JESPAR

Journal of Exercise Sciences
& Physical Activity Reviews

ORIGINAL ARTICLE

Development of a Bagged CART Model for Subclassification of Diabetic Retinopathy Using Metabolomics Data

Fatma Hilal Yagin[1 *] | Badicu Georgian[2]

[1]Department of Biostatistics and Medical Informatics, Faculty of Medicine, Inonu University, Malatya 44280, Türkiye
[2]Department of Physical Education and Special Motricity, Transilvania University of Brasov, 00152 Brasov, Romania

ABSTRACT

In this study, we present the development and evaluation of a predictive model for classifying the subclasses of diabetic retinopathy—No Diabetic Retinopathy (NDR), Non-Proliferative Diabetic Retinopathy (NPDR), and Proliferative Diabetic Retinopathy (PDR)—using metabolomics data. The metabolomics dataset underwent rigorous preprocessing to address missing values, employing the Random Forest algorithm, and was subsequently normalized to ensure comparability across all samples. A bagged Classification and Regression Trees (CART) algorithm was utilized to construct the prediction model, leveraging its robustness and accuracy for classification tasks. Our model demonstrated significant potential in accurately classifying diabetic retinopathy subclasses, suggesting that metabolomics data, when combined with advanced machine learning techniques, can provide valuable insights into the progression and management of diabetic retinopathy. This study underscores the importance of integrating metabolomics biomarkers and machine learning for the advancement of personalized medicine in diabetic care.

Keywords: Type 2 diabetes, diabetes prediction, machine learning, classification, metabolomics

***Corresponding:** Fatma Hilal Yagin; hilal.yagin@inonu.edu.tr
Journal home page: www.e-jespar.com
Academic Editor: Dr. Mehmet Gülü
https://doi.org/10.5281/zenodo.11544643

ARTICLE HISTORY
Received: 20 May 2024
Accepted: 28 May 2024
Published: 01 July 2024

INTRODUCTION

Diabetic retinopathy (DR) is a severe and common complication of diabetes, characterized by damage to the retina's blood vessels, which can lead to vision impairment and blindness. As the global prevalence of diabetes continues to rise, so does the incidence of DR, making it a significant public health concern. DR progresses through various stages, including no diabetic retinopathy (NDR), non-proliferative diabetic retinopathy (NPDR), and proliferative diabetic retinopathy (PDR), each representing increasing severity and risk of vision loss. Early and accurate identification of these stages is crucial for timely intervention and prevention of disease progression (Li et al., 2023; Mohamed, Gillies, & Wong, 2007; Wang et al., 2024).

Metabolomics, the comprehensive study of metabolites within biological systems, has emerged as a powerful tool in understanding the biochemical alterations associated with diseases, including DR. Metabolite profiles can provide insights into the underlying pathophysiological processes and have the potential to serve as biomarkers for disease diagnosis, prognosis, and therapeutic response. Recent advancements in high-throughput metabolomics technologies have enabled the acquisition of large-scale metabolite data, presenting new opportunities for disease characterization and prediction (Filla & Edwards, 2016; Hou, Wang, & Pan, 2021; Sun et al., 2022).

Machine learning (ML), a subset of artificial intelligence (AI), offers robust analytical methods capable of handling complex and high-dimensional data, such as metabolomics datasets. ML algorithms can identify patterns and relationships within the data that may not be apparent through traditional statistical approaches. By leveraging these capabilities, ML models can be developed to predict disease states and classify patients based on their metabolomics profiles, potentially enhancing diagnostic accuracy and enabling personalized treatment strategies (Ali & Mohammed, 2024; Galal, Talal, & Moustafa, 2022; Patra, Disha, Kundu, Das, & Ghosh, 2023).

In this study, we aim to develop and validate an ML prediction model using metabolomics data to classify patients into the subclasses of diabetic retinopathy: NDR, NPDR, and PDR. By integrating metabolomics biomarkers with advanced ML techniques, we seek to improve the early detection and classification of DR stages, thereby contributing to better clinical management and outcomes for diabetic patients.

METHODS

Dataset and related factors

The current work investigated the subtype prediction and biomarkers of DR in T2D patients using a publicly available dataset that examined clinical, biochemical, and metabolomic characteristics (Yun et al., 2020). 317 T2D patients in total—123 NPDR patients, 51 PDR patients, and 143 NDR patients—had open access data used in the study. A retina specialist's dilated fundus examination resulted in the diagnosis of DR. In compliance with global ethical requirements, serum samples were taken from T2D patients with and without DR and kept in a refrigerator at −80 °C. T2D patients' serum samples were assessed using a targeted metabolomics method. Following quality control procedures, 122 metabolites that defined the DR subclass were found, and as a result, these metabolites were chosen for further analysis.i

Methods

In this study, we employed a bagged Classification and Regression Trees (CART) algorithm to develop a prediction model for classifying the subclasses of diabetic retinopathy (NDR, NPDR, and PDR) using metabolomics data. The metabolomics data were preprocessed to remove any missing values based on the Random Forest algorithm and normalized to ensure comparability across samples.

Bagged CART Algorithm

Bagging is an ensemble method designed to improve the stability and accuracy of machine learning algorithms. In the context of this study, the bagged CART algorithm involves creating multiple CART models using different subsets of the training data. These subsets are generated by randomly sampling with replacements from the original dataset, a process known as bootstrapping. Each CART model is trained on a bootstrap sample and then combined to form an ensemble model. The final prediction is made by averaging the predictions of all individual CART models, thereby reducing variance and enhancing the model's robustness (Aydogmus et al., 2015; Zhang et al., 2022).

10-Fold Cross-Validation

To evaluate the performance of our bagged CART model, we utilized a 10-fold cross-validation approach. This technique involves partitioning the entire dataset into ten equal-sized folds. In each iteration of the cross-validation process, nine folds are used for training the model, and the remaining fold is used for testing. This process is repeated ten times, with each fold serving as the test set exactly once. The results from each iteration are then averaged to provide an overall assessment of the model's performance. This method helps

to ensure that the model is not overfitting to a particular subset of the data and provides a more generalized performance evaluation (Fushiki, 2011; Wong & Yeh, 2019).

Performance Metrics

The performance of the bagged CART model was assessed using several metrics, including accuracy, sensitivity, specificity, and F1 score. Accuracy measures the overall correctness of the model, calculated as the proportion of true positive and true negative predictions among the total predictions. Sensitivity (or recall) evaluates the model's ability to correctly identify positive cases (e.g., correctly classifying NDR, NPDR, or PDR). Specificity assesses the model's ability to correctly identify negative cases (e.g., distinguishing between different stages of DR). The F1 score provides a harmonic mean of precision and recall, offering a balanced measure of the model's performance (Yagin et al., 2023).

RESULTS

The performance evaluation of the bagged CART model for predicting the subclasses of diabetic retinopathy (NDR, NPDR, and PDR) yielded promising results. The model achieved an accuracy of 0.72, indicating that 72% of the predictions made by the model were correct. The sensitivity, or the model's ability to correctly identify positive cases, was notably high at 0.919, demonstrating that the model was effective in detecting instances of diabetic retinopathy. The specificity, which measures the ability to correctly identify negative cases, was 0.688, suggesting moderate performance in distinguishing between different stages of the condition. Additionally, the F1 score, which provides a balance between precision and recall, was 0.723, reflecting the overall effectiveness of the model in classifying the subclasses of diabetic retinopathy. These results demonstrate the potential utility of the bagged CART model in clinical settings for the early detection and classification of diabetic retinopathy stages based on metabolomics data.

Table 1. Performance metrics results of machine learning approach in DR prediction.

Metrics	Value
Accuracy	0.72
Sensitivity	0.919
Specificity	0.688
F1-Score	0.723

The analysis of metabolite importance in predicting the subclasses of diabetic retinopathy using the bagged CART model revealed several key biomarkers. Tryptophan (trp) was identified as the most important metabolite with an importance score of 100. It was followed by c4 with a score of 85.697, c3 with 78.865, and c16 with 78.296, indicating their

significant contributions to the model's predictive capabilities. Other important metabolites and factors included total.dma (77.425), age (77.032), creatinine (cr) (70.328), and tyrosine (tyr) (68.638). Glucose, a well-known marker for diabetes, had an importance score of 67.054, while HbA1c, another critical indicator of long-term glucose levels, had a score of 62.521. These findings highlight the diverse range of metabolites and clinical factors that contribute to the classification of diabetic retinopathy stages, emphasizing the complex interplay of biochemical and physiological factors in the progression of this condition.

Table 2. Importance rates of biomarker candidate metabolites contributing to DR prediction
Metabolite Importance

Metabolite	Importance
trp	100
c4	85.697
c3	78.865
c16	78.296
total.dma	77.425
age	77.032
cr	70.328
tyr	68.638
glucose	67.054
hba1c	62.521

DISCUSSION

The findings of this study demonstrate the effectiveness of using a bagged CART model with metabolomics data to classify DR subclasses. The model achieved an accuracy of 0.72, indicating a robust capacity to correctly classify patients into NDR, NPDR, and PDR stages. The high sensitivity of 0.919 reflects the robust performance of the model in detecting diabetic retinopathy cases, which is crucial for early intervention and management. The moderate specificity of 0.688 suggests that although the model is robust in identifying positive cases, there is room for improvement in more accurately distinguishing between different stages of DR.

The importance scores of metabolites provide important insights into biochemical markers associated with DR progression. Tryptophan (trp) emerged as the most critical metabolite with an importance score of 100. Previous studies have highlighted the role of tryptophan metabolism in diabetes and its complications, including retinopathy (Kozieł & Urbanska,

2023). The increasing importance of metabolites such as c4, c3, and c16 underscores their potential role in the pathogenesis of DR and warrants further investigation.

Age was also identified as an important predictor, consistent with the established knowledge that the risk of DR increases with age (Yau et al., 2012). The importance of HbA1c, a marker of long-term glucose control, further confirms its usefulness in DR classification, consistent with previous studies correlating HbA1c levels with DR severity (Control, Interventions, & Group, 2005).

These results highlight the potential of integrating metabolomics data with ML techniques to improve early detection and classification of DR. The use of a bagged CART algorithm provided robust estimates by reducing variance and improving stability through ensemble learning. The application of 10-fold cross-validation provided a comprehensive assessment, increasing the generalizability of the model.

Despite the promising results, there are limitations to this study. The moderate specificity value suggests that the model could benefit from further integration and validation with larger, more diverse datasets. In addition, although the model identified key metabolites, the underlying biological mechanisms linking these metabolites to DR progression need to be further elucidated through experimental studies.

Future research should focus on expanding the dataset to include a more diverse patient population and exploring additional metabolomics and clinical biomarkers. Integrating other omics data, such as genomics and proteomics, may further increase predictive accuracy and provide a more holistic understanding of DR pathophysiology. Furthermore, developing more sophisticated machine learning models, including deep learning techniques, could potentially improve classification performance and reveal more complex patterns in the data.

In conclusion, this study demonstrates the feasibility and potential of using a bagged CART model with metabolomics data to classify DR stages. The findings highlight the importance of metabolomics in improving disease diagnostics and demonstrate the need for continued research to improve predictive models and understand the complex biochemical pathways involved in DR.

Author Contributions

Conceptualization, F.H.Y. methodology, F.H.Y, B.G.; formal analysis, F.H.Y, B.G.; investigation, B.G.; data curation, F.Y.H.; writing—original draft preparation, F.H.Y, B.G.; writing—review and editing, F.H.Y, B.G.

Informed Consent Statement:

The research was conducted in line with the Declaration of Helsinki.

Acknowledgments:

We would like to thank all participants who took part in the research.

Funding:

This research was not funded by any institution or organization.

Conflicts of Interest:

The authors declare that no conflicts interest.

REFERENCES

Ali, A. M., & Mohammed, M. A. (2024). A Comprehensive Review of Artificial Intelligence Approaches in Omics Data Processing: Evaluating Progress and Challenges. International Journal of Mathematics, Statistics, and Computer Science, 2, 114-167.

Aydogmus, H. Y., Erdal, H. I., Karakurt, O., Namli, E., Turkan, Y. S., & Erdal, H. (2015). A comparative assessment of bagging ensemble models for modeling concrete slump flow. Computers and Concrete, 16(5), 741-757.

Control, D., Interventions, C. T. E. o. D., & Group, C. S. R. (2005). Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes. New England Journal of Medicine, 353(25), 2643-2653.

Filla, L. A., & Edwards, J. L. (2016). Metabolomics in diabetic complications. Molecular BioSystems, 12(4), 1090-1105.

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, 21, 137-146.

Galal, A., Talal, M., & Moustafa, A. (2022). Applications of machine learning in metabolomics: Disease modeling and classification. Frontiers in Genetics, 13, 1017340.

Hou, X.-W., Wang, Y., & Pan, C.-W. (2021). Metabolomics in diabetic retinopathy: a systematic review. Investigative ophthalmology & visual science, 62(10), 4-4.

Kozieł, K., & Urbanska, E. M. (2023). Kynurenine pathway in diabetes mellitus—novel pharmacological target? Cells, 12(3), 460.

Li, Z., Tong, J., Liu, C., Zhu, M., Tan, J., & Kuang, G. (2023). Analysis of independent risk factors for progression of different degrees of diabetic retinopathy as well as non-diabetic retinopathy among type 2 diabetic patients. Frontiers in Neuroscience, 17, 1143476.

Mohamed, Q., Gillies, M. C., & Wong, T. Y. (2007). Management of diabetic retinopathy: a systematic review. JAMA, 298(8), 902-916.

Patra, P., Disha, B., Kundu, P., Das, M., & Ghosh, A. (2023). Recent advances in machine learning applications in metabolic engineering. Biotechnology Advances, 62, 108069.

Sun, Y., Kong, L., Zhang, A.-H., Han, Y., Sun, H., Yan, G.-L., & Wang, X.-J. (2022). A hypothesis from metabolomics analysis of diabetic retinopathy: arginine-creatine metabolic pathway may be a new treatment strategy for diabetic retinopathy. Frontiers in endocrinology, 13, 858012.

Wang, Q., Cheng, H., Jiang, S., Zhang, L., Liu, X., Chen, P., . . . Wang, L. (2024). The relationship between diabetic retinopathy and diabetic nephropathy in type 2 diabetes. Frontiers in endocrinology, 15, 1292412.

Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. IEEE Transactions on knowledge and data engineering, 32(8), 1586-1594.

Yagin, B., Yagin, F. H., Colak, C., Inceoglu, F., Kadry, S., & Kim, J. (2023). Cancer metastasis prediction and genomic biomarker identification through machine learning and eXplainable artificial intelligence in breast cancer research. Diagnostics, 13(21), 3314.

Yau, J. W., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., . . . Grauslund, J. (2012). Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care, 35(3), 556-564.

Yun, J. H., Kim, J.-M., Jeon, H. J., Oh, T., Choi, H. J., & Kim, B.-J. (2020). Metabolomics profiles associated with diabetic retinopathy in type 2 diabetes patients. PloS one, 15(10), e0241365.

Zhang, T., Fu, Q., Wang, H., Liu, F., Wang, H., & Han, L. (2022). Bagging-based machine learning algorithms for landslide susceptibility modeling. Natural hazards, 110(2), 823-846.