  |
| --- |
|  |
| Figure 1: Feature importance values of first 200 methylation features. It shows that after 189 features the feature importance value gets to zero, hence top 189 features are of our interest. |

Hence, the total number of methylation features need to be considered in this study is now narrowed down to top 189 features. However, as we are bounded with the n<<p problem (number of sample<<number of features), we need to reduce the maximum number of features up to 70 features. By satisfying all the above conditions in number of features, the accuracies between feature selection methods are compared to select the best feature selection algorithms. For that, different number of features within the limit of 70 features such as top 20, 30, 40, 50, 60 and 70 features are selected using each of these feature selection algorithm, used with cross-validation and their accuracies are compared. Table 1 shows that correlation coefficient features unanimously outperform other feature selection algorithms with any number of features considered here.

Table : Performance comparison between different numbers of features selected using four different feature selection algorithms and Random Forest classification algorithm. Top 20, 30, 40, 50, 60 and 70 methylation features are selected using those algorithms and used with Random Forest and their performances are compared.

**I need this table for 20,30,40,50,60 and 70 features, not this .we need to be constant. While writing it seems like logically there are some problems in the current form**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Top 20 features | 30 features | 40 features | 50 features | 60 features | 70 features |
| Mutual information | 0.47+/-0.07 | 0.46+/-0.09 | 0.49+/-0.7 | 0.49+/-0.07 | 0.47+/-0.07 | 0.49+/-0.07 |
| Correlation Coefficient | 0.81+/-0.12 | 0.79+/-0.13 | 0.86+/-0.12 | 0.86+/-0.10 | 0.87+/-0.08 | 0.87+/-0.05 |
| Chi-square | 0.77+/-0.12 | 0.77+/-0.13 | 0.81+/-0.13 | 0.79+/-0.10 | 0.76+/-0.12 | 0.80+/-0.10 |
| Feature importance | 0.71+/-0.08 | 0.64+/-0.17 | 0.73+/-0.16 | 0.81+/-0.12 | 0.79+/-0.16 | 0.79+/-0.16 |

As per the observation from Table 1, now different numbers of features between 45 to 55 selected using correlation coefficient feature selection algorithm are subjected to different feature selection algorithms and their accuracies are compared. Figure 2 shows that top 48 correlation coefficient selected methylation features along with Logistic Regression algorithm produces the best accuracy value of 0.9571+/-0.0350.

Set of 48 features with the highest accuracy (supplementary material) are subjected to gene ontology (GO) to biologically validate the data. Top biological functions or pathways identified in relation to these features are XXXXXX. Whole set of GO terms related to these features are presented in the supplementary material.

**Now we need to draw a graph illustrating the accuracies by 45,46,....55 features using SVM and logistic regression.**

|  |
| --- |
|  |
| Figure 2: Performance comparison between different numbers of features selected using correlation coefficient along with different machine learning algorithms. |

Now smallRNa data extracted from 79 samples and their performances are compared between machine learning algorithms to reveal the best set of features produces that accuracy along with the name of the corresponding machine learning algorithm. In the same way as described above, important set of features in this diagnosis is tested using feature importance value of each feature.

|  |
| --- |
|  |
| Figure 3: Feature importance values between features are compared to see the number of smallRNA features influences on the diagnosis of diabetic retinopathy. This figure shows that after 643 features the feature importance value reached to zero. |

Now again by considering the same n<<p problem, we select top 20, 30, 40, 50, 60 and 70 features from small RNA data using the same four machine learning algorithms and their performances are compared to see the top feature selection algorithm. Table 2 shows that Feature importance algorithm constantly performs better than other three algorithms. Hence we take this algorithm forwards to the next step.

Table 2: Different numbers of features are selected using four machine learning algorithms, each set is used with Random Forest algorithm in the diagnosis and their accuracies are compared to select the best feature selection algorithm in this prediction.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Top 20 features | 30 features | 40 features | 50 features | 60 features | 70 features |
| Mutual information | 0.40+/-0.09 | 0.35+/-0.05 | 0.44+/-0.13 | 0.42+/-0.11 | 0.38+/-0.09 | 0.39+/-0.05 |
| Correlation Coefficient | 0.38+/-0.08 | 0.38+/-0.08 | 0.38+/-0.08 | 0.37+/-0.06 | 0.37+/-0.06 | 0.38+/-0.08 |
| Chi-square | 0.38+/-0.08 | 0.37+/-0.06 | 0.38+/-0.08 | 0.38+/-0.08 | 0.37+/-0.06 | 0.37+/-0.06 |
| Feature importance | 0.69+/-0.12 | 0.67+/-0.20 | 0.72+/-0.12 | 0.67+/-0.14 | 0.67+/-0.15 | 0.69+/-0.11 |

Now different numbers of features between 45 and 55, selected using Feature importance are used with different machine algorithms and their accuracies are compared. Figure 4 shows that top 46 small RNA features selected using ANN, produces the best accuracy value of 1.9375+/-0.03953 using Cross-Validation.

Doing the same GO analysis on this set of features shows that they are related to XXXX. Name of the features and the whole set of GO terms are presented in the supplementary material.

|  |
| --- |
|  |
| Figure 4: Accuracies are compared between different number of features selected using Feature Importance, with different machine learning algorithms. |

As the final stage, totalRNA data is used in the same prediction. Figure 5 shows that after 668 features the feature important values reached zero, hence they can be neglected from this prediction. Again with the same n<<p problem, we consider top 20, 30, 40, 50, 60 and 70 features selected using four different machine learning algorithm with Random Forest algorithm. Comparing their accuracies shows that Feature Importance algorithm performs better in all cases compared to other three algorithms (Table 3). Hence, that algorithm is considered in the next step.

|  |
| --- |
|  |
| Figure 5: Feature importance value comparison between totalRNA features is compared in the diagnosis of diabetic retinopathy. After 668 features, the feature importance value reached zero. |

Table 3: Four different machine learning algorithms are used to select top 20, 30, 40, 50, 60 and 70 totlRNA features in diabetic retinopathy diagnosis and used with Random Forest. Comparing their accuracies shows that Feature Importance outperforms other three methods in this diagnosis.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Top 20 features | 30 features | 40 features | 50 features | 60 features | 70 features |
| Mutual information | 0.37+/-0.08 | 0.37+/-0.08 | 0.37+/-0.08 | 0.37+/-0.08 | 0.37+/-0.08 | 0.37+/-0.08 |
| Correlation coeff | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 |
| Chi-square | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 | 0.51+/-0.14 |
| Feature importance | 0.72+/-0.17 | 0.70+/-0.14 | 0.72+/-0.16 | 0.72+/-0.18 | 0.70+/-0.17 | 0.72+/-0.18 |

Now, using the observations in Table 3, numbers of features are narrowed down between 10 and 20. Each set of such features is used with two different classification algorithms SVM and Naïve Bayes. Figure 6 shows that while using totalRNA features in this diagnosis, top 14 features selected using Feature Importance along with Naïve Bayes outperform other combinations with the accuracy value of 0.9625+/-0.050.

Now comparing the accuracies between omics data shows that total RNA data outperforms the other omics data with the highest accuracy value of 0.9625+/-0.050. This accuracy is obtained using top 14 features selected using Feature Importance along with Naïve Bayes algorithm. GO analysis on these features shows that they are closely related to XXXXX. Set of features along with the full set of GO terms are presented in supplementary material.

**Discussion:**

**Conclusion:**

This study reveals the role of omics data in the diagnosis of diabetic retinopathy. Three different omics data such as DNA methylation, total RNA and small RNA data are used in this diagnosis using different machine learning algorithms and their performances are compared. As the data we work with has too many numbers of features, this study starts with feature selection, selecting the relevant features of the study. Four different feature selection algorithms are used in this feature selection and their accuracies are compared to select the best feature selection algorithm. Different numbers of features are selected using that particular feature selection algorithm and used with different machine learning algorithms. Comparing those models shows that top 14 total RNA features selected using Feature Importance along with Naïve Bayes algorithm outperforms other models with the accuracy value of 0.9625+/-0.05 in the diagnosis of diabetic retinopathy.

**Data accessibility:**

The datasets used in this study are publicly available in the Gene expression omnibus (GEO) data repository. This is a public data repository where data can be freely downloaded for studies.Transcriptome data with the accession number of GSE160310, and DNA methylation data with the accession number of GSE140842are used here. All the models use pre-built functions in Python libraries, none of them are implemented here from scratch.

**Funding:**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**Conflict of Interest:**

There are no conflicts of interest between authors.