ADScAI
Applied Data Science & Artificial Intelligence Symposium

A Deep Learning Approach for Host Depletion in Metagenomic Samples

1st Mendis D.V.N.
Dept. of Computer Science & Engineering
University of Moratuwa
Moratuwa, Sri Lanka
venukshi.20@cse.mrt.ac.lk

2nd Ratnayake R.M.P.G.C.K.
Dept. of Computer Science & Engineering
University of Moratuwa
Moratuwa, Sri Lanka
chadmi.20@cse.mrt.ac.lk

3rd Rathnasiri W.A.T.N.
Dept. of Computer Science & Engineering
University of Moratuwa
Moratuwa, Sri Lanka
tishani.20@cse.mrt.ac.lk

***Keywords*—Metagenomics, Metagenomic classification, Microbiome, Host depletion, Deep learning**

I. Introduction

Metagenomic studies often struggle with excessive host DNA, which reduces the sensitivity and accuracy of microorganism detection. Traditional lab-based host depletion is costly and time-consuming, while computational methods using reference databases are resource-intensive and often less accurate. To overcome these limitations, there is a growing need for efficient, accurate, and resource-friendly host depletion techniques. Machine learning (ML) offers a promising alternative by enabling read classification without relying on large reference databases, reducing computational load and improving speed and reliability. Such approaches can greatly enhance the effectiveness of metagenomic analyses across diverse host species.

II. Literature Review

Effective host depletion is a critical step in metagenomic workflows, enabling more accurate and efficient microbial identification and characterization. Physical methods of host DNA depletion are commonly employed in metagenomic studies to reduce the proportion of host DNA and enrich microbial DNA. But these methods are costly, lack commercial availability, require fresh samples for each experiment, and are influenced by sample characteristics, making them less efficient and practical. Computational host depletion methods have become a preferred alternative to physical methods, particularly due to their ability to be applied after the sequencing step, directly targeting the sequencing reads rather than manipulating the physical sample. Also these approaches offer key advantages over physical methods, notably in terms of cost efficiency, and avoidance of sample degradation.

HoCoRT [1] is a host contamination removal tool providing a one-step genome indexing method. It removes human short reads accurately, but struggles with long reads, achieving only 59% sensitivity in the best case, emphasizing the necessity of a more robust tool specifically built for precise long read host DNA depletion. As the alignment-based tools are highly storage intensive and they lack the ability of identifying novel species, ML approaches are emerging. AMAISE [2] is using a CNN model to classify the host and microbial long reads in metagenomic samples. Though it supports host depletion of long reads, it is highly time consuming. So to ensure a quick response and avoid overwhelming the computational resources available in most research laboratories, processor and memory requirements must be kept to a minimum.

III. Materials and Methods

A. Datasets

Training data for our tool were composed of 250,000 sequences of length 10,000bp. 50% of those sequences were sampled from reference genomes (50% from human and the other 50% were equally split among 2828 bacterial, 194 archaeal, 5455 viral, 34 protozoan and 57 fungal species) and the other 50% from Nanopore sequencing technology (50% from a human and the other 50% were equally split among 80 bacterial and 8 fungal species). In the reference genome dataset, repeat regions were masked using DustMasker [3], and reads were selected from random locations. Validation data followed the same structure. Reference genomes were obtained from NCBI GenBank, and Nanopore data from NCBI SRA.

For testing, we used 5 Nanopore datasets (2.4M reads each) with varying host fractions (1%, 25%, 50%, 75%, 99%). Microbial sequences in these test sets were sourced from 102 bacterial, 123 viral, and 41 fungal species, with sequence lengths ranging from 1,000 to 1.4M bp. Additionally, we tested our tool on a separate dataset of simulated PacBio reads (human test new) generated using SimLoRD [4]. The dataset comprised 1,682,400 reads equally split between human and 5 microbial classes. Furthermore, we evaluated our tool using a real metagenomic test set containing 155,641 sequences longer than 1,000 bp from the human oral metagenomic sample (SRR11547004) publicly available at the NCBI SRA. The best

https://doi.org/10.31705/ADScAI.2025.44

hits from BLAST [5] were used as the ground truth labels for the real dataset.

B. Data Preprocessing

DNA sequences were encoded into k-mer frequency vectors using Seq2Vec [6], normalizing k-mer counts for length-invariant representation. We tested various k-mer sizes (3-mer, 4-mer, 5-mer, and a 2-mer + 3-mer combination) and selected the 2-mer + 3-mer configuration as the best based on training time.

C. Model Architecture

The proposed tool is based on a Feed Forward Neural Network (FFNN) that takes a k-mer feature vector and outputs a classification label. The architecture consists of an input layer of 42 features, followed by three hidden layers (1024, 256, and 128 neurons) with ReLU activations, and a final output layer with sigmoid activation function for binary classification. The model was trained using the Adam optimizer with an initial learning rate of 1e-3 and a weight decay of 1e-5. BCELoss was used as the loss function. Training was conducted over 30 epochs, and the model that achieved the highest validation accuracy was saved. The training process was carried out on a server equipped with a 32-core Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz, 120 GB of RAM, and one NVIDIA T4 GPU using PyTorch.

IV. Results and Discussion

We evaluated the performance of the proposed tool for both PacBio and Nanopore sequence classification, benchmarking it against existing tools including Kraken2 [7], a state-of-the-art taxonomic classification tool, and AMAISE [2].

When tested on the human_test_new dataset containing only PacBio reads, the proposed tooll outperformed both Kraken2 and AMAISE as shown in Table I.

For Nanopore sequence classification, we tested all three tools on five test sets with different host percentages. For human percentages below 75%, the proposed tool outperformed kraken and AMAISE, achieving a consistent overall accuracy of 95%, while the highest accuracy AMAISE and kraken2 achieved were 89% and 93%, respectively. For host percentages from 75% and above, the proposed tool showed a slightly lower accuracy of 96%, while both AMAISE and Kraken2 reached higher accuracy as the proportion of human sequences increased. However, the proposed tool consistently demonstrated higher microbial recall (93%-96%) in all human host percentages, while Kraken2 and AMAISE showed a recall below 90% for most human percentages. This indicates the effectiveness of the proposed tool in microbial classification,which is crucial for the accuracy and reliability of subsequent steps in the metagenomic analysis pipeline, including assembly and downstream analysis. In terms of peak memory usage, the proposed tool required only 3GB of RAM for each classification, compared to 87GB for Kraken2 due to its extensive reference database and 5GB for AMAISE.

TABLE I
COMPARISON OF CLASSIFICATION PERFORMANCE OF HUMAN_TEST_NEW DATASET

Evaluation Metrics	Kraken2	AMAISE	Proposed Tool
Overall Accuracy (%)	69	95	98
F1 Score - Host	0.74	0.87	0.94
F1 Score - Microbial	0.78	0.97	0.99

Additionally, it classified each test dataset in 3 minutes, while Kraken2 took 15 minutes and AMAISE required 4 hours.

On the real Metagenomic dataset with Nanopore sequences, the proposed tool achieved an overall accuracy of 87%, while Kraken2 and AMAISE achieved slightly higher accuracy of 92%. However, the proposed tool outperformed them in microbial recall and host precision (both 99%), indicating its ability in depleting host sequences while preserving microbial sequences, ensuring accurate downstream analyses.

To provide a fair comparison of resource utilization, we trained AMAISE and the proposed FFNN model on the same dataset. AMAISE required 44 hours and 5.92GB RAM, while the FFNN trained in just 12 minutes using only 0.93GB RAM. This highlights the efficiency of feedforward networks with k-mer encoding over CNNs with one-hot encoding, which are more resource-intensive.

V. Conclusion

We presented a FFNN-based approach for host depletion using normalized k-mer frequency vectors for sequence encoding. This approach provides a more resource-efficient and generalizable alternative to existing host depletion methods. The tool was benchmarked against Kraken2 and AMAISE for PacBio and Nanopore sequences and evaluated on a real metagenomic dataset. Results demonstrated competitive performance while significantly reducing memory and computation time. Currently trained for human host depletion, the model can be extended to support retraining for other host types, making it adaptable to diverse metagenomic samples even in resource-constrained environments. Furthermore, there is a need to evaluate the application of advanced machine learning techniques, including transfer learning and LSTM, for further enhancing host depletion capabilities.

References

[1] I. Rumbavicius, "Tool to remove specific organisms from microbiome sequencing data host contamination removal tool (hocort)," Master's thesis, 2022.

[2] M. Krishnamoorthy, P. Ranjan, J. R. Erb-Downward, R. P. Dickson, and J. Wiens, "Amaise: a machine learning approach to index-free sequence enrichment," *Communications Biology*, vol. 5, no. 1, p. 568, 2022.

[3] A. Morgulis, E. M. Gertz, A. A. Schäffer, and R. Agarwala, "A fast and symmetric dust implementation to mask low-complexity dna sequences," *Journal of computational biology*, vol. 13, no. 5, pp. 1028–1040, 2006.

[4] B. K. Stöcker, J. Köster, and S. Rahmann, "Simlord: simulation of long read data," *Bioinformatics*, vol. 32, no. 17, pp. 2704–2706, 2016.

[5] A. Sf, "Basic local alignment search tool," *J Mol Biol*, vol. 215, pp. 403–410, 1990.

[6] A. Wickramarachchi, V. Mallawaarachchi, V. Rajan, and Y. Lin, "Metabcc-lr: meta genomics b inning by c overage and c omposition for l ong r eads," *Bioinformatics*, vol. 36, no. Supplement_1, pp. i3–i11, 2020.

[7] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with kraken 2," *Genome biology*, vol. 20, pp. 1–13, 2019.