SARS-CoV-2 reservoir prediction using machine learning and natural language processing

Adithya Madhusoodanan[b], Emmanuel Kagning Tsinda[a], Anthony Dunn[c], Alain Zemkoho[c]

[a] Graduate School of Medicine, University of Tohoku, Sendai 980-8575, Japan
[b] National Institute of Technology Karnataka Surathkal, Karnataka 575025, India
[c] School of Mathematical Sciences, University of Southampton, SO17 1BJ Southampton, United Kingdom

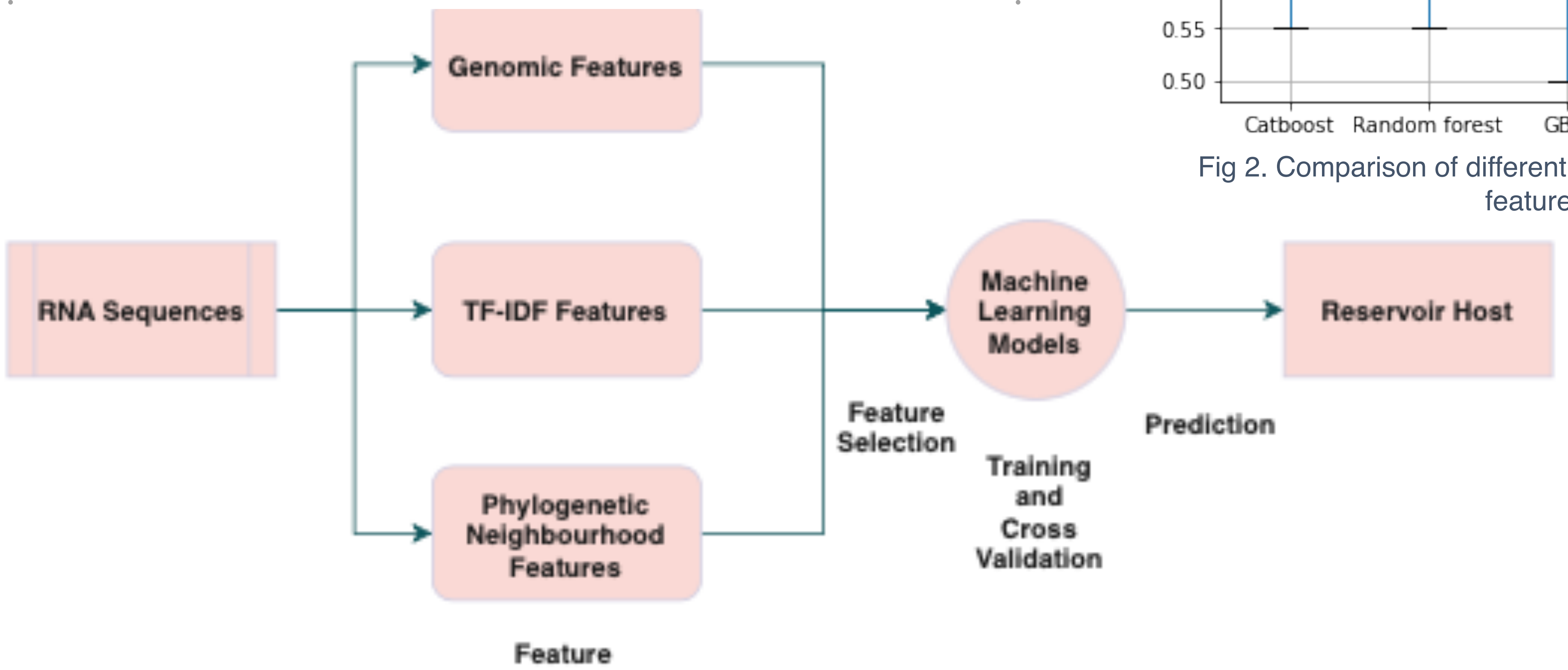
The spread of the SARS-CoV-2 has led to a global pandemic. With the origin of the virus still unknown, some fundamental questions about the it remain open. Traditional methods to identify reservoir hosts involve the analysis of the phylogenetic relationships among viruses. Our approach to achieve this goal in this work is based on using machine learning and natural language processing techniques to predict SARS-CoV-2 genome sequences. The training of our model is done on data sets made of RNA sequences whose reservoirs are known. Term Frequency Inverse Document Frequency (TF-IDF), dinucleotide biases, and codon pair scores techniques are used to generate features from the sequences. These features are then used in the training process of multiple machine learning models. The resulting method could then be used for the effective and time-efficient discovery of reservoirs of unknown or new viruses, which could subsequently help in a better understanding of the virus and support further vaccine development.

OBJECTIVE

This work aims at developing a novel natural language processing based technique for extracting features from RNA sequences. These feature along with other Genomic and Phylogenetic Neighbourhood features was used to train robust machine learning algorithms to predict Reservoir-host of SARS-COV -2 sequences.

MATERIALS AND METHOD

Babayan Et. al [1] collected reservoir hosts for known RNA sequences of viruses. Another dataset that we used which was collected by Brieley L consists of host categories of RNA virus sequences collected from NCBI database. Genomic traits in the sequences was quantified by calculating the Dinucleotide bias for each of the 16 possible dinucleotides And codon pair score for each of the 4,096 possible codon-codon pairs. 64 tri-mers of the RNA sequences were used to generate term frequency inverse document frequency features as well. Phylogenetic neighbourhood traits for the sequencing were also obtained by blasting the sequnces and by calculating the relative support for the reservoirs. Using XGBoost [5] classifier most select the important features. The number of features to be selected was also optimised using cross validation. These selected features were then used to train different machine learning models like extra tree classifiers, gradient boosted machines, XGBoost classifiers and random forests. The best model among them was selected by performing cross validation. These models could then be used to predict the reservoir hosts for new RNA virus sequences. The features generated from SARS-COV-2 sequences collected was then given to the model to predict their Reservoirs.


RNA Sequences

Genomic Features

TF-IDF Features

Phylogenetic Neighbourhood Features

Feature

Feature Selection

Machine Learning Models

Training and Cross Validation

Prediction

Reservoir Host

Fig 1. Feature generation and model training testing workflow.

RESULTS

Among the features generated from the RNA sequences the Phylogenetic neighbourhood traits was ranked the highest. Some of the TFIDF features were also ranked among top 10 important features. Fig 2 shows the comparison between different models for selected 25 important features. The Extra trees[4] classifier model was found to have the highest performance. Fig 3. Compares the accuracies of different number of selected features for extra tress classifier. Model trained with 25 most important features was found to have the highest accuracy.

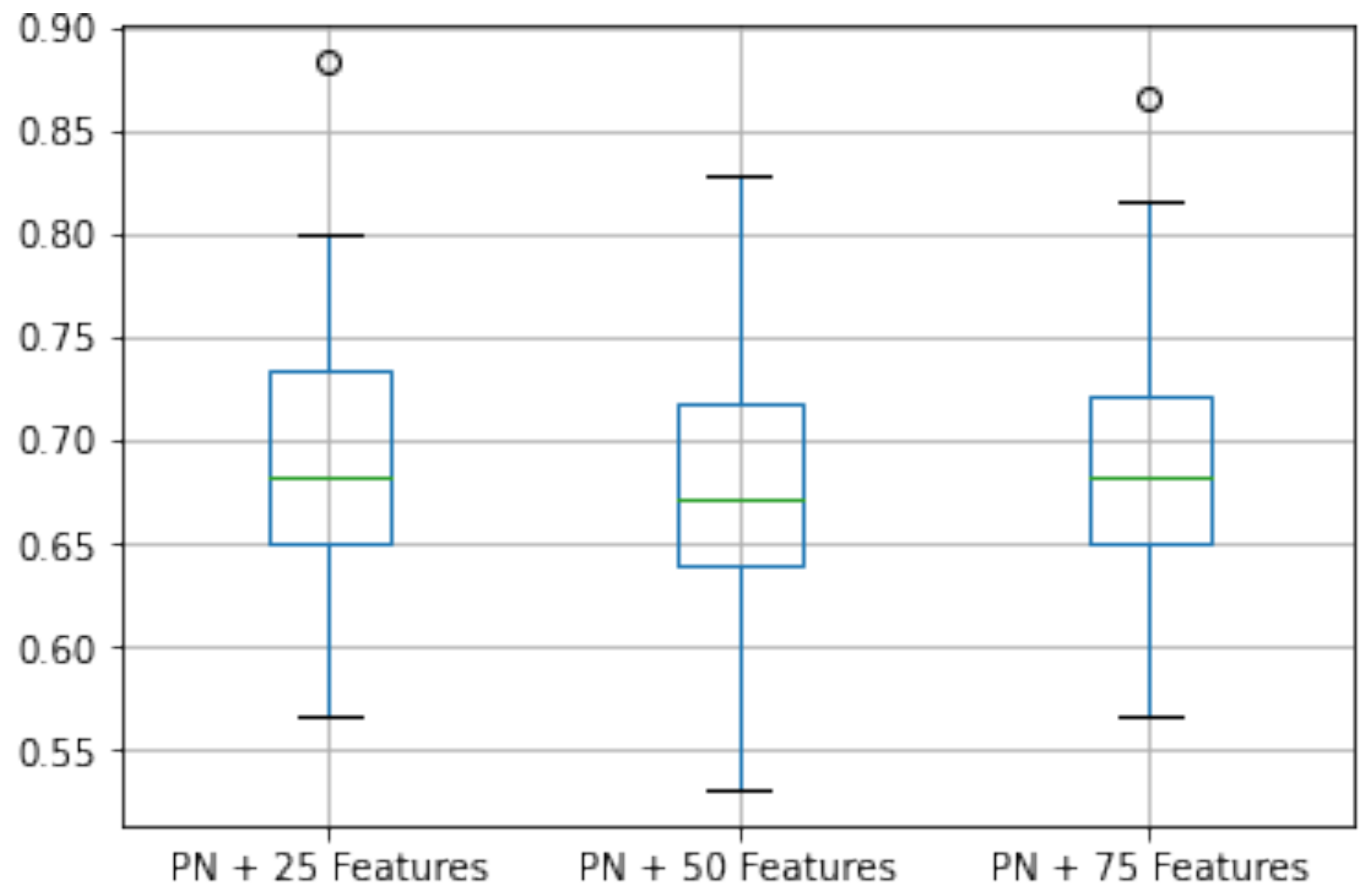


Fig 3. Comparison of extra trees classifier for different number of selected features

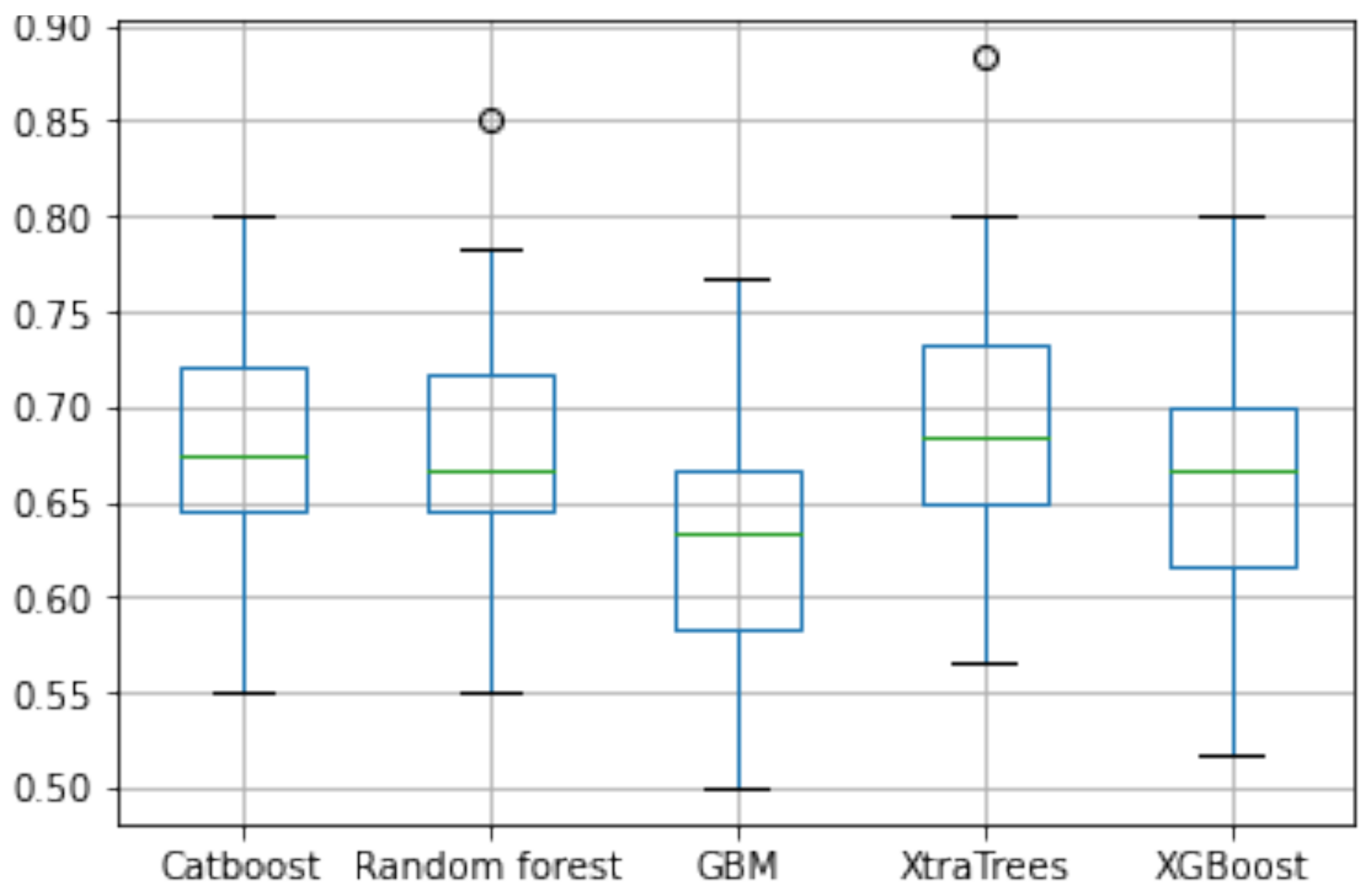


Fig 2. Comparison of different models with 25 selected features

CONCLUSION

The proposed techniques for reservoir prediction and feature extraction is proved to be very effective. Machine learning based reservoir prediction could help in the time-efficient discovery of reservoirs of unknown or new viruses. This could further help in the spread of viral diseases and also in the development of vaccine.

REFERNCES

[1]Babayan SA, Orton RJ, Streicker DG. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. Science. 2018;362: 577–580. pmid:30385576

[2] Brierley L, Fowler A (2021) Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. PLOS Pathogens 17(4): e1009149. https://doi.org/10.1371/journal.ppat.1009149

[3]Amir Jalilifard, Vinicius F. Caridá, Alex F. Mansano, Rogers S. Cristo, Felipe Penhorate C. Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. https://arxiv.org/abs/2001.09896v2

[4] Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach Learn* **63,** 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. DOI:https://doi.org/10.1145/2939672.2939785

[6] Haihao Lu, & Rahul Mazumder. (2020). Randomized Gradient Boosting Machine.

[7] Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001). https://doi.org/10.1023/A:1010933404324