Feature Enhancement and Chaining of Deep Nueral Networks in Colorectal Cancer Classifaction based on Gut-Microbiome Data

Mwenge Mulenga
Business Studies Division,
National Institute of Public Administration,
Lusaka, Zambia
Mwenge.research@gmail.com

Musa Phiri
School of Engineering and Technology,
Mulungushi University,
Kabwe, Zambia
phirimusa@live.com

Luckson Simukonda
School of Engineering and Technology,
Mulungushi University,
Kabwe, Zambia
lucksonsimukonda@hotmail.com

Abstract— **Colorectal Cancer (CRC) is among the top three cancers in world. The current clincal methods for CRC detection have several limitations which range from low accuracy, discomfort and high costs. Availability of next generation sequencing (NGS) technology has opend an opportunity for non invasive detection of CRC which uses gut-microbiome abundance in stool samples. The high dimension of sequence base microbiome data has prompted research interest in the application of machine learning (ML) in order to classify host disease based on microbial counts. However the classification performance of ML methods such data is still limited by factors shuch as high dimensionality and data imbalance. Therefore, in this paper, we propose a deep nueral network based method that combines feature extension and feature and chained execution of deep neural network to improve CRC classifaction based on gut microbiome in stool samples. The proposed method scored a mean area under the receiver operating characteristics curve (AUC) of approximately 95.4%, which is higher than state-of-the-art methods. The proposed method can positively contribute to the development of robust diagnostic and prognostic methods for CRC.**

Keywords—DNN, feature expansion, feature extension, Chaining

Introduction

Colorectal cancer (CRC) is the third most common cancer in the world [1] and among non sex related cancers, it has the second highest mortality rate after lung cancer [2]. However, early detection of the cancer can improve survival chances of a patient [3].

Also, several methods exist for CRC screening of which colonoscopy is the gold standard [4]. In the recent past there has been a growing research interest in the use of gut-microbiome in non-invasive detection of CRC. Research has shown that there is a higher density of microbiome in regions of the human coronary tract that is characterised with inflammations. Higher populations of gut microbiome have also been identified in faecal matter of CRC patients as compared to non-patients [5]. On one hand the availability of next generation sequencing technology has made it possible to generate gut-microbiome based data [6]. On the other hand, development of machine learning (ML) algorithms have created an opportunity for the development of computer aided diagnostic tools. However, conventional ML methods have several limitations such as heavy dependency on manual feature selection to improve predictive performance [7][8].

Therefore, this study seeks to propose a deep neural network (DNN) based method CRC classification that uses the concept of noise replicates for feature augmentation to improve CRC prediction.

The remainder of this paper is organised as follows. Section 2 covers related works, while materials and proposed methods are described in section 3. The results are outlined in section 4, which is followed by the discussion in section 5. Finally, the conclusion and future works are presented in section 6.

Literature Review

Feature engineering has been extensively used to improve the prediction performance of microbiome-based tasks. The work in [9] applied feature extraction to diminish the effect of noise on the learning process. In contrast the work in [10] proposed a method that uses aggregate features and ADABOOST for music classification and showed that aggregation improves the performance of ML algorithms. The work in [8] proposed a method for disease status classification that makes use performed taxonomic abstraction for a reduced feature space to improve classification accuracy and model interoperability. Further, to improve the performance of their model, the work in [11] proposed general regression neural networks (GRNNs) for detecting CRC where they used most predictive microbial species which were filtered by a nonlinear feature selection method. The combination GRNNs with feature selection improved detection accuracy and interpretability of their proposed model.

Some recent methods of have also used the concept of noisy replicates in order to improve classification of microbial based samples. Earlier, Knights, Costello and Knight [12] in their study claimed that replication of training data by adding noise can improve predictive power of ML models. In a very different study in [13] addressed the issue multiple local minima in hyperspace to improve the performance of their model. The authors suggested a technique for combining NNs in order to produce an ensemble that has better accuracy and error tolerance than a usual network.

Existing methods show that there is room for further improvement of DNN models CRC preditive model using microbiome data. This paper a method that applies concepts of noisy replicates, feature expansion, feature expansion and extended search for global minimum in order to increase the performance of a DNN based method for CRC classification.

Materials and Methods

The Section describe data transformations applied to improve the performance of the DNN to classify CRC, namely, dataset extension, column replication and dimensionality reduction. The dataset used in this work is also described.

A. Dataset

The proposed method used a filtered version of curated metagenomic which has a total of 796 samples and can be broken down into 368 CRC samples and 428 controls. Also, the filtered dataset consists of 2033 features of which 2031 contained microbial counts and the other two features contained demographic data namely age and biomass index (BMI). The dataset is a subset of original curated metagenomic data which has been documented by McMurdie and Holmes [14].

B. Feature Manipulation Methods

Several feature manipulations methods are combined in this work, name feature expansion, feature combination and feature extension.

Creation of noisy replicates can be used to improve classification of microiome data [12]. Therefore, the proposed method created noisy replicates, for the same reason, by using simple arithmetic operations to create 8 new additional columns in the dataset which increased the number of features from 2033 to 2041. This method is being referred to as feature expansion. The following arithmetic operations were used to create each of the new columns in the context of feature expansion: the mean, maximum, minimum and standard deviation values from each row was computed to make column number 2034, 2035, 2036 and 2037 respectively; then the average between column 2034 and 2035, 2035 and 2036 and column 2036 and 2037 was computed to make column 2038, 2039 and 2040 respectively; column 2041 was computed as the average of column 2034, 2035 and 2036. Therefore, given that the input dataset has the form shown in (1).

$$A = \begin{bmatrix} X_{0,0} & \cdots & X_{0,m} \\ \vdots & \ddots & \vdots \\ X_{n,0} & \cdots & X_{n,m} \end{bmatrix} \qquad \textit{(1),}$$

a transformation function on matrix A, f(A), will produce a matrix of the form shown in (2).

$$B \\ = \begin{bmatrix} [X_{0,0,0}, X_{0,0,1}, X_{0,0,2}, X_{0,0,3}, X_{0,0,4}, X_{0,0,5}, X_{0,0,6}] & \cdots & [X_{0,m,0}, X_{0,m,1}, X_{0,m,2}, X_{0,m,3}, X_{0,m,4}, X_{0,m,5}, X_{0,m,6}] \\ \vdots & \ddots & \vdots \\ [X_{n,0,0}, X_{n,0,1}, X_{n,0,2}, X_{n,0,3}, X_{n,0,4}, X_{n,0,5}, X_{n,0,6}] & \cdots & [X_{n,m,0}, X_{n,m,1}, X_{n,m,2}, X_{n,m,3}, X_{n,m,4}, X_{n,m,5}, X_{n,m,6}] \end{bmatrix}$$

(2),

Where X represents a corresponding item in in matrix A, s represents the standard deviation of matrix A, max represents the maximum value in matrix A, min represents the minimum value in matrix A, and mean is the arithmetic mean of matrix A, the steps for computing corresponding multiple values in matrix B for every single value in x in matrix A are shown from (3) to (9).

$$x_0 = x \qquad \textit{(3)}$$

$$x_1 = \sqrt{sx_0} \qquad \textit{(4)}$$

$$x_2 = \frac{|x_0 - x_1|}{2} \qquad \textit{(5)}$$

$$x_3 = \frac{|max - x_0|}{2} \qquad \textit{(6)}$$

$$x_4 = \frac{x_0 + mean}{2} \qquad \textit{(7)}$$

$$x_5 = \frac{x_0 + min}{2} \qquad \textit{(8)}$$

$$x_6 = \sqrt{s + x_0} \qquad \textit{(9)}$$

An attempt to dimensionality reduction to reduce the feature space of a dataset was also made by adding the original feature values and the corresponding values of newly generated features in the feature expansion method. This method is being referred to as the combined method. The combined method produces a new dataset using the expression represented by parameter y in (10).

$$y = x_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 \qquad \textit{(10),}$$

Finally, the two datasets generated by the feature extension and feature combination methods are merged horizontally to produce a new dataset which has more features while maintaining the number of records. The new dataset is being referred to as the extended dataset.

C. Deep Neural Network

The proposed method used a four layered DNN which has 2033 nodes in the input layer and 90 nodes in each hidden layer. The root mean square propagation (RMSprop) was used as the optimiser and set to a learning rate of 0.0008 after fine-tuning. The Rectified Linear Unit (ReLU) activation function was used in the connected layers and the sigmoid activation function in the output layer.

D. Iterrative Execution of the DNN Model

In situations where there are multiple local minima, ML algorithms tend to get stack in a local minimum while searching for the global minimum, which reduces classification performance of the algorithms. In order to avoid such a situation, the proposed method uses a solution called chaining. The proposed method simply invokes multiple DNN instances one after another on the hyperplane in order to search for the best solution.

In order to keep the method simple, predictions are not passed between DNN instances, but the algorithm selects the best prediction performance from the DNN instances. Adding logic for selecting the best performance also helps to address the stochastic nature of ML algorithms.

E. Model Validation

The proposed method used k-fold cross validation to evaluate the model. The dataset was split into 10 folds and the model was trained on 9 parts while validation was done on 1 part. The process was repeated 10 times in order to have each fold used to train as well as validate the model. However, in order to prevent overfitting a fold was not used to train and validate the model in the same iteration. We will refer to this technique as looping.

RESULTS

This section discusses the experimental results of the method proposed in the paper. Specifically, the results based on variations applied in the during the experiment, namely, feature expansion, feature combination, feature extension and the looping are analysed. Specifically, the analysis was conducted by comparing how each of the feature manipulation methods and the looping technique cumulatively contributed to the overall performance of the model. Evaluating the model on the original, expanded, combined and extended datasets was meant to give insight on how each aspect of the proposed feature manipulation method contributed to the overall performance of the model. Also, it is worth noting that while original, expanded and combined datasets refer to different version of the same dataset, extended dataset refers to a version of the combined dataset comprising the expanded and combined versions. Furthermore, looping was performed on the extended dataset. The performance of proposed model using receiver operating characteristics area under the curve (ROC AUC) metric is shown inf Fig. 3.

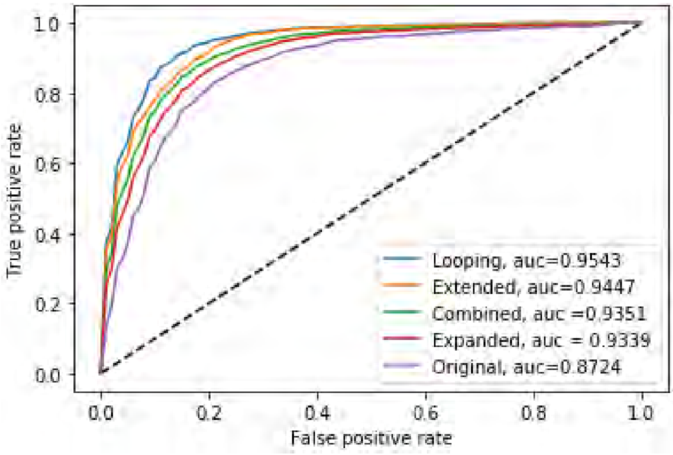

Fig. 3. Receiver Operating Characteristics Area Under the Curve (ROC AUC) for the Proposed Method

Fig. 3 shows while the expanded feature engineering technique significantly increases the performance of the model by 5.15%, combining features, extending the dataset and the looping technique when applied together also able to raise the performance of the model significantly by approximately 2%. Also, confusion matrix was used to analysis how each of the techniques affected the ability of the proposed model to detect both positive and negative cases as shown in Fig 4.

Fig. 4 also shows that while a dataset with expanded features significantly outperforms the original dataset, combined, extended and looping techniques when applied cumulatively to the model, are able to improve the ability of the model to identify both positive and negative cases.

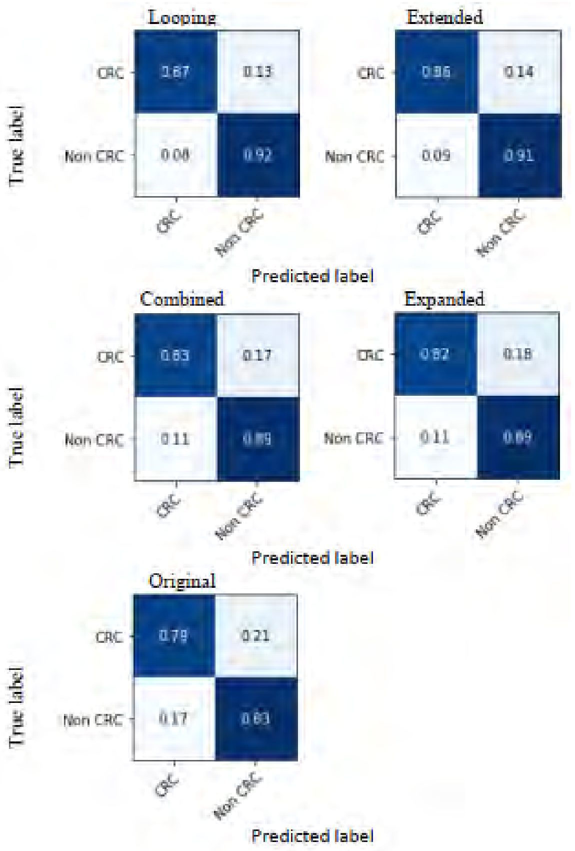

Fig. 4. Analysing the Performance of Feature Manipulation methods and the Looping Technique

DISCUSSION

In this paper we have used a method that applies concepts of noisy replicates, feature expansion, feature expansion and extended search for global minimum in order to increase the performance of a DNN based method for CRC classification.

The results show that while all constituent methods and technique on which overall method is based, positively contributed the improved classification, noisy replication technique of expanding individual features into multiple entities had the highest contribution of approximately 6.15%. Also, feature reduction by recombing features using simple summation had the least contribution of approximately 0.12%. The other techniques namely, extension of the dataset, and looping, contributed 0.96% and 0.96% respectively. Therefore, while mean AUC performance of the model is about 95.43%, the proposed method improved classification performance by approximately 8.77%.

Although feature reduction did not produce very remarkable results in the proposed method, it is a reliable technique for improving both computational and classification performance [9]. Despite research showing that a concatenation of features into a single feature vector is a straight forward method for feature combination [15], the simplicity of the technique could be responsible for the minor performance increment.

Furthermore, aggregate features [10] contributed a higher performance increment to the model than feature reduction. Although the technique had a non-trivial performance contribution, the parameters used to create the aggregate columns were randomly selected. Also, in addition to the generally accepted concept that feature aggregation is a useful strategy for improving classification performance [10], there are other two reasons which prompted the use of aggregate features in the proposed method. Firstly, the technique is similar to feature reduction although the new features are mixed with the original features. Secondly, adding aggregate features to the rest of the dataset can be considered as mechanism for creating noisy replicates in the dataset [16]. The technique has mainly been used in the context of data augmentation [12] [17] which is a proven method for reducing overfitting in datasets that are characterised by high dimensionality [4].

Similarly, the feature expansion technique used in the proposed method is based on the concept of noisy replicates. The technique is explored in [18] and has a higher performance because it adjusts the underlying data distribution. Furthermore, the technique of adding iterations (looping) to the proposed method was meant to increase classification performance by extended the search for the global minimum. The concept is called chaining which is a technique that is used to search for the global minimum in situations where there are multiple local minima [13][19]. Multiple local minima can prevent ML methods from reaching the global minimum which reduces classification performance [13]. Looping over DNN instances also produced a non-trivial performance increment in the model.

Conclusion

In this paper we have been able to increase CRC classification performance of a DNN based method by combining feature manipulation techniques, identification of an appropriate DNN architecture and by iteratively executing several instances of a DNN model. The feature manipulation combined methods that create noisy replicates and feature expansion, feature combination both of which have been used to improve ML classification tasks. Also, iterative execution of the DNN algorithm helped to increase the ability of the model to detect CRC cases by extending the search for the global minimum in the presence of multiple local minima. Although individual contributions of the techniques varied in significance, collectively they produced a model that has a significantly high CRC classification performance. Our investigation has shown that there is great potential in feature manipulation for improved classification of CRC based on microbiome data. The limitations of this work are that it is based on a single dataset and only focuses on binary classification. This may affect the ability of the model to generalise and underperform in multiclassification tasks, respectively. Future methods should consider using several datasets, including those related to multiclassification tasks. It would also be interesting to investigate how to improve on feature reduction, aggregation and their combinations with other techniques such as ensemble DNNs.

7.0 References

[1] J. Yu et al., “Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer,” Gut, vol. 66, no. 1, pp. 70–78, 2017, doi: 10.1136/gutjnl-2015-309800.

[2] J. Ferlay et al., “Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012,” Int. J. Cancer, vol. 136, no. 5, pp. E359–E386, 2015, doi: 10.1002/ijc.29210.

[3] J. P. Zackular, M. A. M. Rogers, M. T. Ruffin, and P. D. Schloss, “The human gut microbiome as a screening tool for colorectal cancer,” Cancer Prev. Res., vol. 7, no. 11, pp. 1112–1121, 2014, doi: 10.1158/1940-6207.CAPR-14-0129.

[4] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, “Automatic colon polyp detection using region based deep CNN and post learning approaches,” IEEE Access, vol. 6, pp. 40950–40962, 2018, doi: 10.1109/ACCESS.2018.2856402.

[5] S. Jahani-Sherafat, M. Alebouyeh, S. Moghim, H. A. Amoli, and H. Ghasemian-Safaei, “Role of gut microbiota in the pathogenesis of colorectal cancer; a review article,” Hepatol Bed Bench, vol. 11, no. 2, pp. 101–109, 2018.

[6] J. C. Wooley and Y. Ye, “Metagenomics: Facts and artifacts, and computational challenges,” J. Comput. Sci. Technol., vol. 25, no. 1, pp. 71–81, 2010, doi: 10.1007/s11390-010-9306-4.

[7] A. M. Thomas et al., “Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation,” Nat. Med., vol. 25, no. 4, pp. 667–678, 2019, doi: 10.1038/s41591-019-0405-7.

[8] M. Oudah and A. Henschel, “Taxonomy-aware feature engineering for microbiome classification,” BMC Bioinformatics, vol. 19, no. 1. 2018. doi: 10.1186/s12859-018-2205-3.

[9] S. Khalid, T. Khalil, and S. Nasreen, “c,” in A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning, 2014, pp. 372–378. doi: 10.1109/SAI.2014.6918213.

[10] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, “Aggregate features and ADABOOST for music classification,” Mach. Learn., vol. 65, no. 2–3, pp. 473–484, 2006, doi: 10.1007/s10994-006-9019-7.

[11] A. Arabameri, D. Asemani, and P. Teymourpour, “Detection of Colorectal Carcinoma Based on Microbiota Analysis using Generalised Regression Neural Networks and Nonlinear Feature Selection,” IEEE/ACM Trans. Comput. Biol. Bioinforma., vol. PP, no. c, p. 1, 2018, doi: 10.1109/TCBB.2018.2870124.

[12] D. Knights, E. K. Costello, and R. Knight, “Supervised classification of human microbiota,” FEMS Microbiol. Rev., vol. 35, no. 2, pp. 343–359, 2011, doi: 10.1111/j.1574-6976.2010.00251.x.

[13] L. K. Hansen and P. Salamon, “Neural network ensembles.,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. October, p. 993— 1001, 1990, doi: 10.1109/34.58871.

[14] A. P. J. Mcmurdie, S. Holmes, G. Jordan, and S. Chamberlain, “Package ‘ phyloseq ,’” 2019.

[15] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, “A survey of audio-based music classification and annotation,” IEEE Trans. Multimed., vol. 13, no. 2, pp. 303–319, 2011, doi: 10.1109/TMM.2010.2098858.

[16] S. S. Lee, “Noisy replication in skewed binary classification,” Comput. Stat. Data Anal., vol. 34, no. 2, pp. 165–191, 2000, doi: 10.1016/S0167-9473(99)00095-X.

[17] C. Lo and R. Marculescu, “MetaNN: Accurate Classification of Host Phenotypes from Metagenomic Data Using Neural Networks,” ACM-BCB 2018 - Proc. 2018 ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics, vol. 20, no. Suppl 12, pp. 608–609, 2018, doi: 10.1145/3233547.3233696.

[18] M. Mulenga et al., “Feature Extension of Gut Microbiome Data for Deep Neural Network Based Colorectal Cancer Classification,” IEEE Access, vol. 9, pp. 1–14, 2021, doi: 10.1109/ACCESS.2021.3050838.

[19] K. Zaamout and J. Z. Zhang, “Improving Neural Networks Classificationthrough Chaining,” in ICANN’12 Proceedings of the 22nd international conference on Artificial Neural Networks and Machine Learning, 2012, pp. 288–295. doi: 10.1007/978-3-642-33266-1_36.