Viral Detection Tools Comparative Analysis

Huiming Zhou
huiming@uoguelph.ca
University of Guelph
Shanghai, China

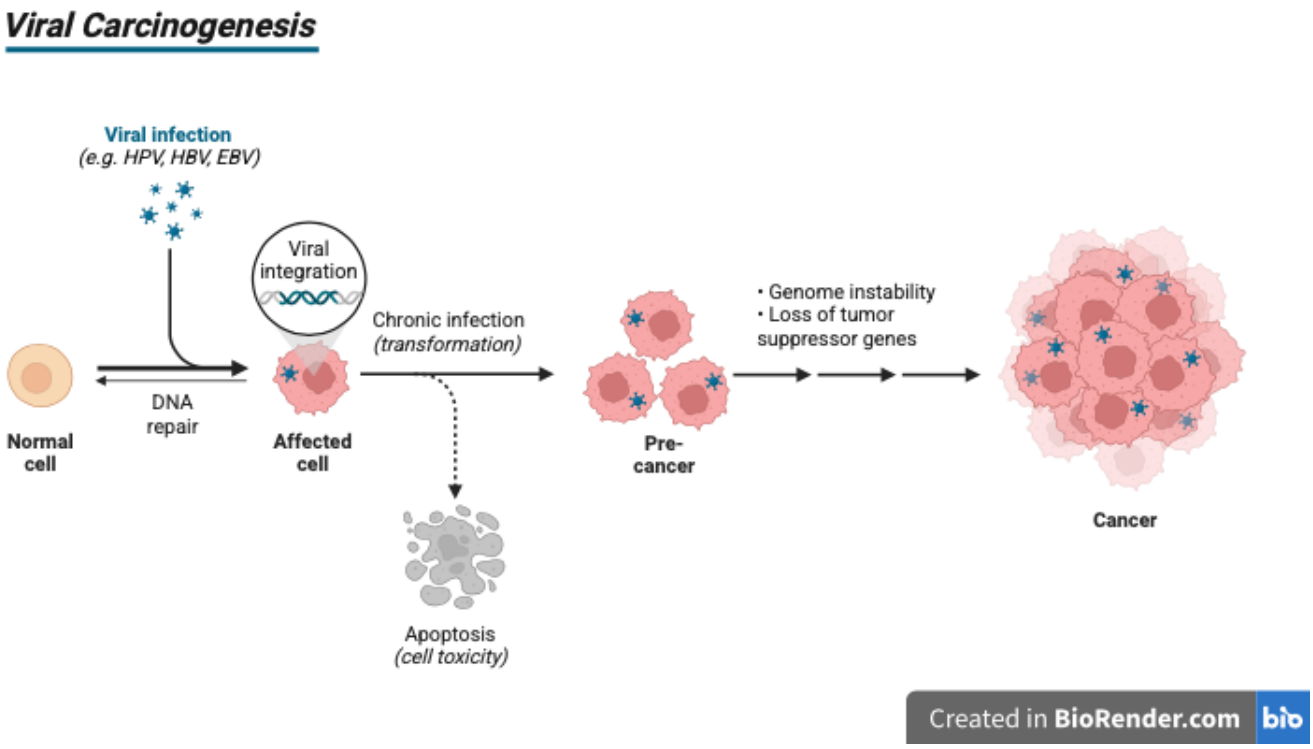

Figure 1: Virus-infected Cancer (Image source: [2])

Abstract

Virus detection is a crucial step for all downstream analysis work, especially in cancer. The development of machine learning-driven bioinformatics tools has significantly advanced this field. Thus, finding an efficient viral detection tool can serve as a helpful methodology in improving early cancer diagnosis. This study presents a comparative evaluation of three deep learning-based viral detection tools: DeepVirFinder, VirHunter, and viRNAtrap. The comparison is tested on a manually collected dataset of viral sequences from NCBI. The viRNAtrap, with a more advanced deep learning structure, achieves an accuracy of 65%, outperforming the other two models.

ACM Reference Format:
Huiming Zhou. 2025. Viral Detection Tools Comparative Analysis. In *Proceedings of Course Project Report (CS6060 Bioinformatics).* Bioinformatics, Guelph, ON, CAN, 3 pages.

CS6060 Bioinformatics , University of Guelph
2025.

1 Introduction

Nowadays, about 15% of human cancer cases are attributed to viral infections. Therefore, identifying viral sequences from metagenomic data is a crucial step for downstream analysis. Until now, most viral expression in tumor tissues has been examined by aligning tumor RNA sequences to known viral databases. However, current databases cover less than 5% of viral diversity, leaving over 90% of viruses uncharacterized. Moreover, existing viral genome databases are markedly biased toward viruses with cultivable hosts. Even when tumor tissue samples are available, host DNA sequences constitute 99% of the sample content, complicating viral detection and causing alignment-based tools to require excessive computational time.

Two main categories[9] of methods were initially used. The first is gene homology-based methods, which identify viral sequences by comparing query sequences against viral protein databases to detect evidence of viral genes (e.g., ViromeScan, DIAMOND, VirSorter2). The second is sequence alignment-based methods, which classify

viral metagenomic reads through alignment and homology searches against reference genomes (e.g., MetaVir, ViromeScan, Kraken). To address these limitations, there is a need for efficient tools capable of identifying known and novel viruses without relying on existing databases. VirFinder [7], an alignment-free k-mer-based machine learning tool, outperformed state-of-the-art gene-based classification methods and advanced metagenomic studies of viral ecology. Three years later, DeepVirFinder (the first model compared in this study) surpassed VirFinder's performance by employing neural network architectures. Subsequently, deep learning tools like VirHunter—which I reviewed in prior research and inspired this study—emerged. Unlike other methods, VirHunter excels at classifying viruses, bacteria, and hosts in plant systems. Finally, viRNAtrap, a state-of-the-art bioinformatics tool identified in Nature, was selected as the third model for comparison.

2 Methodology

As all three methods were developed in different years, it is challenging to compile them in a uniform environment. Fortunately, their source codes are provided with requirement files or `environment.yml`. Therefore, I created three separate virtual environments using Miniconda to accommodate their specific dependencies.

2.1 DeepVirFinder

DeepVirFinder [8] [6] is the oldest tool among the three but represents the first deep learning-based virus detection method, serving as a benchmark for subsequent tools. Built on a CNN framework, its architecture includes a convolutional layer, a max-pooling layer, and two dense layers. The model was trained on viral RefSeq data prior to May 2015 and outperforms the state-of-the-art method VirFinder, which is a logistic regression model with lasso regularization.

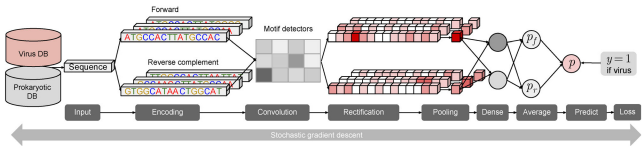

Figure 2: Architecture of DeepVirFinder's CNN framework demonstrating feature extraction layers

2.2 VirHunter

VirHunter [10] [3] is the first bioinformatics tool I studied during my literature review. It employs a neural network framework with a final random forest classifier. Originally trained on plant datasets, it accepts FASTA files as input and predicts probabilities for viral, host, or bacterial classifications. To evaluate its performance on human data, I modified the configuration and training files to enable binary classification. After retraining with DeepVirFinder's training FASTA files and using hyperparameters recommended by VirHunter's authors, the model achieved an accuracy exceeding 95%.

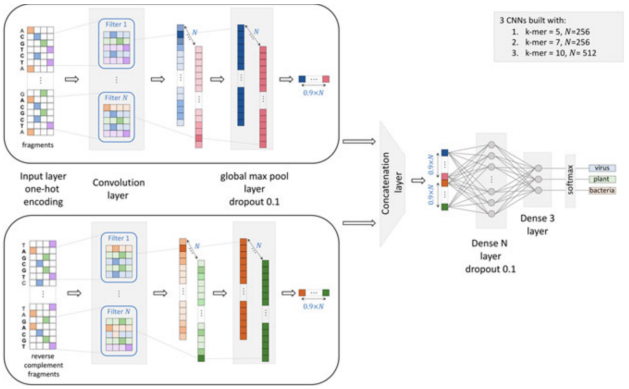

Figure 3: Hybrid architecture of VirHunter combining neural networks with random forest classifier

2.3 viRNAtrap

viRNAtrap [4] [1]a cutting-edge viral detection tool published in *Nature*, is the first model to segment viral data into 48bp reads. It outperforms other methods in all metrics except precision, where DeepVirFinder remains superior. However, precision is less critical for this framework because alignment steps are used to filter out false negatives. viRNAtrap consists primarily of two components: a deep learning model and a contig assembly module for identifying viral contigs.It is the only tool that needs FASTQ file instead of FASTA file.

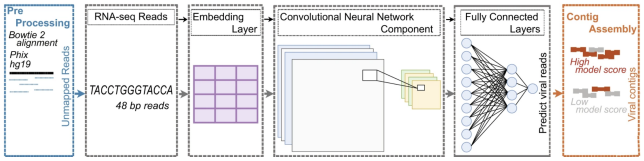

Figure 4: Dual-module architecture of viRNAtrap integrating deep learning with contig assembly

2.4 Validation Dataset

My initial dataset consideration was from the TCGA GDC portal, but due to access controls, this option was abandoned. An additional concern was that viRNAtrap had already been trained on TCGA data including 14 cancer types, potentially leading to data leakage. Consequently, I sourced newly published data from NCBI [5], obtaining six HBV files, six HCV files, and five normal liver samples as negative controls. The final compiled dataset contained 39 sequences. As viRNAtrap uniquely requires FASTQ-formatted input, I developed a script using DeepSeek's assistance to transform the data into compatible FASTQ files.

3 Research Result

After applying the three models to the dataset, I created confusion matrices and computed their accuracy. However, for viRNAtrap — whose output consists of viral contig assemblies rather than complete sequences — manual comparison and accuracy calculation

were required. viRNAtrap achieved 56% prediction accuracy. The confusion matrices are presented below.

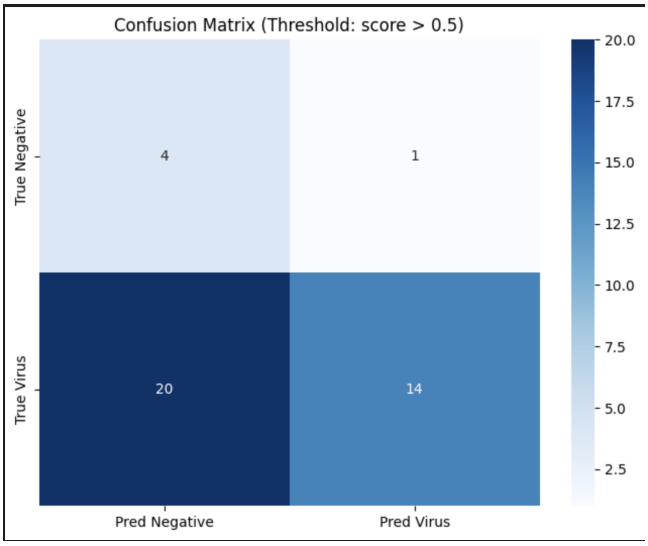

Figure 5: DeepVirFinder performance metrics: 46% overall accuracy with 93% sensitivity (true positive rate) for viral sequence detection

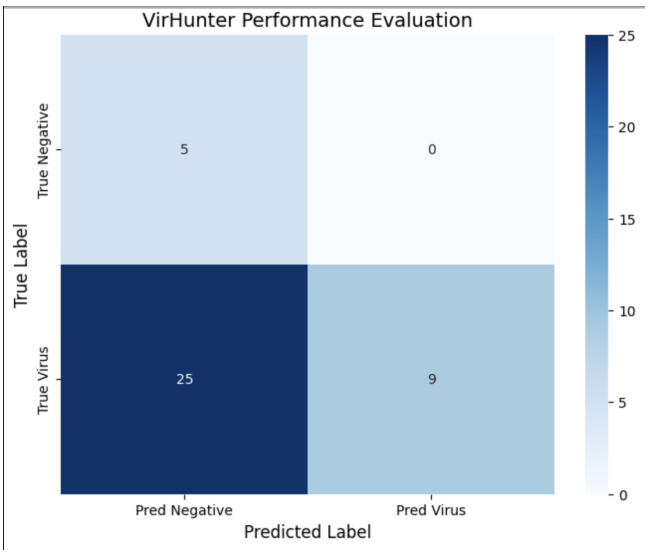

Figure 6: VirHunter performance metrics: 36% overall accuracy with 100% sensitivity (perfect viral sequence recall) at the cost of reduced specificity

4 Discussions and Conclusions

The comparative analysis demonstrates viRNAtrap's superior performance with the highest accuracy (56%) and perfect precision (100%), .So we can say viRNAtrap is still the current state-of-the-art virus detection tool.But there are mainly three limitations need to be solved: First, the dataset composition introduces bias — viral sequences significantly outnumber host sequences in our validation set which is not suitable for evaluation. Second, DeepVirFinder and VirHunter were evaluated on smaller datasets (pre-2015 RefSeq), whereas viRNAtrap benefited from more recent TCGA training data. Third, HBV and HCV are two of the most common cancer that can be easily found on NCBi, which might even be included in some of their training data. And there are also potential furture works for better improvement: Expand methodological comparisons to include structural prediction tools like AlphaFold or Transfomer-based tool, Retrain all models with equivalent training data, Acquire post-2023 viral genomes or synthetic novel viruses for balanced validation.

5 Citations and Bibliographies

References

[1] AuslanderLab. 2023. VirNaTrap: A Tool for Identifying Viral RNA Sequences. https://github.com/AuslanderLab/virnatrap Accessed: 31 March 2025.

[2] BioRender. 2023. Viral Carcinogenesis. https://app.biorender.com/biorender-templates/figures/all/t-5fff0e008e488100a5ebc250-viral-carcinogenesis Accessed: 31 March 2025.

[3] CBiB. 2022. VirHunter: A Tool for Identifying Viral Sequences. https://github.com/cbib/virhunter Accessed: 31 March 2025.

[4] Abdurrahman Elbasir, Ying Ye, Daniel E. Schäffer, Xue Hao, Jayamanna Wickramasinghe, Konstantinos Tsingas, Paul M. Lieberman, Qi Long, Quaid Morris, Rugang Zhang, Alejandro A. Schäffer, and Noam Auslander. 2023. A Deep Learning Approach Reveals Unexplored Landscape of Viral Expression in Cancer. *Nature Communications* 14 (February 2023), Article number: 785. doi:10.1038/s41467-023-36269-2

[5] National Center for Biotechnology Information. 2023. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/ Accessed: 31 March 2025.

[6] Jie Ren. 2019. DeepVirFinder: A Deep Learning-Based Tool for Identifying Viral Sequences. https://github.com/jessieren/DeepVirFinder Accessed: 31 March 2025.

[7] Jie Ren, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2017. VirFinder: A Novel k-mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data. *Microbiome* 5 (July 2017), Article number: 69. doi:10.1186/s40168-017-0283-5

[8] Jie Ren, Kai Song, Chao Deng, Nathan A. Ahlgren, Jed A. Fuhrman, Yi Li, Xiaohui Xie, Ryan Poplin, and Fengzhu Sun. 2020. Identifying Viruses from Metagenomic Data Using Deep Learning. *Quantitative Biology* 8, 1 (March 2020), 64–77. doi:10.1007/s40484-019-0187-4

[9] Ronit Sarid and Shou-Jiang Gao. 2010. Viruses and Human Cancer: From Detection to Causality. *Cancer Letters* 305, 2 (October 2010), 218–227. doi:10.1016/j.canlet.2010.09.011

[10] Grigorii Sukhorukov, Maryam Khalili, Olivier Gascuel, Thierry Candresse, Armelle Marais-Colombel, and Macha Nikolski. 2022. VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. *Frontiers in Bioinformatics* 2 (12 May 2022), 867111. doi:10.3389/fbinf.2022.867111