PathoTransformer: Advancing Pathogen Identification with Transformers

Khalid Saifullah
khalids@umd.edu

Jiuhai Chen
jchen169@umd.edu

Davit Soselia
dsoselia@umd.edu

1 Introduction

The rapid and accurate classification of pathogens from genomic sequences is a cornerstone of modern infectious disease control and outbreak management. Traditional methods, including sequence alignment and comparison against databases like GenBank, have been instrumental in identifying known pathogens. However, these methods can be computationally intensive and less effective in the face of novel or rapidly mutating viruses. With the advent of deep learning techniques, specifically transformer models, there is an opportunity to significantly enhance the accuracy and speed of pathogen classification by leveraging large-scale genomic data.

Our project aims to harness the power of transformer models to develop a more robust tool for classifying pathogens from genomic sequences. Building on previous work that relied on traditional machine learning models such as Long Short-Term Memory (LSTM) networks, we propose the PathoTransformer model. This model utilizes advances in attention mechanisms inherent in transformer architectures to improve pathogen classification. In our work we contribute by combining a comprehensive pathogen genomic sequence dataset with a transformer model trained to identify pathogens. Furthermore, we establish a pipeline for evaluating the results. Using the DeepPredictor dataset, supplemented with additional pathogen sequences from NCBI and ENA, we conducted robustness evaluations, employing metrics like the F1 score to measure performance against various form of sequence perturbations.

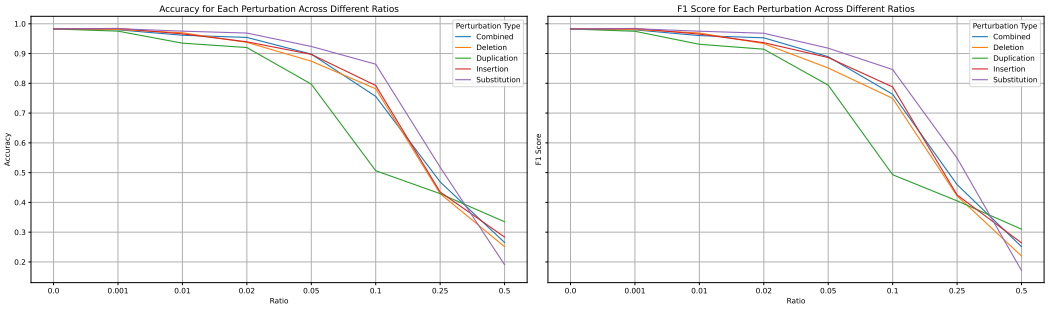

Figure 1: The figure shows the change in performance using average accuracy (left) and average F1 scores (right) as the ratio of perturbed nucleotides increases. The possible perturbations compared are Substitution, Deletion, Duplication, Insertion, and Combination, the latter of which randomly selects among all others. We observe gradual decline in both accuracy and F1 scores, with the highest drop for Duplication and the least for Substitution.

2 Related Work

The rapid spread of the SARS-CoV-2 virus, which causes COVID-19, has led to a global pandemic, underscoring the pressing need for early and accurate prediction of this and other pathogenic viruses from their genomic sequences. Traditional approaches for pathogen identification have relied on alignment-based techniques such as BLAST [1], which can be computationally expensive and may fail to detect similarities in highly divergent sequences.

In recent years, several studies have explored the use of machine learning and deep learning models for pathogen prediction from genomic data. Koohi-Moghadam et al. [7] employed a deep learning approach to predict disease-associated mutations in metal-binding sites of proteins, achieving an area under the curve (AUC) of 0.90 and an accuracy of 0.82. Özturk et al. [8] utilized deep neural networks with X-ray images for automated detection of SARS-CoV-2 cases, reporting an accuracy of 98.08% for binary classification (COVID vs. No-Findings) and 87.02% for multi-class classification (COVID vs. No-Findings vs. Pneumonia).

More recently, Saha et al. [10] proposed COVID-DeepPredictor, a Long Short-Term Memory (LSTM) based recurrent neural network for SARS-CoV-2 and other pathogenic virus prediction using an alignment-free k-mer approach. COVID-DeepPredictor achieved impressive accuracy ranging from 99.51% to 99.94% on various test datasets. However, the authors acknowledged the potential for further improvement by exploring alternative deep learning architectures.

Transformer models, particularly the Bidirectional Encoder Representations from Transformers (BERT) [5], have achieved state-of-the-art performance in various natural language processing tasks. While originally designed for text data, BERT has also shown promise in biological sequence analysis tasks such as protein structure prediction [9] and functional genomics [2]. Inspired by the success of BERT in capturing long-range dependencies and leveraging pre-trained representations, this work aims to explore the application of BERT for SARS-CoV-2 and other pathogenic virus prediction from genomic sequences.

Building on the foundation laid by BERT, recent explorations have extended transformer models to a variety of biomedical applications. De Silva and Brown et al. [4] demonstrated the efficacy of Vision Transformers (ViTs) combined with Convolutional Neural Networks (CNNs) for plant disease identification using multispectral imaging. Their study achieved high accuracy, precision, recall, and F1 scores, showcasing the potential of transformer models in the domain of precision agriculture and plant pathology. This hybrid approach underscores the adaptability of transformer architectures beyond their initial scope, suggesting a promising avenue for pathogen identification in complex biological datasets.

Furthermore, the integration of genomics and transformer-based models is gaining traction as a method for enhancing pathogen surveillance and identification. Vashisht et al. [12] reviewed the prospects and challenges of using genomic techniques for the detection and monitoring of emerging pathogens. They highlighted the role of high-throughput DNA sequencing tools in conjunction with advanced computational models to rapidly and precisely characterize pathogens. The review emphasizes the importance of continuous innovation in genomic analysis, which, when paired with transformer models, could revolutionize the field of pathogen identification and outbreak management.

3 Approach

We base our work primarily on COVID-DeepPredictor dataset. [10] COVID-DeepPredictor, is a compilation of genomic sequences of various pathogenic viruses, including SARS-CoV-1, MERS-CoV, Ebola, Dengue, Influenza, and SARS-CoV-2. The datasets for SARS-CoV-1, MERS-CoV, Ebola, Dengue, and Influenza were obtained from the National Center for Biotechnology Information (NCBI) [10]. Additionally, the dataset for SARS-CoV-2 was sourced from both NCBI and the Global Initiative on Sharing All Influenza Data (GISAID) [10].

The initial dataset comprised a total of 4,643 complete or near-complete genomic sequences of these pathogenic viruses. For training purposes, a subset of 1,500 samples was randomly selected from the initial dataset. To ensure representation from each pathogenic virus and mitigate class imbalance, 250 samples from each virus were included in the training dataset.

For testing, five distinct test datasets were created, denoted as Testdata-1 through Testdata-5. Testdata-1 consisted of the remaining 3,143 sequences from the initial dataset. Testdata-2 comprised 200 sequences each for MERS-CoV, SARS-CoV-2, Ebola, Dengue, and Influenza, along with 90 sequences of SARS-CoV-1 from diverse sources. Testdata-3, Testdata-4, and Testdata-5 contained recent SARS-CoV-2 sequences from NCBI and GISAID, along with sequences of other pathogenic viruses.

We also explore the genomic sequences for additional pathogens from the National Center for Biotechnology Information (NCBI) European Nucleotide Archive (ENA) a[3]. To achieve this we used NCBI genome explorer 2, to fit with the previous dataset we used the average number of sequences per class as a filter, meaning looking for pathogens with at least 750 sequences, resulting in selecting Monkeypox, African swine fever, Porcine circovirus, Rotavirus A from NCBI and adding to the existing dataset.

The raw nucleotide sequences are then processed via k-mers as suggested in [10], more specifically, they are split into overlapping substrings of length $K = 3$. This method transforms the sequences into a format that captures local sequence information effectively. For example, consider the nucleotide sequence "ATCGATCG". This sequence would be transformed into the following k-mers: "ATC", "TCG", "CGA", "GAT", "ATC", "TCG". These k-mers are then joined into a single string separated by spaces: "ATC TCG CGA GAT ATC TCG". This transformation is crucial for encoding the sequence information in a way that is compatible with downstream machine learning models, particularly those using transformers, which require tokenized input.

We utilize a BERT-based model, specifically DNA-BERT, to classify pathogens from genomic sequences. The architecture of DNA-BERT includes multiple transformer layers, each consisting of attention heads that allow the model to focus on different parts of the sequence simultaneously. This multi-headed attention mechanism is critical for understanding the complex relationships within genomic data.

Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation
ViralProj15142	GCA_000857045.1	GCF_000857045.1	Monkeypox virus	Zaire-96-I-16 (strain)	NCBI RefSeq Submitter
ASM2546247v1	GCA_025462475.1		Monkeypox virus		Submitter
ASM2579107v1	GCA_025791075.1		Monkeypox virus		Submitter
ASM2546246v1	GCA_025462465.1		Monkeypox virus		Submitter
ASM2546242v1	GCA_025462425.1		Monkeypox virus		Submitter
ASM2937617v1	GCA_029376175.1		Monkeypox virus		Submitter
ASM647192v1	GCA_006471925.1		Monkeypox virus		Submitter
ASM646594v1	GCA_006465945.1		Monkeypox virus		Submitter

Figure 2: We viewed different sequences available at NCBI explorer and viewed number of sequences available, downloading ones above threshold one at a time.

Following the approach in [10], the performance of the proposed BERT-based model is evaluated using accuracy, precision, recall, and G-mean. Accuracy measures the ratio of correctly predicted instances. Precision quantifies the fraction of true positives among predicted positives, while recall measures the fraction of true positives captured. The G-mean balances performance across classes by taking the geometric mean of sensitivity.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{G-mean} = \sqrt{\frac{\text{TP}}{(\text{TP} + \text{FP})(\text{TP} + \text{FN})}} \tag{4}$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Performance is compared against COVID-DeepPredictor.

We originally selected DNA-BERT [13], BERT-Plant-Genome [11], and Nucleotide-transformer-2.5b-multi-species [6] as possible base models to use for fine-tuning on our dataset, however due to compute limitations chose to focus on DNA-BERT.

To evaluate the robustness and reliability of the model we tested random perturbations of sequences. for the distribution of perturbations, we use random uniform distribution for a chance of perturbation with perturbation affecting ratio of chromosomes from 0.01 to 0.5, gaussian distribution with a random mean. The motivation for uniform distribution is to measure general robustness, while for the Gaussian to simulate a situation where a specific segment is incorrect. For perturbations, we used the following methods:

- **Deletion** - Remove a targeted ratio of nucleotides at randomly selected indices according to the specified distribution.
- **Insertion** - Add nucleotides at randomly selected indices.
- **Duplication** - Duplicate randomly selected nucleotides in the sequence.
- **Substitution** - Randomly substitute one or more nucleotides with different nucleotides. For example, change 'A' to 'T'.
- **Combination** - Apply a mix of the above operations with equal chances of each occurring.

Once applied to the test set we test the original model using the previously defined metrics and observe the change in the performance.

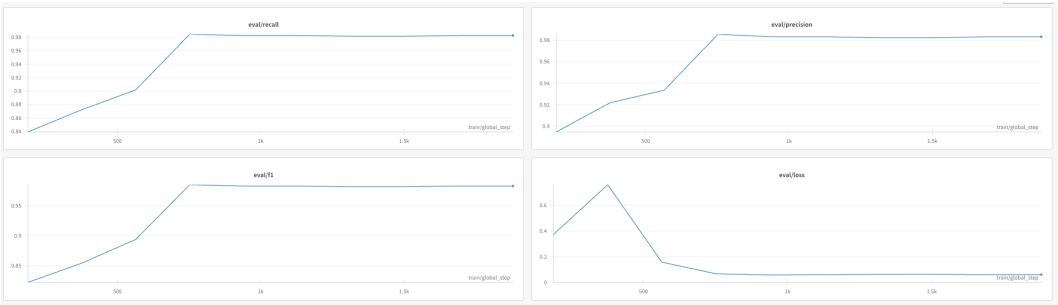

Figure 3: The figure shows the change in recall, precision, F1 score, and loss on the y-axis, evaluated on the validation set as the training progresses, denoted by the training steps on the x-axis.

4 Experiments

Our PathoTransformer model was trained using a combination of the COVID-DeepPredictor dataset and additional genomic sequences from NCBI and ENA. The model was evaluated using a set of metrics including accuracy, precision, recall, and F-1 score across multiple test datasets:

- **Accuracy:** The overall accuracy of PathoTransformer in classifying the genomic sequences of the viruses was significantly improved compared to traditional models, achieving an average accuracy of 98% from Figure 3 and 4.
- **Precision and Recall:** The model demonstrated a precision of 98% and a recall of 98%, indicating its effectiveness in identifying true positive cases without a substantial number of false positives from Figure 3 and 4.

The experiments address the question of whether Bert-based model architecture provides a meaningful increase in performance of recognizing genomic sequences of rapidly mutating viruses. As well as how robust a transformer-based model is to minor perturbations as discussed in the previous section. For the perturbations, we find that the drop is relatively gradual, with little loss at the ratio of 0.1% perturbations, and increasing loss above 10%. However, values beyond few percentages

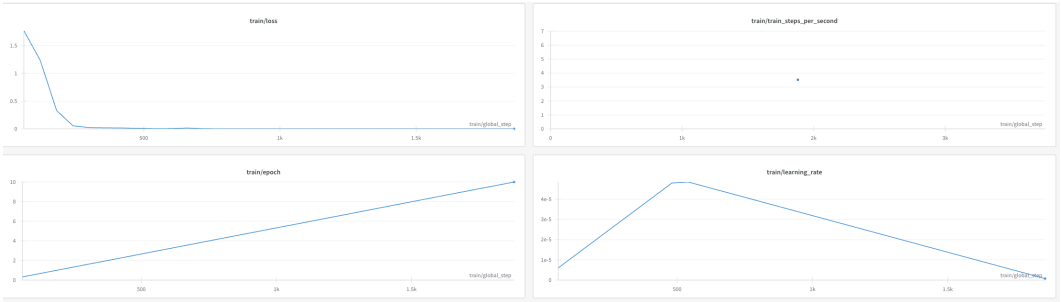

Figure 4: The figure shows the training process, change in the loss on the training dataset, learning rate, and training speed as the training progresses.

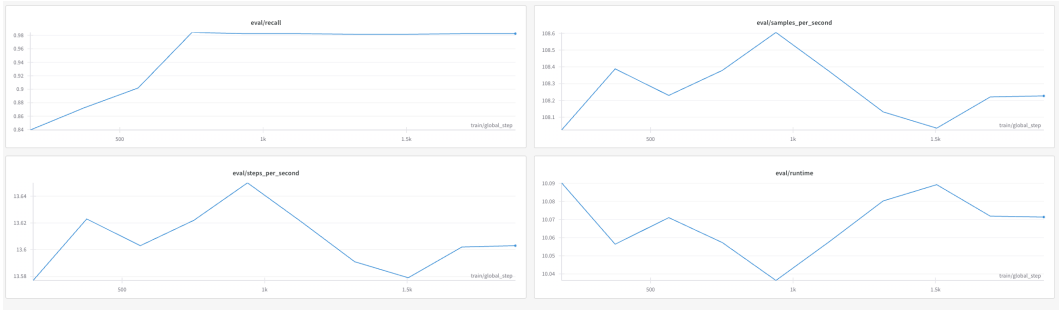

Figure 5: The figure shows the change in recall, samples processed per second, steps per second, and the runtime on the y-axis, evaluated on the validation set as the training progresses, denoted by the training steps on the x-axis.

are reasonably expected to cause degradation, because the sequence might actually not resemble the original at that point. We see that the Duplication, which chooses random indices and duplicates the nucleotides found there, causes the highest level of degradation, and on the opposite end is substitution which causes the least. However this trend only holds for ratios up to 0.25. The Combined perturbation seems to cause around the average of the others and could indicate that the model is not highly sensitive towards a mixture of perturbations.

5 Conclusion/Discussion

The primary takeaway from our perturbation experiments is the robustness of our model to minor sequence alterations and its sensitivity to more substantial changes. Our findings in figure 1 indicate that the model maintains high accuracy with up to 1% perturbation, but performance degrades significantly with higher perturbation levels, dropping to approximately 25% accuracy at 50% perturbation. Among the types of perturbations tested, the "duplication" operation had the most detrimental effect on model performance, suggesting that the model is particularly sensitive to changes that alter sequence length. Conversely, "substitution" perturbations had the least impact, which may indicate that the model relies heavily on positional encoding to memorize sequences, rather than understanding them in a meaningful way to be able to generalize better.

For future work, it is crucial to explore the underlying reasons behind the model's differing sensitivities to various perturbations. One potential direction is to integrate perturbation strategies during the training phase to enhance the model's robustness. Additionally, further investigation into the model's reliance on positional encodings versus actual sequence content could provide insights into improving model architecture for better generalization. The broader implications of our study are significant for pathogen identification and classification tasks, where robust models can enhance the detection and monitoring of rapidly mutating viruses. By improving the model's resilience to sequence alterations, our approach can be adapted for real-world applications in pathogen surveillance, outbreak management, and precision medicine.

References

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[2] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[3] Carla Cummins, Alisha Ahamed, Raheela Aslam, Josephine Burgin, Rajkumar Devraj, Ossama Edbali, Dipayan Gupta, Peter W Harrison, Muhammad Haseeb, Sam Holt, et al. The european nucleotide archive in 2021. *Nucleic acids research*, 50(D1):D106–D110, 2022.

[4] Malithi De Silva and Dane Brown. Multispectral plant disease detection with vision transformer–convolutional neural network hybrid approaches. *Sensors*, 23(20):8531, 2023.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] InstaDeepAI. Nucleotide-transformer-2.5b-multi-species. `https://huggingface.co/InstaDeepAI/nucleotide-transformer-2.5b-multi-species`, 2023.

[7] Mohamad Koohi-Moghadam, Haibo Wang, Yuchuan Wang, Xinming Yang, Hongyan Li, Junwen Wang, and Hongzhe Sun. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. *Nature Machine Intelligence*, 1(12):561–567, 2019.

[8] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine*, 121:103792, 2020.

[9] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[10] Indrajit Saha, Nimisha Ghosh, Debasree Maity, Arjit Seal, and Dariusz Plewczynski. Covid-deeppredictor: recurrent neural network to predict sars-cov-2 and other pathogenic viruses. *Frontiers in genetics*, 12:569120, 2021.

[11] Suke Sho. Bert-plant-genome-6. `https://huggingface.co/suke-sho/BERT-plant-genome-6`, 2023.

[12] Vishakha Vashisht, Ashutosh Vashisht, Ashis K Mondal, Jaspreet Farmaha, Ahmet Alptekin, Harmanpreet Singh, Pankaj Ahluwalia, Anaka Srinivas, and Ravindra Kolhe. Genomics for emerging pathogen identification and monitoring: Prospects and obstacles. *BioMedInformatics*, 3(4):1145–1177, 2023.

[13] Zhihan Zhang. Dnabert-2-117m. `https://huggingface.co/zhihan1996/DNABERT-2-117M`, 2023.