Ensemble and Transformer Models for Infectious Disease Prediction

Blessing Isoyiza ADEIKA
Department of Computer Science
Morgan State University
Baltimore, USA
https://orcid.org/0009-0007-5600-8051

Joseph Aina
Department of Computer Science
Morgan State University
Baltimore, USA
joain1@morgan.edu

Temileye Ibirinde
Department of Public Health
Morgan State University
Baltimore, USA
teibi2@morgan.edu

Tijesunimi Adeyemi
Department of Computer Science
Morgan State University
Baltimore, USA
adtij1@morgan.edu

Md mahmudur Rahman
Department of Computer Science
Morgan State University
Baltimore, USA
md.rahman@morgan.edu

Saroj Pramanik
Department of Biological Sciences
Morgan State University
Baltimore, USA
saroj.pramanik@morgan.edu

***Abstract*—Infectious diseases persist as an urgent global challenge, necessitating innovative strides in prediction and monitoring. This study delves into the intricate realm of infectious disease prediction, employing three transformer models—BERT, XLNET, and RoBERTa. The central objective of this research was to craft a framework for infectious disease prediction, significantly enhancing capabilities in disease monitoring, detection, and outbreak response. The approach entailed receiving a set of translated protein sequences from various infectious diseases and leveraging these sequences to predict each disease with the models. This methodology advanced infectious disease prediction and monitoring by expediting the analysis of genomic data, enabling the identification of distinctive patterns, mutations, and signatures associated with specific infectious agents. The dataset comprised genomic sequences from diseases such as Zika, Ebola, SARS-CoV-2, Influenza A, Influenza B, Tuberculosis, along with sequences from non-infected individuals. Model evaluation encompassed essential metrics, including accuracy, precision, recall, and the F1 score. In our quest for heightened precision, we also devised ensemble techniques to harness the collective power of all three models, yielding accuracies of 92% (Majority Voting) and 85% (Weighted Average). Leveraging DNA sequences translated into protein sequences, this study contributed to advancing our understanding and management of infectious diseases on a global scale.**

Keywords— Infectious diseases, BERT, XLNET, RoBERTa, Disease Prediction, Ensemble techniques, Genomic sequence

I. INTRODUCTION

Infectious diseases are a constant threat to humanity, and they have the potential to suddenly manifest or recur, as we have seen most recently with the COVID-19 pandemic, Ebola outbreaks, Zika epidemics, and ongoing issues with influenza [1]. It is essential to quickly and precisely identify the infectious agents that cause these diseases to protect the health and safety of the entire world's population. Conventional laboratory tests are still the primary component of the methods used to identify and study pathogens today. Even though these techniques work, they have several drawbacks, such as high costs, protracted processing times, a lack of diagnostic kits, and a lack of biosafety facilities [2]. Conventional methods have limitations [2][3], prompting the integration of innovative technologies. Deep learning, a subset of machine learning [15], and Transformer models, originally designed for natural language processing [7], offer transformative capabilities.

The need for an innovative framework to predict infectious agents that make use of genomic sequencing is the driving force behind this research project. Deep learning autonomously identifies complex patterns in data, aiding infectious disease prediction [15]. Transformer models, known for their attention mechanisms, excel in pattern recognition [7]. Together, they enable rapid pathogen identification and outbreak dynamics prediction, revolutionizing public health preparedness [4]. Without the need for prior knowledge of the target organism or sequence, we could use technologies capable of quickly and thoroughly analyzing DNA molecules within any biological sample [4].

The combined presence of this technological frontier creates opportunities for our objectives:

- To make it easier to find and study pathogens in-depth, to find patterns of transmission and risk factors, and to empower proactive public health interventions.
- To improve our ability to respond to threats of infectious diseases quickly and effectively by making it possible to predict the dynamics and outcomes of outbreaks.
- To develop and improve a transformer-based Genomic sequencing transformer enabled models that can quickly find and analyze infectious agents in samples taken from sick people.

II. LITERATURE REVIEW

In the field of infectious disease prediction and monitoring, several pioneering studies have harnessed the power of deep learning and transformer-based models, significantly contributing to our understanding and response to infectious

2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE) | 979-8-3503-9311-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBE60311.2023.00068

2471-7819/23/$31.00 ©2023 IEEE
DOI 10.1109/BIBE60311.2023.00068

diseases. Deep learning, exemplified by Cho et al.'s use of CNNs for disease classification from medical images [5], and Li et al.'s application of RNNs for epidemic forecasting [6], has significantly advanced disease prediction. Transformer-based models, like Vaswani et al.'s "Transformer" and Devlin et al.'s "BERT," have revolutionized healthcare applications, including infectious disease monitoring, by effectively capturing complex patterns in clinical data [7], [8]. Infectious disease surveillance has benefited from deep learning techniques, as seen in studies such as "EpiDeep" by Ren et al. and Long et al.'s real-time surveillance framework, aiding in understanding disease spread dynamics and early detection [9], [10]. The COVID-19 pandemic spurred research in deep learning for infectious diseases, with Wu et al. utilizing deep neural networks for COVID-19 diagnosis from chest X-ray images [11], and Yan et al. introducing "COVID-Net" for pandemic response [12]. Genomic sequencing, a pivotal tool, as demonstrated by Gardy et al. and Hadfield et al., has enabled tracing the transmission of infectious agents and tracking the spread of diseases like SARS-CoV-2 [13], [14]. These studies collectively underscore the transformative potential of deep learning, transformer-based models, and genomics in enhancing global public health through improved infectious disease prediction and monitoring.

III. METHODOLOGY

The research database was meticulously curated, drawing from an extensive collection of genomic and biomedical data sourced from the National Center for Biotechnology Information (NCBI). This comprehensive dataset encompassed a wide array of genomic sequences spanning across seven distinct disease categories, which included Ebola, Influenza A and B, Normal genome, Tuberculosis, Zika, and SARS-CoV-2. Within this dataset, a total of 26 samples were selected, each containing a wealth of detailed information pertaining to its respective disease category. This careful curation resulted in the creation of a rich and diverse database that served as the foundation for subsequent research endeavors.

The significance of data preprocessing cannot be overstated when striving to attain high-quality and precise predictions. Therefore, considerable emphasis was placed on the pre-processing of the dataset to eliminate any sources of noise and outliers. Notably, a pivotal step in this process involved the transformation of raw genomic sequences into their corresponding protein sequences.

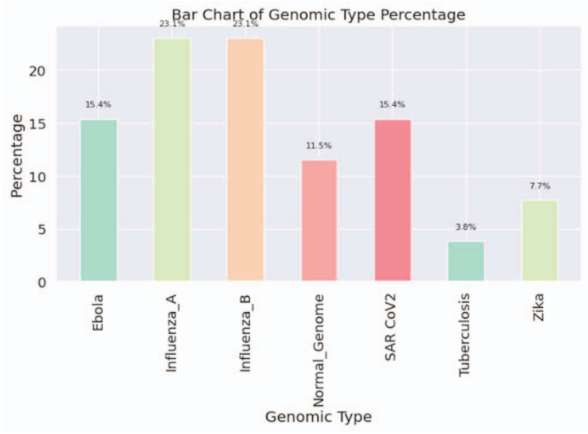

Figure 2: Infectious Disease Dataset Distribution

This process commenced with the transcription of genomic sequences into RNA sequences, a critical step for capturing essential genetic information encoded within DNA. Subsequently, the RNA sequences underwent translation into protein sequences. This conversion to protein sequences held paramount importance, as proteins function as the workhorses within biological systems, unveiling concealed genetic information that remains inaccessible through raw genomic sequences alone. This transformation paved the way for more accurate and early disease detection. The resulting protein sequences were characterized by alphanumeric representations, encompassing a range of protein counts. Each of these sequences was attributed to a label, signifying the specific disease category to which it belonged. The methodology adopted for this study is depicted by Figure. 2.

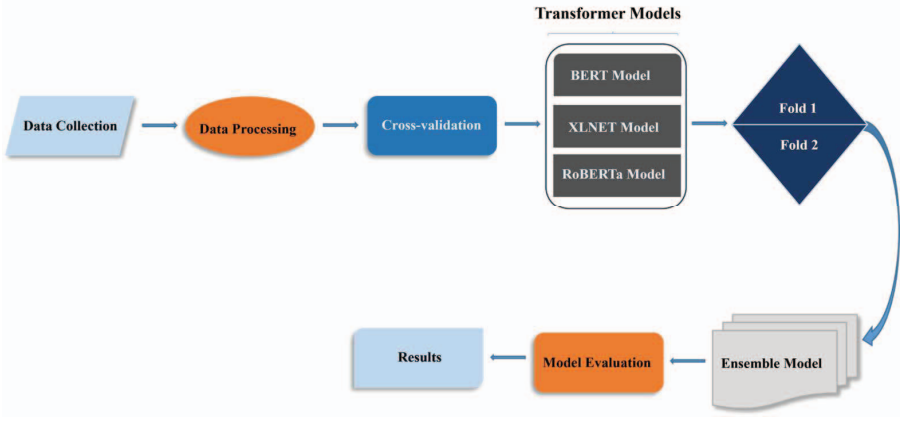

Figure 1: Process flow diagram of the proposed architecture

In preparation for model training, these labels underwent a transformation through the label encoding technique, which assigned a unique positive integer to each label. Given that machines cannot process string-based data, it was imperative to further convert the protein sequences into a machine-readable format. To accomplish this, model-compatible tokenization techniques were applied to the protein sequences. This method translated the sequences into numerical vectors, with each vector serving as a representation of a particular protein sequence.

During the training phase, three pretrained language models were considered, namely BERT, XLNET, and RoBERTa. To ensure optimal performance during training, essential hyperparameters were carefully configured, including the utilization of the AdamW optimizer, a batch size of 64, and training for 20 epochs. The performance of these models was thoroughly assessed and compared to identify the most effective model for the given task.

The accuracy of individual models and ensemble methods is comprehensively assessed using various metrics such as precision, recall, and the F1-score, all of which are critical for the precise classification of diseases based on genomic sequences. When trained individually, BERT, XLNET and RoBERTa models achieved accuracies of 92%, 96% and 85% respectively. We decided to perform some ensemble techniques to see how it can better boost the accuracy of predictions. The majority voting ensemble combined predictions, achieving an accuracy rate of 92%, underscoring its effectiveness in enhancing predictive capabilities. Furthermore, the mean average approach, boasting an accuracy rate of 85%, assigns weighted contributions from individual models based on their performance. Both ensemble strategies exhibit promising outcomes in enhancing the overall accuracy of disease prediction.

To mitigate the risk of overfitting, a phenomenon wherein a model fails to generalize to unseen data, a strategic approach was adopted. This involved employing a 2-fold K-fold cross-validation strategy, which entailed varying the training and test datasets systematically. Additionally, an experimental approach was explored, which involved combining the predictions generated by each of these transformer models using techniques such as majority voting and mean averaging. The performance of these ensemble techniques was also rigorously evaluated and compared with the results derived from the previous approach, providing a comprehensive assessment of their efficacy in enhancing predictive accuracy.

The k-fold validation approach when used with the transformer models, had its accuracies from each fold combined using ensemble techniques. Using this approach, BERT and RoBERTa models each achieved an accuracy of 96.2% while the XLNET model achieved an accuracy of 88.5%.

IV. RESULTS

In this section, the performance of each model is presented and discussed. The performance of the BERT, XLNET, and RoBERTa models are assessed, and furthermore compared to the performance of the Ensemble techniques such as the Majority voting and Weighed average. Visualization of the results from each model can also be found in this section.

A. BERT Performance

Two methods were used to access the BERT model's performance on the genomic sequence of various infectious diseases. The first method entailed training the BERT model without splitting into test and train dataset. Here we allowed the model to split the datasets itself and perform prediction. The summary of the model's performance is shown in Table 1.

Table 1. Classification Report for BERT

Class	Precision	Recall	F1-score	Support
Ebola	0.80	1.00	0.89	4
Influenza_A	0.86	1.00	0.92	6
Influenza_B	1.00	0.83	0.91	6
Normal_Genome	1.00	1.00	1.00	3
SAR CoV2	1.00	1.00	1.00	4
Tuberculosis	0.00	0.00	0.00	1
Zika	1.00	1.00	1.00	2

From the classification report in Table 1, it can be observed that the model performed well on all but one of the datasets having precision scores between 80% and 100%. However, it is noteworthy that the precision score for the Tuberculosis dataset was 0%. This result stems from the fact that only one dataset was available for model training (support = 1), limiting its ability to generalize.

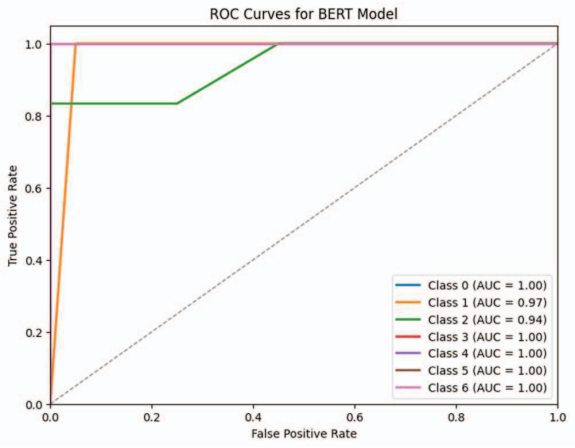

Figure 3: ROC curve and AUC score for BERT

The ROC Curve in Figure 3 shows that the model performed well on all classes with a closer tilt to the top left of the graph. In Figure 4 we can see that only two of the datasets **Influenza_B** and Tuberculosis were not well classified. In the case of Influenza_B, it is suspected that the model detected certain similarities in the protein sequences of **Influenza_A**, resulting in the misplacement of one of the datasets. Overall, the model achieved an impressive overall accuracy of 92%.

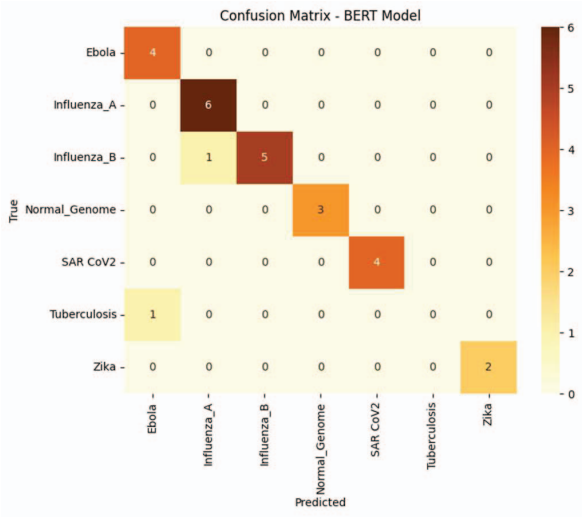

Figure 4: Confusion Matrix BERT

For us to ensure validity with the model predictions, we employed a second method which made use of cross validation. The data sets were split into n_folds = 2 and the model was trained in two folds after which the accuracies were ensembled to give an average accuracy score. The accuracy score for this was 96.2%.

Table 2. Classification Report - BERT with cross validation ensemble

Class	Precision	Recall	F1-score	Support
Ebola	1.00	1.00	1.00	4
Influenza_**A**	0.86	1.00	0.92	6
Influenza_B	1.00	0.83	0.91	6
Normal_Genome	1.00	1.00	1.00	3
SAR CoV2	1.00	1.00	1.00	4
Tuberculosis	1.00	1.00	1.00	1
Zika	1.00	1.00	1.00	2

As can be observed in Table 2, the BERT model performed better with the classes and prediction when the cross validation and ensemble of its result is employed. The Tuberculosis class now has a precision accuracy of 100% as opposed to the result gotten in Table 1. Despite these enhancements, Figure 5 reveals that the Influenza_B dataset continued to pose challenges for proper classification even after the implementation of cross-validation and result ensembling. In summary, this alternative methodology demonstrated superior performance compared to the approach outlined in Table 1.

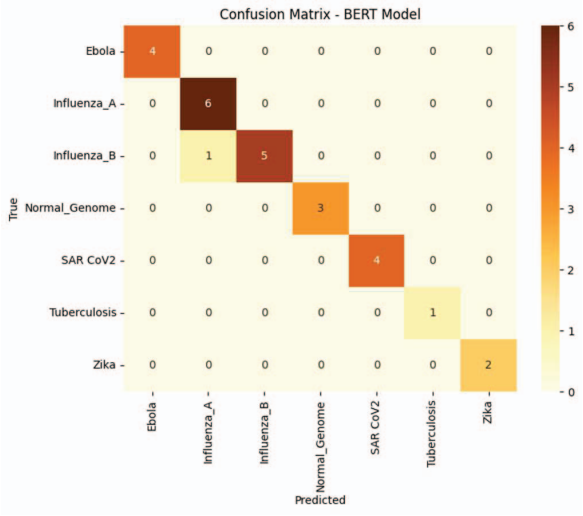

Figure 5: Confusion Matrix - BERT with cross validation ensemble

B. XLNET Performance

In this context, we assessed the XLNET model's performance on genomic sequences associated with diverse infectious diseases through the application of two distinct methodological approaches. Initially, we pursued a strategy in which the XLNET model was trained without the conventional division of data into explicit test and training sets. Instead, the model was granted the autonomy to independently partition the dataset and perform predictive tasks. The summarized outcomes of this performance evaluation are presented in Table 3.

Table 3. Classification Report XLNET

Class	Precision	Recall	F1-score	Support
Ebola	1.00	1.00	1.00	4
Influenza_**A**	1.00	1.00	1.00	6
Influenza_B	1.00	1.00	1.00	6
Normal_Genome	1.00	1.00	1.00	3
SAR CoV2	0.80	1.00	0.89	4
Tuberculosis	1.00	1.00	1.00	1
Zika	1.00	0.50	0.67	2

Analyzing the classification report in Table 3 reveals that the XLNET model exhibited strong performance across most datasets, exhibiting precision scores ranging from 80% to 100%.

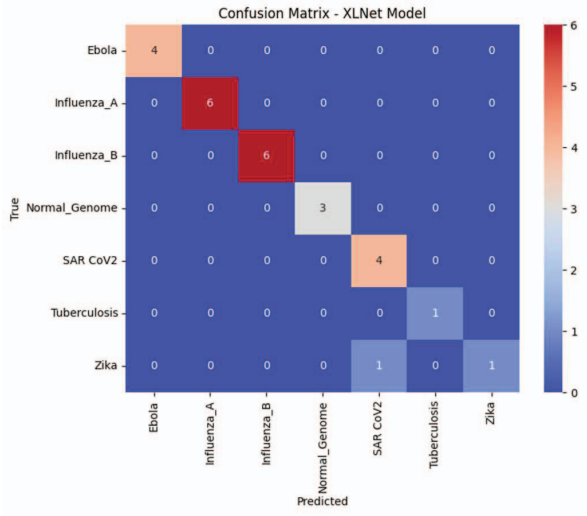

Figure 6: Confusion Matrix XLENT

Figure 6 highlights that only one dataset, namely Zika, exhibited suboptimal classification results. Overall, the model achieved an impressive overall accuracy of 96%.

Table 4. Classification Report - XLNET with cross validation ensemble

Class	Precision	Recall	F1-score	Support
Ebola	0.80	1.00	0.89	4
Influenza_**A**	1.00	1.00	1.00	6
Influenza_B	1.00	1.00	1.00	6
Normal_Genome	1.00	1.00	1.00	3
SAR CoV2	1.00	1.00	1.00	4
Tuberculosis	1.00	1.00	1.00	1
Zika	1.00	0.50	0.67	2

To bolster the validity of our model predictions, we adopted a second methodology that incorporates cross-validation. The dataset was divided into n_folds = 2, and the model was trained separately on each fold. Subsequently, the accuracy scores were combined to yield an average accuracy score of 96.2% which is the same as the accuracy achieved when similar method was employed in the BERT model.

As elucidated in Table 4, the utilization of cross-validation and ensemble techniques substantially improved the XLNET model's performance in terms of class prediction. Notably, the SAR CoV2 class achieved a precision accuracy of 100%, a marked improvement from the result presented in Table 3. Nevertheless, Figure 7 highlights that the Zika dataset remained problematic for accurate classification even with the incorporation of cross-validation and result ensembling. In summary, this alternate approach showcased a moderately improved performance when contrasted with the method detailed in Table 3.

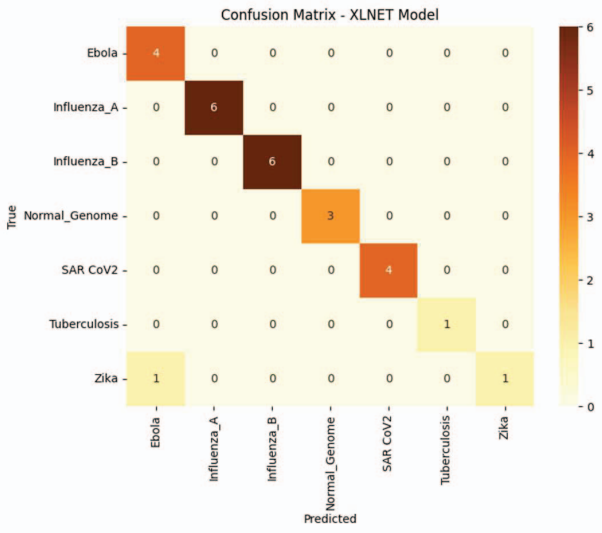

Figure 7: Confusion Matrix - XLNET with cross validation ensemble

C. RoBERTa Performance

In continuation of the research, we undertook an evaluation of the RoBERTa model's effectiveness in analyzing genomic sequences associated with a variety of infectious diseases using two distinctive methodological paradigms.

Table 5. Classification Report RoBERTa

Class	Precision	Recall	F1-score	Support
Ebola	0.80	1.00	0.89	4
Influenza_**A**	0.75	1.00	0.86	6
Influenza_B	1.00	0.83	0.91	6
Normal_Genome	1.00	1.00	1.00	3
SAR CoV2	0.80	1.00	0.89	4
Tuberculosis	0.00	0.00	0.00	1
Zika	0.00	0.00	0.00	2

Initially, we adopted an unconventional approach by training the RoBERTa model without the customary segregation of data into explicit test and training sets. Instead, we empowered the model to autonomously partition the dataset and execute predictive tasks. The concise summary of the outcomes derived from this performance assessment can be found in Table 5.

An examination of the classification report in Table 5 unveils that the RoBERTa model exhibited a reasonably good performance across most datasets, displaying precision scores spanning from 75% to 100%. It's worth noting that a few classes attained precision scores of 0%. Figure 8 underscores that three datasets, specifically Influenza_B, Tuberculosis, and Zika, yielded suboptimal classification results. Overall, the model attained an aggregate accuracy rate of 85%.

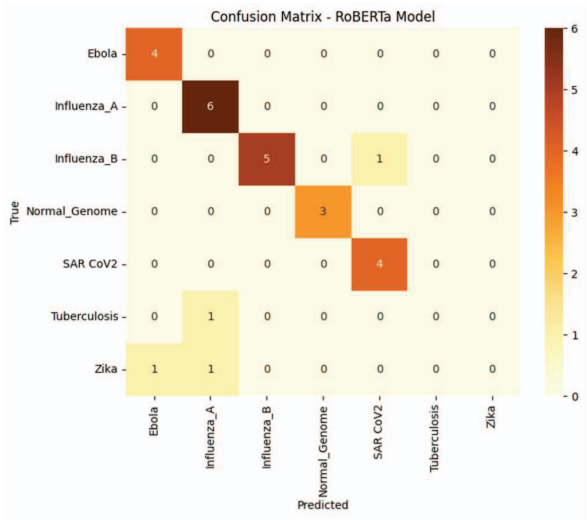

Figure 8: Confusion Matrix RoBERTa

To enhance the credibility of our model's predictions, we employed a secondary methodology that incorporates cross-validation. We divided the dataset into n_folds = 2, and the model was trained separately on each fold. Subsequently, the accuracy scores were amalgamated to yield an average accuracy score of 88.5%.

Table 6. Classification Report - RoBERTa with cross validation ensemble

Class	Precision	Recall	F1-score	Support
Ebola	0.80	1.00	0.89	4
Influenza_**A**	0.86	1.00	0.92	6
Influenza_B	0.86	1.00	0.92	6
Normal_Genome	1.00	1.00	1.00	3
SAR CoV2	1.00	1.00	1.00	4
Tuberculosis	0.00	0.00	0.00	1
Zika	0.00	0.00	0.00	2

As elucidated in Table 6, the implementation of cross-validation and ensemble techniques led to a modest improvement in the RoBERTa model's class prediction performance. Notably, the **Influenza_B** class was successfully classified, whereas Zika and Tuberculosis continued to pose classification challenges, as evidenced by the results presented in Figure 9. Even with the inclusion of cross-validation and result ensembling, these two datasets remained difficult to classify accurately. In summary, this alternative approach demonstrated a moderate enhancement in performance when compared to the methodology outlined in Table 5.

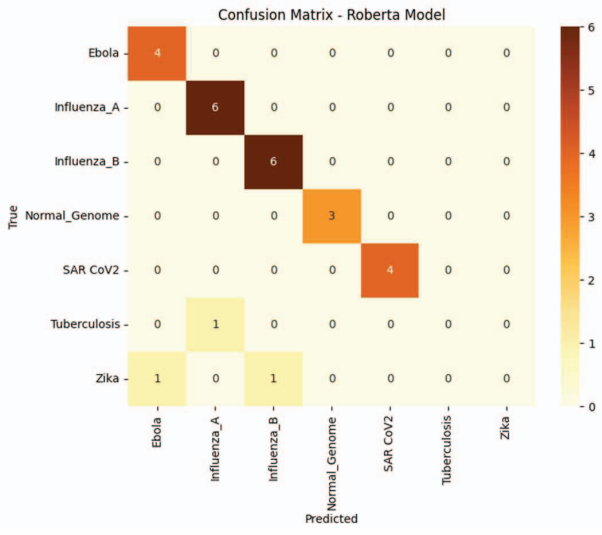

Figure 9: Confusion Matrix - RoBERTa with cross validation ensemble

D. Transformer Ensemble Performance

Initially, an evaluation of various transformer models was conducted to assess their effectiveness in analyzing genomic sequences associated with infectious diseases. It is noteworthy that our research approach diverged from conventional practices, as we omitted the standard practice of partitioning data into distinct test and training sets. Instead, each model autonomously partitioned the dataset and performed predictive tasks independently. A succinct overview of the outcomes derived from these preliminary evaluations can be found in Tables 1, 3, and 5.

In our continuous endeavor to enhance the accuracy of disease classification, we proceeded to implement ensemble techniques across all transformer models utilized in our study. This ensemble methodology was exclusively applied to the results obtained from models for which cross-validation had not been utilized.

Table 7. Classification Report Ensemble Majority Voting

Class	Precision	Recall	F1-score	Support
Ebola	0.80	1.00	0.89	4
Influenza_**A**	0.86	1.00	0.92	6
Influenza_B	1.00	1.00	1.00	6
Normal_Genome	1.00	1.00	1.00	3
SAR CoV2	1.00	1.00	1.00	4

Tuberculosis	0.00	0.00	0.00	1
Zika	1.00	0.50	0.67	2

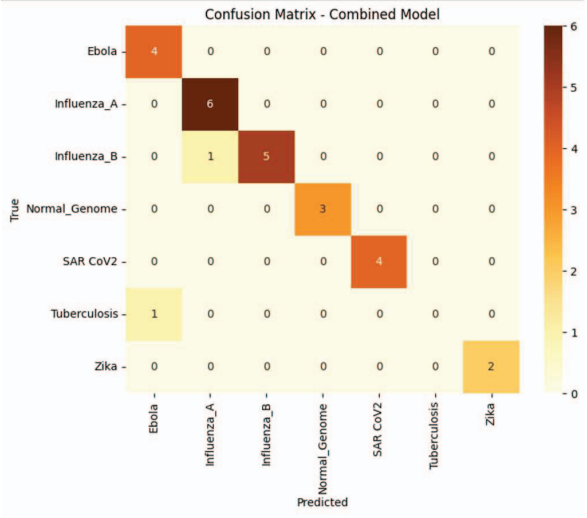

Figure 10: Confusion Matrix Ensemble Majority Voting

The two ensemble techniques employed in this context were Majority Voting and Weighted Average. Comprehensive summaries of the classification reports for each technique are provided in Tables 7 and 8. Remarkably, Majority Voting outperformed Weighted Average, achieving a remarkable accuracy rate of 92%, while Weighted Average achieved an accuracy of 85%. It is imperative to note that the Weighted Average technique incurred classification errors in three distinct classes, as depicted in Figure 11, specifically (***Tuberculosis, Zika,*** and ***Influenza_B***), whereas Majority Voting exhibited misclassifications in two classes, illustrated in Figure 10, namely ***Tuberculosis*** and ***Influenza_B***.

In summary, our ensemble methodology did not necessarily yield more favorable results when compared to the initial approaches outlined in Tables 1, 3, and 5. It is worth emphasizing that several methodologies were explored in our pursuit to determine the most effective approach for infectious disease prediction. Of all the methods evaluated thus far, the XLNET and BERT models, when integrated with cross-validation ensembles, have emerged as the top-performing models in our comprehensive study with accuracies of 96.2% each.

Table 8. Classification Report Ensemble Weighted Average

Class	Precision	Recall	F1-score	Support
Ebola	1.00	1.00	1.00	4
Influenza_A	1.00	1.00	1.00	6
Influenza_B	0.83	0.83	0.83	6
Normal_Genome	0.75	1.00	0.86	3
SAR CoV2	0.67	1.00	0.80	4
Tuberculosis	0.00	0.00	0.00	1
Zika	0.00	0.00	0.00	2

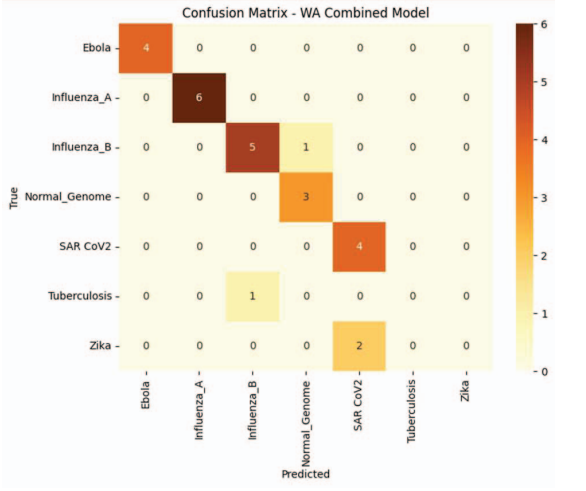

Figure 11: Confusion Matrix Ensemble Weighted Average

V. CONCLUSION

Infectious disease prediction and monitoring represent critical challenges for global public health. This research leveraged transformer models, including BERT, XLNET, and RoBERTa, to predict infectious agents from genomic sequences. Overall, the results demonstrated the efficacy of the transformer models – BERT and XLNET, particularly when integrated with cross-validation ensembling, with accuracies of 96.2% each. While ensemble techniques showed promise, they did not consistently outperform individual models. Moreover, the model's performance on diseases with low dataset, such as zika and tuberculosis, was limited. Only two datasets were found for zika and for tuberculosis – a bacteria, we decided to use only one dataset to see how the model might perform on it. Future studies should expand the dataset to include a broader range of infectious diseases and develop real-time prediction and visualization systems for rapid disease identification from human DNA sequences.

ACKNOWLEDGEMENT

This work is supported by the National Science Foundation (NSF) grant (ID 2131307) under CISE-MSI program.

REFERENCES

[1] D. M. Morens and A. S. Fauci, "Emerging Pandemic Diseases: How We Got to COVID-19," *Cell*, vol. 182, no. 5, pp. 1077–1092, Sep. 2020, doi: 10.1016/j.cell.2020.08.021.

[2] S. Yang and R. E. Rothman, "PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings," *The Lancet Infectious Diseases*, vol. 4, no. 6, pp. 337–348, Jun. 2004, doi: 10.1016/S1473-3099(04)01044-8.

[3] D. Jungkind, “Molecular Testing for Infectious Disease,” *Science*, vol. 294, no. 5546, pp. 1553–1555, Nov. 2001, doi: 10.1126/science.294.5546.1553.

[4] R. V. Francis *et al.*, “The Impact of Real-Time Whole-Genome Sequencing in Controlling Healthcare-Associated SARS-CoV-2 Outbreaks,” *The Journal of Infectious Diseases*, vol. 225, no. 1, pp. 10–18, Jan. 2022, doi: 10.1093/infdis/jiab483.

[5] Cho, Kyong Hwan, et al. "Convolutional neural networks for human action recognition." In Proceedings of the 25th international conference on neural information processing systems, pp. 2673-2681. 2012.

[6] Li, Rui, et al. "Epidemic dynamics of cholera in Haiti: model for the 2010-2011 epidemic." PLoS neglected tropical diseases 6.5 (2012): e1483.

[7] Vaswani, Ashish, et al. "Attention is all you need." In Advances in neural information processing systems, pp. 30-38. 2017.

[8] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[9] Ren, Zhida, et al. "EpiDeep: Exploiting gis information in convolutional neural networks for the prediction of epidemic outbreaks." ISPRS International Journal of Geo-Information 8.11 (2019): 486.

[10] Long, Yindong, et al. "Real-time epidemic surveillance using social media data with deep learning approach." Journal of medical internet research 20.7 (2018): e185.

[11] Wu, Yijun, and Jonghyun Choi. "Coronavirus disease 2019 (COVID-19) diagnosis using a multi-feature classification model." Sensors 20.22 (2020): 6456.

[12] Yan, Li, et al. "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images." Scientific Reports 10.1 (2020): 1-12.

[13] Gardy, Jennifer L., et al. "Whole-genome sequencing and social-network analysis of a tuberculosis outbreak." New England Journal of Medicine 364.8 (2011): 730-739.

[14] Hadfield, James, et al. "Nextstrain: real-time tracking of pathogen evolution." Bioinformatics 34.23 (2018): 4121-4123.

[15] Goodfellow, I., et al. (2016) Deep Learning. MIT Press, Cambridge, MA. http://www.deeplearningbook.org