Using Data Mining for Analysis and Classification of Necrotizing Enterocolitis  
Guy Shpringer, *Department of Software and Information Systems Engineering  
Ben Gurion* *University of the Negev, Beer Sheva, Israel*  
guyshp@post.bgu.ac.il

**Abstract**

Necrotizing Enterocolitis (NEC) is an acute inflammatory-gastrointestinal disease which affecter premature infants. These days, the main root causes for the disease have not been discovered yet, which affected also the treatment methods for this mortified disease. The last advancements in the artificial intelligence and machine learning fields, these tools may be implemented on microbiome data to try and discover different important features for differentiation between affected and non-affected infants. In this research, we present features extraction for microbiome gut data sampled from affected and non-affected infants. In the results it can be seen that the features extraction in this research help with the differentiation and might be also enhance and improve on future works where another features will be extracted, such as operational taxonomical units (*OTUs*) and hierarchical features space using.

1. INTRODUCTION

Necrotizing Enterocolitis (*NEC*) is an acute inflammatory gastrointestinal disease affecting premature infants. Intestinal microbial [1], [2]. The prevalence of the disease is about 7% among infants born in <32 weeks gestation with a birth weight of <1,500 grams in the US [3]. Treatment strategies for this disease, have remained limited and haven’t changed significantly for decades [4].

Despite the NEC field spending great amounts of sources in early diagnosis, still it is not clear what are the main causes for the disease, and most of the biomarkers discovered so far are not sufficient by themselves [5]. Looking at the Artificial Intelligence (*AI*) and Machine Learning (*ML*) tools and techniques, easily it can be seen that they may be implemented and help in the research for the causes of the disease and early treatment, there is a high potential to help the medical community. AI has the ability to identify the importance of different features in the input data, which can help to define better the disease and the root causes for it.

In this work we will examine the ability to extract features from the microbiome of affected samples, and to classify according to the sampled microbiome between affected and non-affected samples. This would be an enhancements to different possibilities to improve classifiers in order to help and reveal another mysterious takeaways about the deadly disease.

1. RELATED WORK

Olm et al [6] processed the microbiome data to the operational taxonomic units of the samples, to investigate the combination of the guy microbiome in NCE and non-NCE infants, through their 60 days of life. They also looked at the bacterial replication rates before the NEC development, and from the conclusion suggested the replication rate as a differentiation between the NEC and non-NEC infants. Another critical conclusion in this research was about the complexity of the disease, and the conclusion that it is multi-factorial disease, as no single predictor was enough to differentiate between NEC and non-NEC infants. The design of Olm’s study is very similar to the one of Warner’s et al [7].

Hooven et al. [8], as Olm et al., used also the taxonomy and the relative abundance of different classes in the microbiome. Together with the taxonomy data (after pre-processing it to handle zero values, log-transformation and feature selection), also the demographic data of the patients were assigned as inputs for the neural network in the model.

Lin et al. [9] tried to improve Olm’s [6] and Warner’s previous works. In this research they are looking at the special property of microbiome data, as it is constructed of hierarchical featured space, where each bacterium might be classified using different descriptors, i.e. kingdom, phylum, class, order, family, genus and species levels. The data was an input for the multiple instance learning (*MIL*) architecture of their model and resulted with a prediction model for the NEC risks.

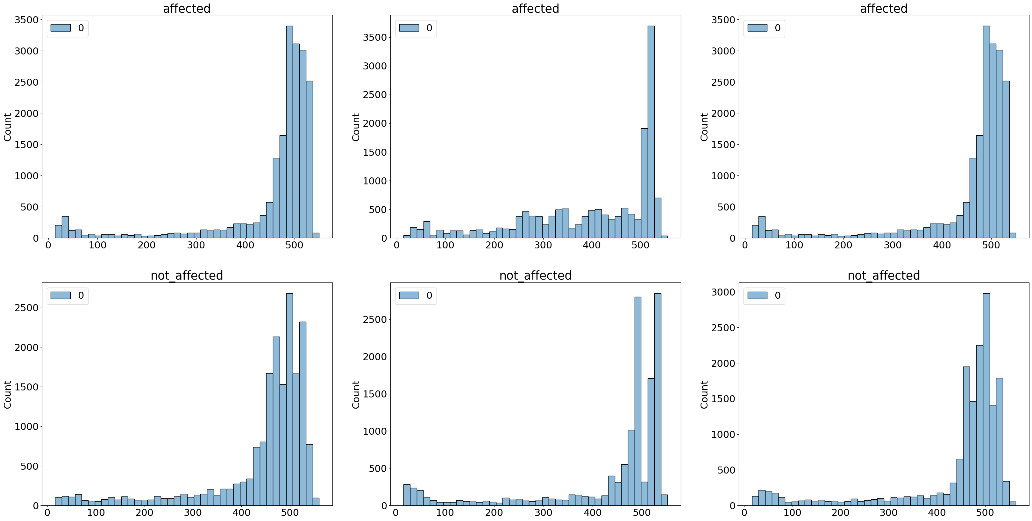
One of the most modern works in the field is of Casaburi et al. [10]. Their research is the first one to link between premature newborn gut dysbiosis, the acquired newborn diseased NEC and a mechanism of biochemical mediated intestinal injury conferred through microbial fermentation and the over production of formate [10]. The link they discovered might help us understand and reveal new ways and methods to differentiate and predict NEC illness, in non-invasive ways, and improve the different machine-learning based methods mentioned in the works above.

Figure 1: Histogram plots of sequences lengths in a sample file. Top row shows the histograms for 3 different samples of affected infants, and the lower bottom for 3 samples of non-affected infants.

1. METHODS
2. *Data collection and analysis*

The data used in this research was acquired from the Human Microbiome Project (*HMP*) database [11]. The data was collected originally for a comprehensive study for the NEC microbiome and used control group of non-affected as well as affected infants’ microbiome to perform the analysis.

The data was collected using the Roche 454 system, from 163 subjects, resulting in 3055 samples of 16S microbiome mapping. For this research particularly, we used only 33 microbiome samples form affected infants, and 127 microbiome samples of non-affected infants (due to unavailability of the whole database online). Also, no metadata was used for the analysis done in this research (pure DNA sequencing from the samples).

1. *USUM and UMAP*

USEARCH is a useful tool in the bioinformatics field, which was first introduced by Edgar and Bateman in 2010 [12]. Using this tool for analysis of microbiome files to extract taxonomies or component extraction methods, many other computational methods were developed. Uniform Manifold Approximation (*UMAP)* is a dimension reduction method, like t-SNE method, but also for general non-linear dimension reduction [13].

USUM is a python package using for plotting DNA sequences similarity embeddings used UMAP and USEARCH libraries. In this package, the distance matrix is calculated between each sequence in the two samples we want to compare between (therefore used the USEARCH library). The distance matrix is embedded as a metric using UMAP to calculate the principal features and components of the two samples, and then the results is shown in 2-dimensional plot to compare the two samples.

1. *Data Feature Extraction*

In this section we will articulate the details of our analytical approach. In this research, the given inputs are ‘.fasta’ files of affected and non-affected infants with NCE. The ‘.fasta’ file format is a text-based format representing the nucleotide sequences in the 16S sequence mappings. The file format is consisted of different entries, each one of them has a line representing the scanning parameters and the sample data (following the ‘>>>’ sign), and the second line representing sequence itself, represented by the letter of the different nucleotides (A, C, G & T for DNA sequences).

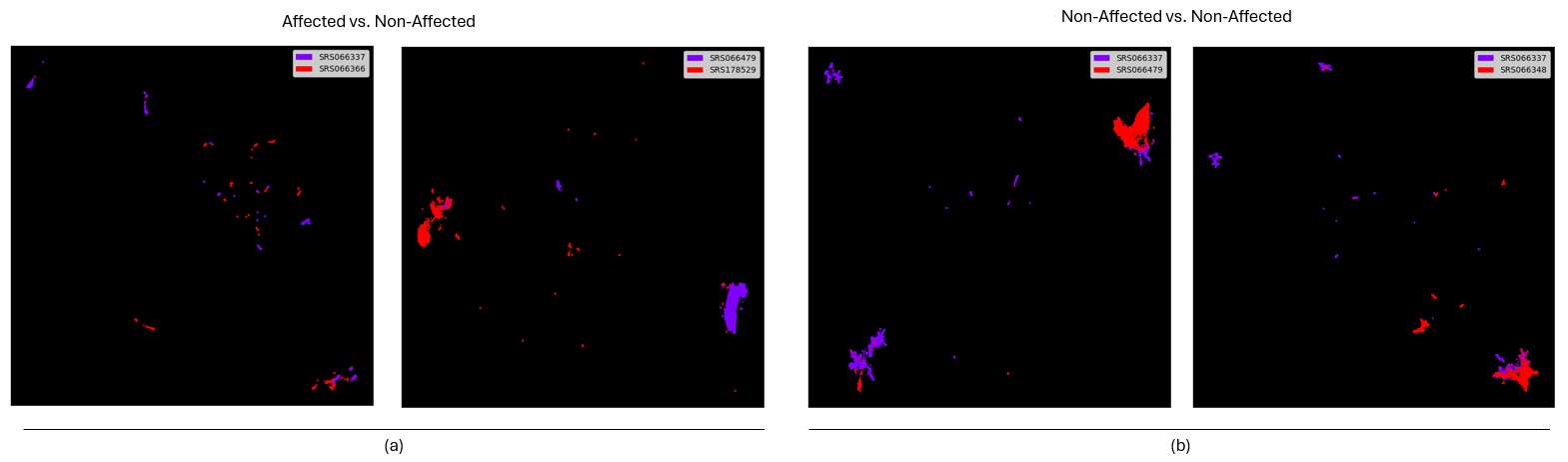
To visually looking for features between the affected and non-affected samples, we constructed histograms of the sequence’s lengths. We divided the histogram to 40 bins and compared the samples between the samples (both in-groups, affected-affected samples, and cross-groups).

Figure 2: The UMAP plots for different couples of samples. (a) Two plots, each one for couple of two affected versus non-affected infant’s sample. (b) 2 plots representing two couples of non-affected samples. one versus the other.

For the analysis of the data, we are using several approaches. The MathFeature is a feature extraction package for using with DNA, RNA and protein sequences data. The resulted features from applying the package on the microbiome data are both mathematical descriptors and conventional descriptors for the sequences. The mathematical descriptors including features based on genomic signal processing (GSP) using Fourier transform (FT), and the different features that can be extracted of it (first and second peaks of the FT, amplitude, standard deviation, mean value, etc.). As there are many sequences in each sample file, we looked for the mean values of all the features across the same ‘.fasta’ file of the sample.

The second features extraction from the MathFeature package was using the *k*-mers method analysis [14]. The DAN *k*-mers, are the DNA ‘words’ of length *k* in the DNA sequence. The spectrum of the *k*-mers might reveal relevant data for the sequenced microbiome sample. As in the data we have ‘.fasta’ files with many sequences from the same sample. The *k*-mers values were calculated for each sequence independently, and then the feature table we build concluded the mean value, the standard deviation and the maximal value of the *k*-mers frequency in the whole file (among all the mapped DNA sequences of the same sample).

1. RESULTS ANALYSIS
2. *Histogram analysis*

To understand the importance of the sequence lengths of the samples, we drew the histograms of the lengths, to see the main differences between the affected and non-affected samples. Fig.1 shows examples of 6 histograms of the samples, 3 affected and 3 non-affected, with their histograms:

As seen in Fig. 1, there are some bars that differentiates the histograms between affected and non-affected used samples, where it is visually around the 500 lengths. At the affected samples we can distinguish a bell shape around the 500, while in the non-affected samples we can see it is not increasing and the increasing but have some decreasing points in the shape of the bars around the 500 (in 510 and 480).

1. *USUM analysis*

In Fig.2 we can see the USUM plots that were done between affected and non-affected infants’ samples (Fig.2(a)) and also non-affected-to-non-affected analysis as a control (Fig.2(b)).

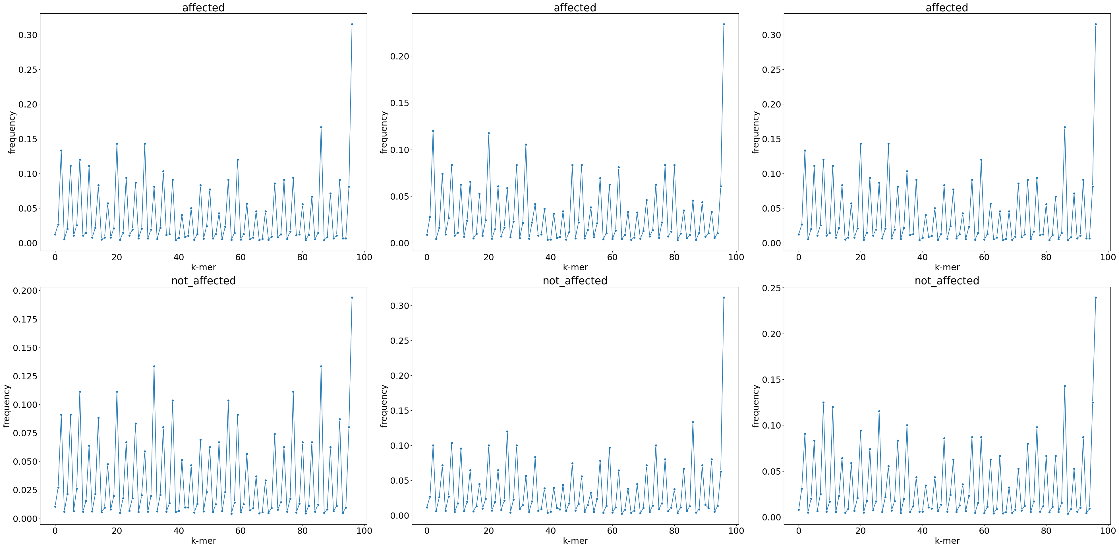


Figure 3: k-mers spectrum graphs for different samples sequences. In Upper row, spectrum for the 3 different affected samples, and in lower row spectrums of 3 different non-affected samples.

תמונה שמכילה תרשים, צילום מסך, טקסט, קו

התיאור נוצר באופן אוטומטיVisually from the UMAP plots seen in Fig2, it is possible to see the overlapping (or similarity and closeness) between the samples, which were randomly chosen from the sequences sample files, is clearer visually in the comparison between the non-affected and non-affected samples than in the comparison between the affected and non-affected samples. On this research there were no features extracted from those maps, but on future works this might discover and show another relevant features for the classifiers between the non-affected and affected infants.

Figure 4: Features analysis extracted from the accumulate nucleotide frequency spectrum.

1. *Features Analysis*

As mentioned, the *k*-mers spectrum was calculated to each of the samples, to see if the extracted features from the spectrums might be relevant and important for the classifier. In Fig.3, the *k*-mers spectrums of 6 samples (3 of affected infants and 3 of non-affected infants)

are shown. It is possible to visualize the change   
in the spectrum between the two groups. From the graphs itself, it is hard to distinguish relevant information and differentiators between the affected and non-affected samples.

For the spectrum analysis features (Fig.4), although a lot of the features have similar values (in their mean value), we can see there are some features with relevant difference between the affected and non-affected samples. For example, for the variance coefficient value, amplitude, maximum and average we can see difference they might be interpreted as an option to differentiate between the samples. All these features mentioned above, are ones that are interpreted for the spectrum shape and are holding a data about the dispersion of the different spectrum values around the average.

1. *Classification evaluation*

Using the features extracted in the analysis, both using histograms and the MathFeature package, we build a classificatory using the CatBoost package. In this classificatory, we used the whole features extracted of the sequences. The data was split to train and test sets, with proportions of 77% and 33% respectively. The classifier model was trained with 100 iterations and depth of 15. The accuracy resulted of the model train was 0.86. and the corresponding confusion matrix of the model is shown in Fig.5.

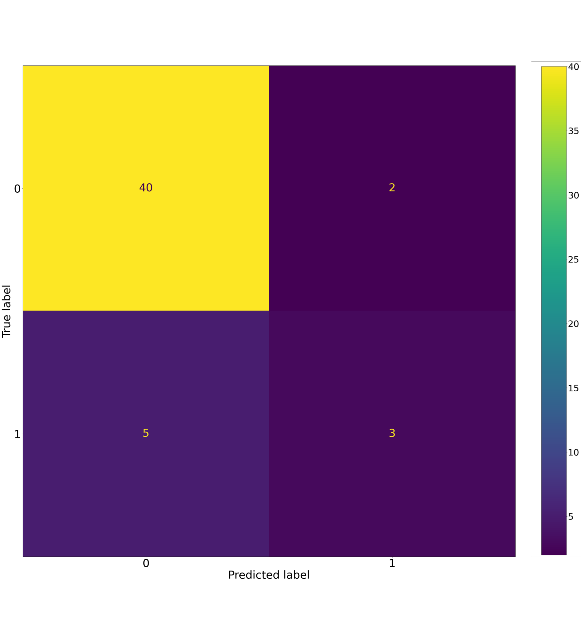


Figure 5: Confustion matrix for the modeled classifier

It is possible to see of the confusion matrix that the classifier works very-well on the non-affected data, and with 50% accuracy for the affected samples. These results make sense, as the amount of non-affected samples is bigger than the non-affected (and we didn’t use augmentation on the sequences data in this research).

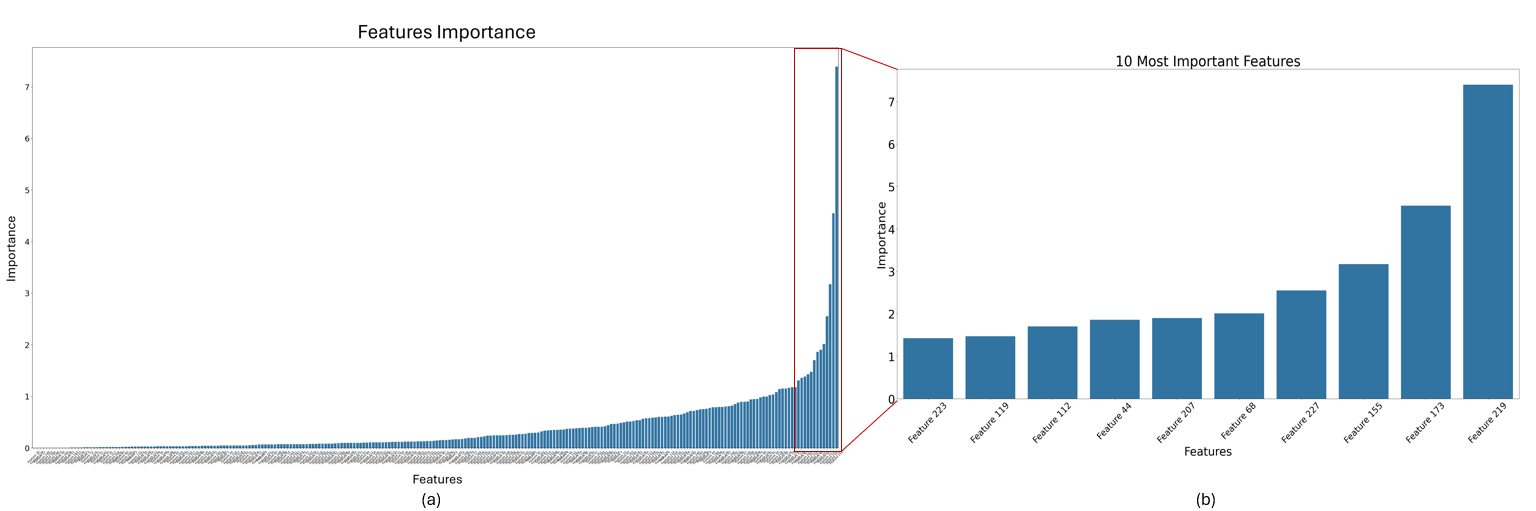
To understand the importance of each feature in the model for the prediction, the features’ importance values were plotted in Fig.6.

Figure 6: Importance of different features for the classifier. (a) The importance of all the features in the classifier, (b) the importance values only for the 10-most important features

The features we are using in the classification process are divided to 3 groups, i.e. features 0-39 (for the histogram’s 40 bins values of sequence lengths), features 40-58 (for the spectrum analysis features), and features 59-253 for the features resulted from the *k*-mers analysis. As seen in Fig.6, in the group of the 10 most important features both the second and third groups are featuring, which may give us some insights for their importance for differentiating between affected and non-affected samples.

1. CONCLUSIONS AND FUTURE DIRECTIONS

From the results we can see a clear potential to track the NCE using stool microbiome of the infants. On future works, there is an opportunity to extend the bank of features using for classification, e.g. operational taxonomic units (*OTU*) abundance tables and amplicon sequence variants (*ASV*) tables, etc., in order to use also the hierarchical feature space as already used in relevant other works in this field.

On the data side of the research, next recommended stages might be to conduct extensive data collecting from patients, in different stages of the development of disease. That way, one might be able to build a predictor both for potential upcoming illness in infants, but also building the pseudo-time trajectory of the illness. Using the pseudo-time trajectory, we might discover new logics of the illness development and maybe target different stages of the disease with the appropriate treatment to reduce mortality.

1. BIBLIOGRAPHY

[1] J. Duchon, M. E. Barbian, and P. W. Denning, “Necrotizing Enterocolitis,” *Clin Perinatol*, vol. 48, no. 2, pp. 229–250, Jun. 2021, doi: 10.1016/J.CLP.2021.03.002.

[2] J. Bellodas Sanchez and M. Kadrofske, “Necrotizing enterocolitis,” *Neurogastroenterology and motility*, vol. 31, no. 3, Mar. 2019, doi: 10.1111/NMO.13569.

[3] J. Neu and W. A. Walker, “Necrotizing Enterocolitis,” *New England Journal of Medicine*, vol. 364, no. 3, pp. 255–264, Jan. 2011, doi: 10.1056/NEJMRA1005408.

[4] E. Wejryd *et al.*, “Low Diversity of Human Milk Oligosaccharides is Associated with Necrotising Enterocolitis in Extremely Low Birth Weight Infants,” *Nutrients*, vol. 10, no. 10, Oct. 2018, doi: 10.3390/NU10101556.

[5] B. D. Garg, D. Sharma, and A. Bansal, “Biomarkers of necrotizing enterocolitis: a review of literature,” *The Journal of Maternal-Fetal & Neonatal Medicine*, vol. 31, no. 22, pp. 3051–3064, Nov. 2018, doi: 10.1080/14767058.2017.1361925.

[6] M. R. Olm *et al.*, “Necrotizing enterocolitis is preceded by increased gut bacterial replication, Klebsiella, and fimbriae-encoding bacteria,” *Sci. Adv*, vol. 5, pp. 5727–5738, 2019.

[7] B. B. Warner *et al.*, “Gut bacteria dysbiosis and necrotising enterocolitis in very low birthweight infants: a prospective case-control study,” *The Lancet*, vol. 387, no. 10031, pp. 1928–1936, May 2016, doi: 10.1016/S0140-6736(16)00081-7.

[8] T. A. Hooven, A. Lin, and A. Salleb-Aouissi, “Multiple Instance Learning for Predicting Necrotizing Enterocolitis in Premature Infants Using Microbiome Data,” 2020, doi: 10.1145/3368555.3384466.

[9] Y. C. Lin, A. Salleb-Aouissi, and T. A. Hooven, “Interpretable prediction of necrotizing enterocolitis from machine learning analysis of premature infant stool microbiota,” *BMC Bioinformatics*, vol. 23, no. 1, pp. 1–29, Dec. 2022, doi: 10.1186/S12859-022-04618-W/FIGURES/7.

[10] G. Casaburi *et al.*, “Metabolic model of necrotizing enterocolitis in the premature newborn gut resulting from enteric dysbiosis,” *Front Pediatr*, vol. 10, p. 893059, Aug. 2022, doi: 10.3389/FPED.2022.893059/BIBTEX.

[11] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, “The Human Microbiome Project”, doi: 10.1038/nature06244.

[12] R. C. Edgar and A. Bateman, “Search and clustering orders of magnitude faster than BLAST,” *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, Oct. 2010, doi: 10.1093/BIOINFORMATICS/BTQ461.

[13] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Feb. 2018, Accessed: May 18, 2024. [Online]. Available: http://arxiv.org/abs/1802.03426

[14] B. Chor, D. Horn, N. Goldman, Y. Levy, and T. Massingham, “Genomic DNA k-mer spectra: Models and modalities,” *Genome Biol*, vol. 10, no. 10, pp. 1–10, Oct. 2009, doi: 10.1186/GB-2009-10-10-R108/FIGURES/5.