NEU PRESS

FEN VE MÜHENDİSLİK BİLİMLERİ DERGİSİ

Vol: 6 No: 2 Year: 2024 Araştırma Makalesi/Research Article e-ISSN: 2667-7989

https://doi.org/10.47112/neufmbd.2024.52

İnflamatuar Bağırsak Hastalığında (İBH) 16s Sekans Verilerinden Oto-Kodlayıcı ile Muhtemel Biyobelirteç Keşfi

Ayşenur SOYTÜRK PATAT [1,2*] **Eda DAĞDEVİR** [2]

[1] Necmettin Erbakan University, Faculty of Science, Department of Molecular Biology and Genetics, Konya, Türkiye

[2] Kayseri University, Technical Sciences Vocational School, Department of Electronics and Automation, Program of Biomedical Device Technologies, Kayseri, Türkiye

Makale Bilgisi

Geliş Tarihi: 09.01.2024
Kabul Tarihi: 21.04.2024
Yayın Tarihi: 31.08.2024

Anahtar Kelimeler:
İnflamatuar Bağırsak Hastalığı (İBH),
Oto-kodlayıcı,
16S sekanslama.

ÖZET

Bir habitattaki fonksiyonel ve ekolojik dengeye (homeostasis) sahip mikrobiyal komünitelere mikrobiyota denir. Mikrobiyom ise bu komüniteyi oluşturan toplam genetik materyal ve bu genetik materyallerin çevre ile etkileşimine verilen isimdir. Mikrobiyotamız ile simbiyotik bir ilişki içinde olduğumuz yapılan çalışmalarla gösterilmiştir. Yeni nesil DNA dizileme teknolojilerinin yaygınlaşması ve hesaplama kabiliyeti yüksek bilgisayarların gelişmesi ile insan mikrobiyomunu ve sağlığa etkilerini keşfetmeye yönelik çalışmalar artmıştır. Yakın zamandaki araştırmalar birçok hastalıkla, kişinin mikrobiyom profilinin ilişkili olduğunu göstermiştir. Hastalık durumunda tedavinin yöntemini değiştirebilecek nitelikteki mikrobiyom çalışmaları, yüksek potansiyelli translasyonel çıktıları öncelikli alan haline gelmiştir. Ancak oldukça karmaşık olan bu verinin içerisinde hastalığın tanı ya da tedavisinde kullanılabilecek yüksek doğrulukta özniteliklerin bulunması oldukça zordur. Derin öğrenme teknikleri ise çeşitli çalışmalarda özellikle sınıflandırma alanında karmaşık verilerde ilham verici başarılar elde etmektedir. Oto-kodlama (AE) tekniklerinin ortaya çıkışı ise özellik seçme görevi için tasarlanmış bir sinir ağı mimarisidir. Bizim çalışmamızda veriyi yeniden temsil etmede önemli olarak görülen öznitelikler bir oto-kodlayıcısı tarafından belirlenmiş ve sadece belirlenen bu özniteliklerin gruplarda görülme sıklığına bakılarak İBH hastaları ve sağlıklı kontroller XGBoost algoritmasıyla %88.89 doğruluk değeri ile başarılı bir şekilde sınıflandırılmıştır. Önerilen yöntemle İBH hastalığını temsil eden mikrobiyol türler hastalığın tanısı için muhtemel biyobelirteçleri oluşturduğu düşünülmektedir.

Potential Biomarker Discovery with Auto-Encoder from 16s Sequence Data in Inflammatory Bowel Disease (IBD)

Article Info

Received: 09.01.2024
Accepted: 21.04.2024
Published: 31.08.2024

Keywords:
Inflammatory Bowel Disease (IBD),
Auto-encoder,
16S sequencing.

ABSTRACT

Microbial communities with functional and ecological balance (homeostasis) in a habitat are called microbiota. Microbiome is the name given to the total genetic material that makes up this community and the interaction of these genetic materials with the environment. Studies have shown that we have a symbiotic relationship with our microbiota. With the widespread use of new generation DNA sequencing technologies and the development of computers with high computational capabilities, studies to explore the human microbiome and its effects on health have increased. Recent studies have shown that a person's microbiome profile is associated with many diseases. Microbiome studies, which can change the method of treatment in case of disease, and high-potential translational outputs have become a priority area. However, it is very difficult to find high accuracy features that can be used in the diagnosis or treatment of the disease in this very complex data. Deep learning techniques, on the other hand, achieve inspiring success in various studies on complex data, especially in the field of classification. The emergence of auto-coding (AE) techniques is a neural network architecture designed for the feature selection task. In our study, the attributes that were considered important in representing the data were determined by an auto-encoder, and IBD patients and healthy controls were successfully classified with the XGBoost algorithm with an accuracy value of 88.89%, just by looking at the frequency of occurrence of these determined attributes in the groups. With the proposed method, microbial species representing IBD disease are thought to constitute possible biomarkers for the diagnosis of the disease.

To cite this article:
Soytürk Patat, A. & Dağdevir, E. (2024). Potential biomarker discovery with auto-encoder from 16s sequence data in inflammatory bowel disease (IBD). *Necmettin Erbakan University Journal of Science and Engineering*, *6*(2), 347-357. https://doi.org/10.47112/neufmbd.2024.52

***Sorumlu Yazar:** Ayşenur SOYTÜRK PATAT, *asoyturk@erbakan.edu.tr*

INTRODUCTION

Microbiome refers to the microbial communities living in an environment [1]. While this habitat sometimes creates an environmental example, sometimes it represents humans [2]. This community living in the human intestine is called the human intestinal microbiome and is considered our second brain [3]. We can be called a superorganism in this sense, as we host many different microbial species [4]. While there are approximately 20,000 genes in the human genome, there are approximately 2,000,000 protein-coding genes in the human intestinal microbiome [5]. This shows that it is very important to identify and understand these microbial communities that have much more gene coding capacity than ours. Various studies have shown that our intestinal microbiome, which has a high protein-coding capacity, changes in disease, reduces its taxonomic diversity, or changes in the frequency of microbial communities, that is, their relative abundance. So far, one of the most important limitations in microbiome studies is that the information about the genetic material that makes up this community has not been obtained completely and accurately [6,7,8]. Developing DNA sequencing technologies allow the genetic material of this community to be obtained, and the increase in microbiome studies is parallel with new generation DNA sequencing technologies. Figure 1 shows the numbers of the studies conducted. High-throughput new generation DNA sequencing technologies are used as basic tools for microbiome studies. Two approaches to sequencing are adopted here, metagenome sequencing is the uncovering of the entire DNA sequence of the microbial community. 16S sequencing is the determination of the 16S rRNA sequence, which contains variable and conserved regions for species assignment in bacteria. 16S rRNA sequencing is a standard approach used for species assignment. Sequencing tools can basically be classified as Sanger sequencing method (Primary Generation), Second Generation DNA Sequencing and Third Generation DNA Sequencing. Although the Sanger sequencing method is still used, second generation DNA (NGS) [9] sequencing tools are used more frequently due to the cost of sequencing and the long laboratory processes involved. Second Generation DNA sequencing tools adopt the reading approach while synthesizing DNA and read the DNA in short pieces. Reading multiple samples simultaneously, relatively large output data, short reading time and high accuracy are advantages for Second Generation DNA Sequencing platforms. Combining these short fragments and determining their functions requires a series of complex bioinformatics processes [10]. Third Generation DNA sequencing approaches aim to create longer reads in order to reduce the computational cost of Second Generation Sequencing approaches. Some of the approaches here pass the DNA strand through an electrical circuit and determine the graph of the changing mains current in the circuit and provide reading in this way [11]. Thanks to these technologies, large volumes of DNA data (typically several GB per sample) of microbial species that are not characterized in terms of phylogenetic or genetic function are produced for each microbiome sample. This microbiome data can be obtained for a large study group. In this way, species that differ for the disease state of interest can be identified, but an effective algorithm for classification and characterization has not yet been developed to make sense of this high-dimensional data. Developing computer technologies have provided us with computational capacity and the success of deep learning approaches, especially artificial intelligence applications, in image processing and text processing shows that these complex data are applicable and necessary to be used to make sense of them. Algorithms such as Naïve Bayes Classifiers [12], one of the basic machine learning approaches for the discovery of microbiological diversity from sequencing data, that is, for creating the taxonomic profile, and Random Forest [13] and Support Vector Machine [14] for the detection of disease dysbiosis and disease classification are now widely used. These algorithms are frequently used in other data and provide successful results [15,16]. However, machine learning-based approaches cannot identify important species with high enough accuracy through the complex microbiome profile. Deep learning methods are artificial intelligence applications and are known to perform much better than machine learning approaches when trained on enough data [17].

Auto-encoders (AE) are deep learning-based artificial neural networks, especially for dimensionality reduction and selection of important features. AE assigns the input data to a lower dimensional set of features (auto-encoding phase) to identify important features, and the generated set of features is assigned back to the input data (decoding phase) and the data is restructured. Artificial neural networks consist of neurons grouped into different layers. AE has a special hidden layer, called the hidden layer, which has fewer nodes than the input layer, forcing the network to develop a good representation of the input data [18]. The working architecture of AE is shown in Figure 2. Especially considering the success of deep learning in biological data, autoencoders are commonly used for feature determination. Our aim in this study is to identify important attributes with the auto-encoder method, one of the new deep learning algorithms that will model the microbiome data by taking into account the biological characteristics of the microbiome data, and to classify IBD, an intestinal disease, based on the determined attributes.

MATERIALS AND METHODS

Deep learning algorithms are making great progress these days, and one of these areas is the auto-encoder and decoder learning areas. An autoencoder algorithm will be used to determine specific microbial species for use in the diagnosis and treatment of the disease. This algorithm first converted microbiome data into small codes to represent it and thus found hidden attributes (microbial species) in the data. Then, the groups were classified using machine learning approaches based on the determined attributes. Python language and libraries written there were used to classify the applied approaches, necessary calculations and working group. The working order of the proposed method is given in Figure 3.

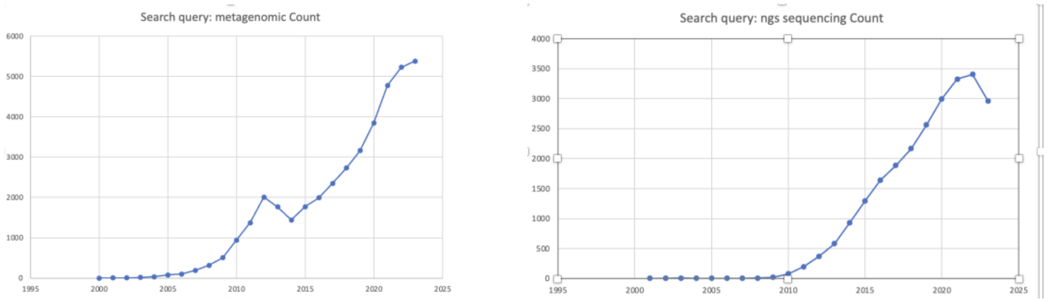
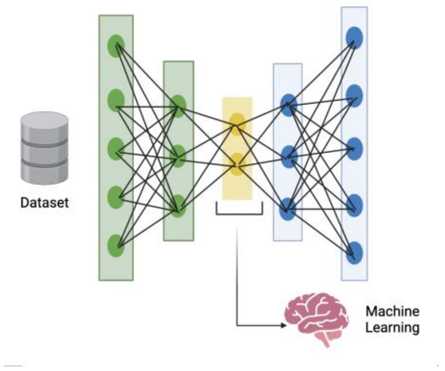


Figure 1
Number of Microbiome and NGS Based Studies in NCBI



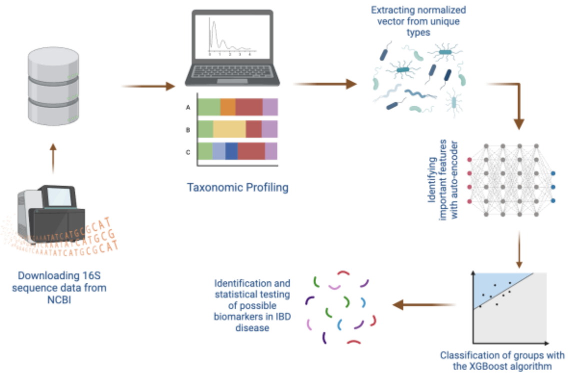

Extracting normalized vector from unique types
A
B
C
Taxonomic Profiling
Identifying important features with auto-encoder
Downloading 16S sequence data from NCBI
Identification and statistical testing of possible biomarkers in IBD disease
Classification of groups with the XGBoost algorithm

Figure 2
Auto-Encoder Working Architecture

Figure 3
Working Pipeline of the Proposed Method

Creating the Data Set

First, a data set consisting of IBD patients, an intestinal disease, and healthy controls was used to apply the developed method. For the discovery of biomarkers, a 16S rRNA sequencing dataset of 45 healthy controls and 131 IBD patients was created. The data set was generated using the Illumina MiSeq platform from the V3-V4 variable region. This dataset used is open access data and can be accessed at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA398089/.

Creating a Taxonomic Profile

The created data set consists of raw data and the taxonomy needs to be determined for each example. For this purpose, the Qiime2 tool, which is the most used bioinformatics tool in the 16S sequencing approach, was used. Qiime2 is an extensible, free, open source, and next-generation microbiome bioinformatics platform [19]. Taxonomic profiles for 176 specimens were created with this tool. 4144 different taxonomic species were uniquely found in 176 patients. To identify important attributes and classify them based on these attributes, a vector was created from the relative abundances of 4144 different taxa in each sample. Then, this created vector was normalized to avoid creating false positives.

Identifying Important Attributes

It is very difficult to identify specific species for disease within the complex gut microbiome data. The process of identifying important attributes from microbiome studies is based on classical machine learning approaches, but this is not sufficient to identify hidden attributes in complex data. For this purpose, the autoencoder block, which has recently yielded very successful results, was used to determine important features [20]. The operation of the autoencoder is shown in Figure 4. This algorithm will first convert the microbiome data into small codes that will represent it, and thus will be able to find hidden attributes (microbial species) in the data. Then, the decoding block in the algorithm will try to reconstruct the data from the attributes. 79 taxonomic species were identified as important for re-representing the data by the autoencoder.

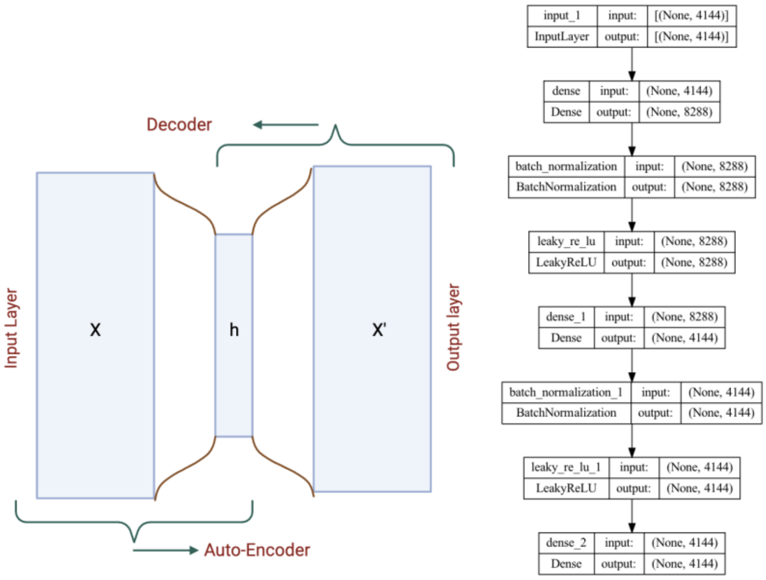

Decoder
Input Layer
X
h
X'
Output layer
Auto-Encoder
input_1 | input: | [(None, 4144)]
InputLayer | output: | [(None, 4144)]
dense | input: | (None, 4144)
Dense | output: | (None, 8288)
batch_normalization | input: | (None, 8288)
BatchNormalization | output: | (None, 8288)
leaky_re_lu | input: | (None, 8288)
LeakyReLU | output: | (None, 8288)
dense_1 | input: | (None, 8288)
Dense | output: | (None, 4144)
batch_normalization_1 | input: | (None, 4144)
BatchNormalization | output: | (None, 4144)
leaky_re_lu_1 | input: | (None, 4144)
LeakyReLU | output: | (None, 4144)
dense_2 | input: | (None, 4144)
Dense | output: | (None, 4144)

Figure 4
Auto-Encoder Layers.

Classification of IBD Patients and Healthy Controls

First, the success of frequently used classification algorithms was measured to see how successfully the groups were classified over the entire vector created from IBD and healthy controls. The usability of a model depends on its performance values, and the calculation of the accuracy value, which is the performance criterion we used for our study, is as follows:

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

True Positives (TP): Positive values that are predicted to be true. It refers to the number of cases where the actual class value is positive and the predicted class value is also positive.

True Negatives (TN): These are negative values that are predicted to be true. It refers to the number of cases where the actual class value is negative and the predicted class value is also negative.

False Positives (FP): The number of cases where the actual class is negative and the predicted class is positive.

False Negatives (FN): The number of cases where the actual class is positive and the predicted class is negative.

Then, the success of the XGBoost algorithm [21] in classifying IBD patients and healthy controls was compared with its previous success, using only the species represented as important in reconstructing the data by the autoencoder blog. Then, t-test was applied for two independent groups, IBD patients and healthy controls, and the accepted p value was taken as 0.05 as stated in the literature. 79 taxonomic species determined to be important are below the literature value.

RESULTS

The first thing to do in microbiome studies is to determine which taxa the DNA sequencing data in the environment, which consists of a very complex structure, consists of. The taxonomic profiling resulting from the Qiime2 tool run to determine taxonomic species can be seen in Figure 5. The given graph is a classic output of Qiime 2, with each different color representing a different taxonomic class. Samples are displayed on the x-axis (CD, UC, nonIBD), and the relative abundances of taxa indicated by different colors are shown on the y-axis. The fact that the bar plot is highly colored indicates the diversity of taxa in the sample and the scarcity of unclassified data can be observed.

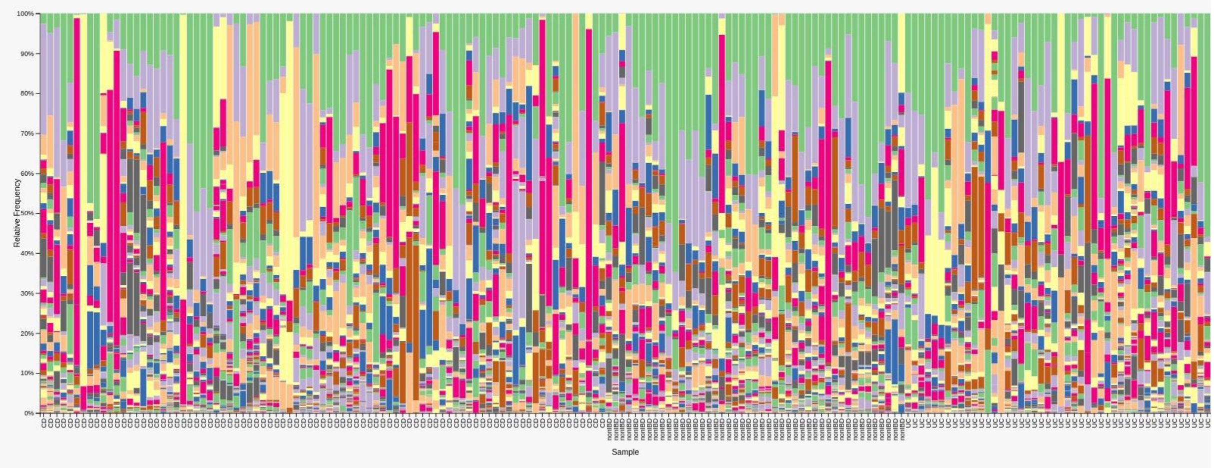


Figure 5

Taxonomic Profile of the Study Group

Samples are displayed on the x-axis (CD, UC, nonIBD), and the relative abundances of taxa indicated by different colors are shown on the y-axis.

Then, the success of the classification algorithms was tested on all data without feature selection. The accuracy values of machine learning models over all data are given in Table 1. The highest accuracy value is seen in the XGBoost algorithm. XGBoost is one of the implementations of gradient boosting algorithms [21]. The algorithm adds a new model to minimize the errors of the existing model and aims to increase classification accuracy. In this way, it can provide more successful results than other machine learning approaches.

The graph showing the operating performance of the auto-encoder in the transfer learning method used to determine important features and drawn according to loss functions is given in Figure 6. The taxonomic vector of the working group was separated into 80% training and 20% testing with the train-test separation function of scikit learn, a Python library.

Table 1
Classification Results

CLASSIFICATION ALGORITHMS	ACCURACY VALUE
Random Forest (RF)	80.56%
XGBoost (XGB)	**83.33%**
Logistic Regression (LR)	75.00%
Support Vector Machines (SVM)	71.88%
Decision Trees (DT)	68.75%
Gaussian Naive Bayes (GNB)	63.75%
K-Nearest Neighbors (KNN)	62.50%

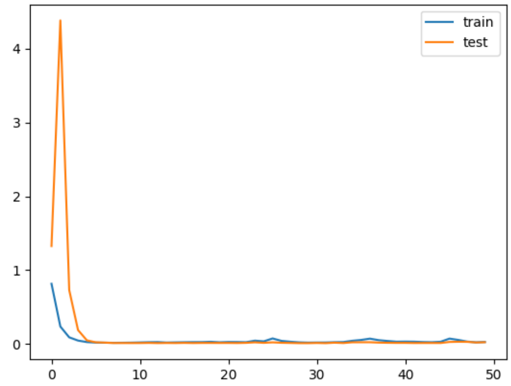


Figure 6
Training-Test Loss Function Graph

Some taxonomic classes selected by the auto-encoder created with the proposed method and considered important in re-representing the data are shown in table 2. With the autoencoder, 79 taxonomic species were selected as important features for data reconstruction. Data were reclassified only based on these selected species. The XGBoost algorithm increased the accuracy by 83.33% and 88.89%. Among the species listed in the table, *Clostridium ramosum* is an anaerobic, nonmotile, slender, spore-forming, gram-positive bacterium found in the human intestinal flora. *Clostridium ramosum* is rarely associated with disease in humans and has a low GC content. Therefore, its relationship with IBD is quite surprising. However, what really matters is the potential of these species as biomarkers. Furthermore, the key observation here is that some species identified in the data were not determined using classical statistical methods, as stated in the publication where the data were obtained [25].

Table 2

Some of The İmportant Taxonomic Classes Identified by The Autoencoder

TAXONOMIC CLASSES
k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Clostridium; s__ramosum
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__[Ruminococcus]; s__
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Dorea; s__
k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__[Odoribacteraceae]; g__Odoribacter; s__
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Faecalibacterium; s__prausnitzii
k__Bacteria; p__Verrucomicrobia; c__Verrucomicrobiae; o__Verrucomicrobiales; f__Verrucomicrobiaceae; g__Akkermansia; s__muciniphila
k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Roseburia; s__inulinivorans
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Coprococcus; s__
k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__ovatus

DISCUSSION AND CONCLUSIONS

Inflammatory Bowel Disease (IBD) is a group of diseases in which there is chronic inflammation in the digestive system. People with IBD may experience intermittent or persistent symptoms that make it difficult to perform daily activities [22]. The prevalence of IBD is increasing, and IBD is one of the most common diseases in the United States and Europe. The causes, diagnosis and treatment methods of this disease, which is observed at a very high rate, are still being investigated [23]. The development of new generation DNA sequencing methods and the obtaining of DNA sequences of non-culturable microorganisms are of great importance in finding microbial flora elements that can be used in the diagnosis and treatment of this disease, which has been determined to be in close relationship with the intestinal flora in previous studies. The beginning of microbiome-based studies is to understand and define the complex microbiome profile from the total DNA sequence and try to find its relationship with the host by determining its functions. However, finding the taxa that make up the complex microbiome data, determining the importance of the obtained taxa, or understanding the functions of the taxa in the host remains challenging. Developments are continuing in many areas such as microbiome-based diagnosis, treatment, nutrition, and drug use, but determining the taxa that truly contribute to the dysbiosis of the disease from this complex data set remains a problem. The developing computers with high computational capabilities and the increasing amount of data allow the development of this field. The auto-encoding method, which is quite popular among these technologies, shows higher performance than classical dimension reduction or feature determination algorithms [24]. It is observed that the decoding blog we used in our study increases the accuracy of the classifiers through the features determined. In addition, many previously unidentified subspecies were found with this method as an important feature when classifying groups in the 16S sequencing data, which comprised the training and test data of the study. Some of these species are *Roseburia inulinivorans*, *Akkermansia muciniphila*, *Collinsella aerofaciens*. These identified species are thought to constitute possible biomarkers to be used in the diagnosis of this disease [25]. The obtained species are thought to constitute potential biomarkers, but it is considered necessary to expand the study group to determine whether these species are truly biomarkers of the disease. Additionally, adding samples from different regions with the same disease to the study group is crucial for the reliability of the results [26]. In this context, it is believed that adding more layers to the encoder structure will increase the success of the proposed method. In future studies, diagnosis of the disease can be made using kits developed based solely on these species, and the course of the disease can be altered by diets prepared to change the abundance of these identified species. Furthermore, it sheds light on the development of probiotics, which have made significant progress in this field. The proposed method can be applied not only for IBD but also for many other diseases.

Ethics Committee Approval

The data in the study is open source and does not require ethics committee approval.

Authors Credit

Research Design (CRediT 1) Author 1 (%90) - Author 2 (%10)
Data Collection (CRediT 2) Author 1 (%90) - Author 2 (%10)
Research - Data Analysis - Validation (CRediT 3-4-6-11) Author 1 (%90) - Author 2 (%10)
Writing the Article (CRediT 12-13) Author 1 (%100) - Author 2 (%00)
Revision and Improvement of the Text (CRediT 14) Author 1 (%50) - Author 2 (%50)

Financing

This study was supported by TÜBİTAK with program 2209 and project number 1919B01221710.

REFERENCES

[1] L.-H. Lee, S. H. Wong, S.-F. Chin, V. Singh, and N.-S. Ab Mutalib, Editorial: Human Microbiome: symbiosis to pathogenesis, *Frontiers in Microbiology*. 12 (2021), doi: 10.3389/fmicb.2021.605783.

[2] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, The human microbiome project, *Nature*, 449(7164) (2007), 804–810, doi: 10.1038/nature06244.

[3] C. G. Buffie, M.Equinda, L.Lipuma, A.Gobourne, A.Viale, C.Ubeda, J.Xavier, E.G.Pamer, Profound alterations of intestinal microbiota following a single dose of clindamycin results in sustained susceptibility to Clostridium difficile-induced colitis, *Infection and Immunity*. 80(1) (2012), 62–73, doi: 10.1128/IAI.05496-11.

[4] A. Gundogdu, Bir 'Süper Organizma' olarak insan; mikrobiyomun genetik kontrolü, *Türk Mikrobiyoloji Cemiyeti Dergisi*. 46(4) (2016), doi: 10.5222/TMCD.2016.147.

[5] X. Yang, L. Xie, Y. Li, C. Wei, More than 9,000,000 unique genes in human gut bacterial community: estimating gene numbers inside a human body, *PLoS ONE*. 4(6) (2009), doi: 10.1371/journal.pone.0006074.

[6] G. Berg, D. Rybakova, D. Fischer, T. Cernava, M.C. Vergès, T. Charles, X. Chen, L. Cocolin, K. Eversole, G.H. Corral, M. Kazou, L. Kinkel, L. Lange, N. Lima, A. Loy, J.A. Macklin, E. Maguin, T. Mauchline, R. McClure, B. Mitter, M. Ryan, I. Sarand, H. Smidt, B. Schelkle, H. Roume, G.S. Kiran, J. Selvin, R.S.C. Souza, L. van Overbeek, B.K. Singh BK, M. Wagner, A. Walsh, A. Sessitsch, M. Schloter, Microbiome definition re-visited: old concepts and new challenges, *Microbiome*. 8(1) (2020) 119. doi: 10.1186/s40168-020-00875-0

[7] F & H Löchel, D. Heider, Comparative analyses of error handling strategies for next-generation sequencing in precision medicine, *Scientific Reports*. 10(1) (2020), 5750. doi: 10.1038/s41598-020-62675-8

[8] A.L. Lapidus, A.I. Korobeynikov, Metagenomic Data Assembly - The Way of Decoding Unknown Microorganisms, *Frontiers in Microbiology*. 12 (2021), 613791. doi:10.3389/fmicb.2021.613791

[9] S. Jünemann, N. Kleinbölting, S. Jaenicke, C. Henke, J. Hassa, J. Nelkner, Y. Stolze, S.-P. Albaum, A. Schlüter, A. Goesmann, A. Sczyrba, J.Stoye, Bioinformatics for NGS-based metagenomics and the application to biogas research, *Journal of Biotechnology*. 261 (2017), 10–23. doi: 10.1016/j.jbiotec.2017.08.012.

[10] J. Reinartz, E. Bruyns, J. Z. Lin, T. Burcham, S. Brenner, B. Bowen, M. Kramer, R. Woychik, Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms, *Briefings in Functional Genomics & Proteomics*, *1*(1), (2002), 95–104. doi:10.1093/bfgp/1.1.95

[11] E.L. van Dijk, Y. Jaszczyszyn, D. Naquin, C. Thermes, The Third Revolution in Sequencing Technology, *Trends in Genetics*. *34*(9) (2018), 666–681. doi:10.1016/j.tig.2018.05.008

[12] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy, *Applied and Environmental Microbiology*, 73(16), (2007), 5261–5267. doi: 10.1128/AEM.00062-07.

[13] F.-H. Karlsson, V. Tremaroli, I. Nookaew, G., Bergström, C.-J. Behre, B. Fagerberg, J. Nielsen, F. Bäckhed, Gut metagenome in European women with normal, impaired and diabetic glucose control, *Nature*. *498*(7452) (2013), 99–103. doi: 10.1038/nature12198.

[14] K. Forslund, F. Hildebrand, T. Nielsen, G. Falony, E. Le Chatelier, S. Sunagawa, E. Prifti, S. Vieira-Silva, V. Gudmundsdottir, H.-K. Pedersen, M. Arumugam, K. Kristiansen, A.Y. Voigt, Vestergaard, H., Hercog, R., P. I. Costea, J. R. Kultima, J. Li, T. Jørgensen, F. Levenez, O. Pedersen, Corrigendum: Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota, *Nature*. *545*(7652), (2017), 116. doi: 10.1038/nature22318.

[15] H. İ. Ayaz ve Z. Kamışlı Öztürk, Shilling attack detection with one class support vector machines, *Necmettin Erbakan University Journal of Science and Engineering*. 5(2) (2023), 246–256. doi: 10.47112/neufmbd.2023.22.

[16] M. Hacıbeyoğlu, M. Çelik, Ö. Erdaş Çiçek, Energy efficiency estimation in buildings with K nearest neighbor algorithm, *Necmettin Erbakan University Journal of Science and Engineering*. 5(2) (2023), 65–74. doi: 10.47112/neufmbd.2023.10.

[17] A. Mathieu, M. Leclercq, M. Sanabria, O. Perin, A. Droit, Machine learning and deep learning applications in metagenomic taxonomy and functional annotation, *Frontiers in Microbiology*. *13* (2022). doi:10.3389/fmicb.2022.811495

[18] P. Li, Y. Pei, and J. Li, A comprehensive survey on design and application of autoencoder in deep learning, *Applied Soft Computing*. 138 (2023). doi:10.1016/j.asoc.2023.110176

[19] E. Bolyen, J. R. Rideout, M. R. Dillon, N. A. Bokulich, C. C. Abnet, G. A. Al-Ghalith, H. Alexander, E. J.Alm, M. Arumugam, F. Asnicar, Y. Bai, J. E. Bisanz, K. Bittinger, A. Brejnrod, C. J. Brislawn, C. T. Brown, B. J. Callahan, A. M. Caraballo-Rodríguez, J. Chase, E. K. Cope, J. G. Caporaso, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, *Nature Biotechnology*. *37*(8) (2019), 852–857. doi: 10.1038/s41587-019-0209-9.

[20] P.P. Líndez, J. Johansen, S. Kutuzova, Adversarial and variational autoencoders improve metagenomic binning, *Commununications Biology*. 1073 (2023). doi:10.1038/s42003-023-05452-3

[21] T. Chen and C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: *Volume 16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, 785–794. doi: 10.1145/2939672.2939785.

[22] J. Carriere, A. Darfeuille-Michaud, HTT. Nguyen, Infectious etiopathogenesis of Crohn's disease, *World Journal of Gastroenterology*, (2014). doi: 10.3748/wjg.v20.i34.12102

[23] S.C. Ng, C.N. Bernstein, M.H. Vatn, P.L. Kakatos, E.V. Loftus, C. Tysk, Geographic variability and environmental risk factors in inflammatory bowel disease, *Gut*, 62, (2013),630–49. doi: 10.1136/gutjnl-2012-303661

[24] M. Leon, T. Markovic, and S. Punnekkat, Feature Encoding with Autoencoder and Differential Evolution for Network Intrusion Detection Using Machine Learning, in: *GECCO '22: Proceedings of the Genetic and Evolutionary Computation Conference Companion*, New York, NY, USA, 2022, 2152–2159. doi: 10.1145/3520304.3534009.

[25] J. Lloyd-Price, C. Arze, A.N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N.J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, A, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, C. Huttenhower, Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases, *Nature*. 569(7758) (2019) 655–662. doi:10.1038/s41586-019-1237-9

[26] G. S. Navgire, N. Goel, G. Sawhney, M. Sharma, P. Kaushik, Y.K. Mohanta, T. K. Mohanta, A. Al-Harrasi, Analysis and ınterpretation of metagenomics data: an approach, *Biological Procedures Online*. 24(1) (2022) 18. doi:10.1186/s12575-022-00179-7