Journal of Advanced Health Informatics Research (JAHIR)
Vol. 2, No. 3, 2024, pp. 127-137
ISSN: 2985-6124, DOI: 10.59247/jahir.v2i3.298

Virus Host Prediction with Metagenomic Features using Support Vector Machine Algorithm and Grid Search Cross Validation Optimization

Purwono Purwono [1], Annastasya Nabila Elsa Wulandari [1], Novieta Hardeani Sari [2]
[1] *Department of Informatics, Universitas Harapan Bangsa, Indonesia*
[2] *Newcastle University, United Kingdom*

ARTICLE INFO

Article history:

Received September 13, 2024
Revised November 27, 2024
Published December 30, 2024

Keywords:

Virus;
Host;
Metagenomics;
Support vector machine;
Grid search

ABSTRACT

Viruses and bacteria continue to evolve alongside humans. Viruses are spreading too fast and causing a huge loss of life in the world. Viruses play an important role as dangerous pathogens that continue to spread various infectious diseases. Metegenomics is the application of large sequencing technology to genetic material obtained directly from one or more environmental samples, resulting in at least 50Mb random samples and multiple long sequences. It is important to identify the origin of the virus to prevent the spread of outbreaks. Understanding the biology of these viruses and how they affect their ecosystems depends on knowing which host they infect. We can use metagenomic features derived from the viral genome to determine the type of virus host. The activity of predicting virus hosts has traditionally taken a lot of time and effort in the process. Technology can be one of the solutions that can be used to predict virus host types. One of the technologies that can be used is machine learning. We chose one of the machine learning algorithms, SVM, to predict viral hosts with metagenomics features, namely genome size, GC% and number of CDS from viral genomes derived from 7326 viral genomes. The SVM model was further optimised with GS and K-CV methods. This optimisation resulted in an increase in the accuracy value of the model when predicting virus hosts from 80% to 84%.

Corresponding Author:

Purwono, Department of Informatics, Universitas Harapan Bangsa, Indonesia
Email: purwono@uhb.ac.id

1. INTRODUCTION

Viruses and Bacteria continue to evolve in the world [1]. The evolution of the virus apparently occurred simultaneously with humans and has resulted in a very high number of deaths [2]. One example is the Covid-19 virus, which has evolved into various variants. Based on data from the World Health Organisation (WHO),

the Covid-19 pandemic has hit the world with more than 236 million confirmed cases and nearly 5 million deaths occurring until October 2021 [3].

Viruses play an important role as pathogens that continue to spread various infectious diseases [4]. They have a complex molecular architecture and have morphed into highly destructive parasites [5]. Viruses consist of a genome enclosed in a protein or proteolipid shell [5].

Nontargeted metagenomics is the application of large sequencing technologies to genetic material obtained directly from one or more environmental samples, resulting in at least 50Mb random samples and multiple long sequences [6]. Metagenomic sequencing data provides a rich resource for expanding our understanding of differential protein functions involved in human health [7]. The use of metagenomic sequencing is dramatically increasing our understanding of the evolution and ecology of microbial systems in a variety of environments, from water and soil to the human body [8].

The host of the virus needs to be found to know where a virus comes from. It is important to identify the origin of the virus to prevent the spread of an outbreak [9]. Understanding the biological side of these viruses and how they affect their ecosystems depends on knowing which hosts they infect [10]. We can use metagenomic features derived from the viral genome to determine the virus host type. The activity to predict the virus host by traditional means is time-consuming and labour-intensive [11].

Technological developments have been steadily increasing in the healthcare sector in the last 40 years [12]. Technology can be one of the solutions that can be used to predict the type of virus host [13]. Popular technologies such as machine learning, which is one of the fields of artificial intelligence, can be applied to this [9]. Machine learning has various types of learning including supervised learning, unsupervised learning, deep learning, ensemble learning, semi-supervised learning, reinfocement learning, outlier detection and metric learning [14]. Supervised learning allows us to acquire or generate data based on prior knowledge. In supervised learning, the data used for training must be selected and handled appropriately [15]. One type of supervised learning machine learning algorithm is Support Vector Machine (SVM).

Virus host prediction will be predicted based on the features of genome size, GC (guanine-cytosine) %, and the number of CDS (coding sequence) of the virus genome using the support vector machine algorithm. SVM will process the predefined features to find out where a virus comes from.

The machine learning model created in predicting the virus host can then be optimised with various types of hyperparameter tunning methods. These methods are used to improve the performance of previously created machine learning models [16]. This optimisation aims to help get better performance faster than tuning with randomly selected hyperparameters. One type of hyperparameter optimisation that can be used is grid search cross validation. Grid search is used to generate a suitable evaluation index and select the best hyper parameter [17].

2. METHODS

Several types of research related to virus host prediction have been conducted by previous researchers. We summarise the different types of research and attempt to explain our contribution to this work. Research conducted by Alakus [18] took a computational approach to predicting virus hosts. FIBHASH, AVL-tree and entropy-based methods were applied to this field. As a result, the viral host interaction between SARS-CoV-2 and human proteins was effectively classified through deep learning. Research that has been conducted by Holcomb [19] created a pipeline that identifies coagulation-related proteins that interact with the SARS-CoV-2 protein. The pipeline further searched databases such as COVID-19 HGI for genetic variants of its host proteins. This research found similar motifs and regions of protein-protein interactions in the viral host system. Research that has been conducted by Das [20] conducted a possible SARS-COV-2 attack of proteins in 17 signalling pathways. This research resulted in a scheme to infer viral host associations based on codon usage patterns identifying the most affected signalling pathways during COVID-19. Research that has been conducted by Dey [13] has built an ensemble voting classifier using Radial SVM, Polynomial SVM, and Random Forest techniques that provides better accuracy, precision, specificity, recall, and F1 score compared to all other models used. A total of 1326 potential human target proteins of SARS-CoV-2 were predicted by the proposed ensemble model and validated using gene ontology and KEGG pathway enrichment analysis. Research that has been conducted by Xu [9] has created a machine learning model to predict influenza virus hosts. The results show that the 5-gram artificial neural network is the most effective algorithm for predicting the origin of virus sequences, with about 99.54% AUCPR, 98.01% score and 96.60% MCC at higher classification levels, and about 94.74% AUCPR, 87.41% score and 80.79% MCC at lower classification levels. In contrast to previous studies, we seek to apply optimisation to the hyperparameters used by the SVM

algorithm. A grid search method with cross validation is proposed to produce better prediction model performance.

3. PROPOSED METHOD

3.1 Dataset

The data used in this study is taken from the NCBI website which will be adjusted to the needs of the machine learning model. We used 7326 viral genomes that are unique and their characteristics are unknown. The data can be obtained at the following NCBI website https://www.ncbi.nlm.nih.gov/genome/browse/#!/viruses/. The unique virus hosts in this dataset arebacteria, fungi, plants, vertebrates, invertebrates, protozoa, vertebrates, invertebrates, human, invertebrates, plants, algae, vertebrates, invertebrates, vertebrates, human, archaea, human, nan.

The main metagenomics features used in this dataset are genome size, GC%, and CDS count of the viral genome used as features to predict the virus host. The distribution of genome size can be seen in Figure 1. The distribution of GC% of the viral genome can be seen in Figure 2 and the distribution of CDS of the viral genome can be seen in Figure 3.

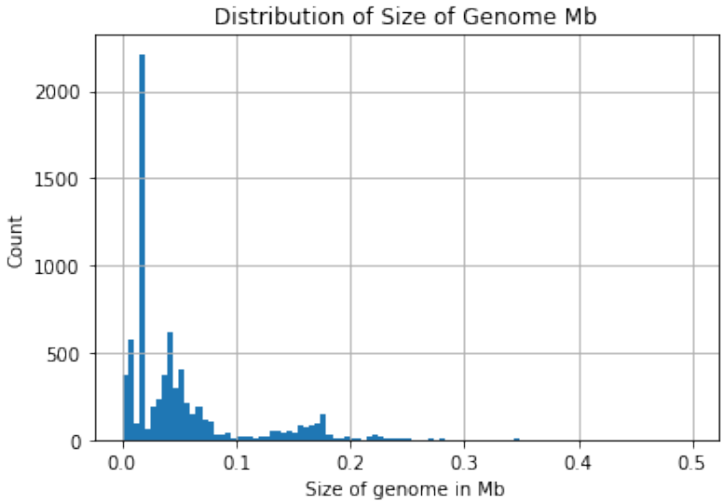

Fig. 1. Distribution of Size of Genome

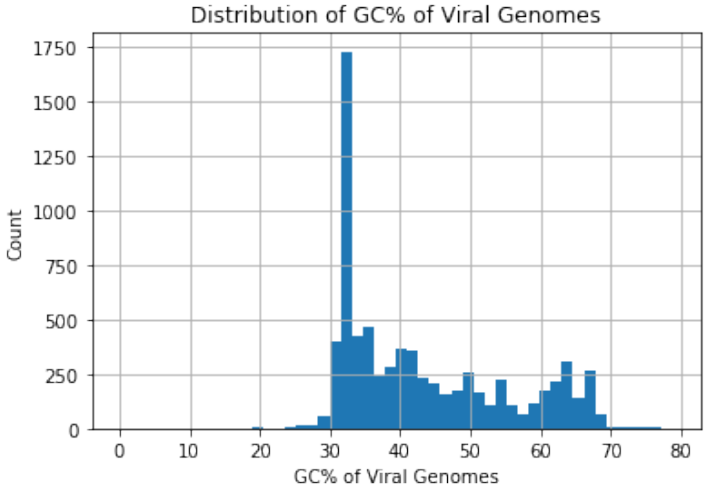

Fig. 2. Distribution of GC% of Viral Genomes

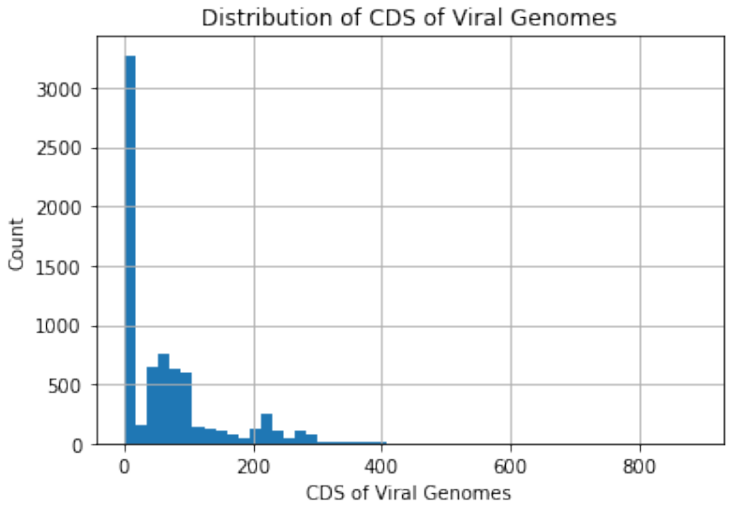

Fig. 3. CDS of Viral Genomes

The distribution of genome size consists of 5,233 unique data points, with a mean value of 0.051478. This dataset provides insights into the variation in genome size across different viral genomes. The distribution of GC content in viral genomes includes 2,789 unique data points, with an average GC percentage of 43.230327. This indicates the proportion of guanine (G) and cytosine (C) bases in the genome, which plays a crucial role in genome stability and functionality. The coding sequence (CDS) distribution of viral genomes comprises 363 unique data points, with a mean value of 67.090736. The CDS represents the protein-coding regions of the viral genome, which are essential for understanding viral gene expression and functional characteristics. These statistical distributions provide valuable insights into the structural and functional properties of viral genomes, supporting further analysis in comparative genomics and evolutionary studies.

3.2 Pre-processing

Datasets first pass through a data pre-processing stage in order to be recognised by machine learning models. Data pre-processing, such as normalisation, feature extraction, and dimensionality reduction, is necessary to better complete data classification [21]. Data processing stages that may be used by the SVM algorithm are categorical to numerical processes such as Percentage Categorical Pruned (PCP), Inverse Document Frequency (IDF) and the simpler One-Hot-Encoding method [22]. The data that has been processed at this stage is then processed with the support vector machine algorithm.

3.3 Support Vector Machine

Support Vector Machine is a widely used algorithm in supervised prediction and classification [23]. In classification SVM generates a function that learns from the training dataset to find the best *hyperplane* by maximising the distance between classes on different data. In SVM prediction depends on a subgroup of the original dataset, whose elements are interpreted as support vectors and are in charge of margin selection [24].

The data used in the SVM algorithm is represented in an n-dimensional space that is in charge of predicting whether new training instances belong to the same or different class categories [25]. An important goal of the SVM algorithm is to find the best hyperplane in the n-dimensional space that is able to classify or predict the data points [25]. The pre-processed data is transformed into a numerical format with encoding labels so that it is easily recognised by this algorithm. With this numerical format, the process of comparing the same or different classes becomes easier [1].

SVM considers a set Ω of vectors that are divided into two different classes. Vectors are denoted as pairs $(x_i, y_i) \epsilon\ \mathbb{R}^n x\{-1,1\}$where n is the number of features observed by each vector, x_i contains the feature value for vector i and y_i indicates which of the two classes Ω vector i belongs to [23]. SVM uses a straight-line kernel that divides two different classes with a linear equation that can be seen in Equation 1 [26].

$$w * x - b = 0 \quad (1)$$

Based on equation (1), w is the *hyperplane* parameter whose value is being sought, x for input data and b is bias. Furthermore, to produce the optimum value of the hyperplane can be used the following equation [27].

$$min\frac{1}{2}||\omega||^2 \quad (2)$$

$$y_i(wx_i + b) \geq 1, i = 1, \ldots, \lambda \quad (3)$$

Equation (3) is utilised as a way of optimising the value of $||\omega||^2$ which focuses on the boundary line $y_i(wx_i + b) \geq$. If the resulting output data is $y_i = +1$, then the boundary line becomes $(wx_i + b) \geq 1$ and vice versa if $y_i = -1$, then the value of the boundary line becomes $(wx_i + b) - 1$.

3.4 Hyperparameter Tunning

SVM models can be optimised for performance using *Grid Search* (GS) and *K-Cross Validation* (K-CV) methods. GS is a type of hyperparameter tuning used to optimise complex machine learning problems [28].

GS is a method used to select the best model from a combination of parameters of a machine learning algorithm by testing and validating each combination. The main goal of GS itself is to search and find the combination of model parameters with the best performance selected for the effectiveness of the prediction model [29]. GS is commonly associated with K-CV which will create an evaluation index for prediction and classification models [30]. K-CVcan repeat the training data and test data as many as krepetitions and*1/k*division ofthe test data [25]. The accuracy of k modelscan be obtained, and the performance ofK-CV prediction modelis evaluated based on the average accuracy score of*k*models. Subsequently, the parameters of the classifier are changed based on GS, andthe predictionaccuracyis recalculated. In general, the stages of K-CV can be seen in Figure 4 below.

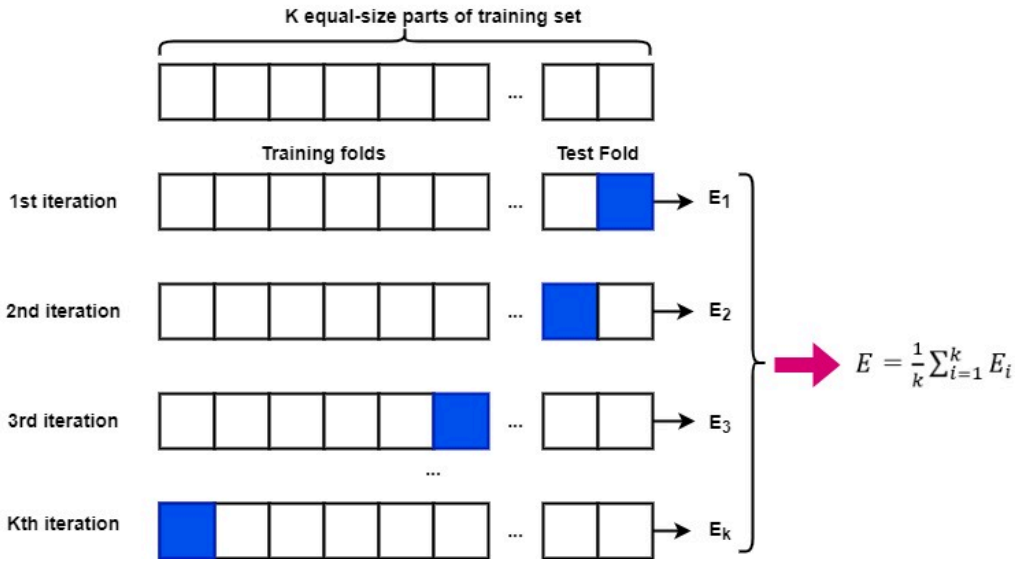

Fig. 4. K-Fold Cross-Validation Procedure

The model optimization process progresses further during the parameter selection stage, as illustrated in Figure 5. In this phase, Grid Search systematically explores various combinations of hyperparameters to identify the optimal configuration that enhances model performance. By exhaustively evaluating different parameter sets, GS ensures that the most effective combination is selected to maximize accuracy while maintaining computational efficiency. Each combination undergoes comparative analysis based on predefined evaluation metrics, allowing the model to achieve better generalization and robustness. This iterative process is crucial in fine-tuning the model, as selecting the right set of hyperparameters significantly impacts the learning dynamics and overall predictive capability. Furthermore, the optimization process not only improves

accuracy but also helps in minimizing overfitting and balancing the trade-off between model complexity and computational cost.

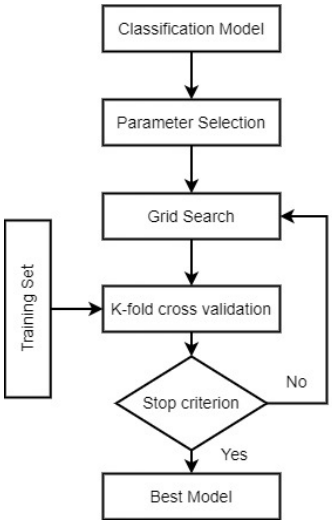

Fig. 5. K-Fold Cross-Validation and Grid Search

3.5 Model Evaluation

The model evaluation used is *confusion matrix*. Some types of criteria for measuring *confusion matrix* can be seen in Table 1 [31].

Table 1. Confusion Matrix

Class Data	Classed as Positive	Classed as Negative
+	True Positive (TP)	False Negative (FN)
-	False Negative (TN)	True Negative (FN)

Table 1 is the *confusion matrix* table. TP means the model correctly labelled the number of positive tuples. TN means the model correctly labelled the number of negative tuples. FP means the model incorrectly labelled the number of negative tuples. FN means the model mislabelled the number of positive tuples.

Accuracy is also called the performance measure of a prediction or classification model and is the percentage of correctly predicted data out of the total data [32]. Calculation of the accuracy value, Equation 4 can be used.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

In prediction models, the types of *performance metrics* commonly used consist of *precission*, *recall*, and *f1-score* [31]. *Precission* is the positive ratio or can be called the degree of reliability, which is the proportion of prediction results that have the correct positive label value to the total positive predict ions [33]. The equation for calculating the precision value can be seen in equation 5.

$$Precission = \frac{TP}{TP + FP} \tag{5}$$

Recall or commonly referred to as the *true positive rate* or also known as sensitivity. *Recall* is also the degree of reliability of the model in detecting positively labelled data correctly [33]. The equation for calculating the recall value can be seen in equation 6.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

F1-score becomes the summarised value of all precision results and recall calculations by making the harmonic mean [31]. The equation for calculating the *F1-score* value can be seen in equation 7.

$$F1 - Score = \frac{2\ x\ Precission\ x\ Recall}{Precission + Recall} \quad (7)$$

4. RESULT AND DISCUSSION

4.1 Pre-processing

In the pre-processing stage, there is a transformation of the form of the virus data. We changed the string dataset into a numeric form so that it can be easily processed optimally by the SVM model. The results of data transformation can be seen in Table 2.

Table 2. Data Transformation

Host	Size_Mb	Gc_Percent	Cds
0.0	0.073089	0.557107	0.060948
1.0	0.010327	0.667513	0.002257
0.0	0.066560	0.581218	0.057562
2.0	0.007142	0.560216	0.002257
3.0	0.014095	0.488579	0.007901
3.0	0.317524	0.569797	0.086907
0.0	0.079869	0.770305	0.056433
0.0	0.082039	0.601523	0.076749
0.0	0.162887	0.571066	0.129797
…	…	…	…

4.2 SVM Model

The SVM model is created by first separating the dataset into training data and test data. Comparison of the amount of training data and test data in the composition of 80% versus 20% [34]. This modeling uses the default kernel which is linear. The performance of the SVM model then produces an accuracy value of 80% which can be seen in Table 3.

Table 3. Result SVM Model Before Optimization

Precission	Recall	F1-score	Support
0.78	0.95	0.86	744
0.00	0.00	0.00	11
0.00	0.00	0.00	71
0.00	0.00	0.00	111
0.00	0.00	0.00	22
0.00	0.00	0.00	10
0.00	0.00	0.00	5
0.00	0.00	0.00	9
0.00	0.00	0.00	5
0.82	0.96	0.89	480
0.00	0.00	0.00	2
Accuracy : 80%			

4.3 Grid Search Optimization

Based on Table 3, the SVM model produces an accuracy value of 80%. This accuracy value is further optimised by using GS and K-CV *hyperparameter tuning*. The initial stage of model optimisation by first

determining the type of kernel that will be used by the SVM model. *Grid-shaped* parameters used include *linear, rbf, poly kernels* with degrees 2 and 3. K-CV is run on the parameters n_splits = 10 with n_repeats = 10 in 43.7 s. The result of using GS and K-CV is a parameter that can be seen below.

```
GridSearchCV(estimator=SVC(),
             param_grid={'C': [0.1, 1, 10, 100, 1000],
                         'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
                         'kernel': ['linear']},
             verbose=3)
```

The last step is to apply the best parameters to the SVM model. The result is an increase in accuracy to 84%. The performance of the SVM model after optimisation with GS and K-CV can be seen in Table 4.

Table 4. Result SVM Model After Optimization

Precission	Recall	F1-score	Support
0.91	0.94	0.92	744
0.00	0.00	0.00	11
0.43	0.13	0.20	71
0.43	0.65	0.52	111
0.00	0.00	0.00	22
0.00	0.00	0.00	10
0.00	0.00	0.00	5
0.00	0.00	0.00	9
0.00	0.00	0.00	5
0.90	0.96	0.93	480
0.00	0.00	0.00	2
Accuracy: 84%			

The evaluation results of the model that has been optimised with GS and K-CV can then be illustrated in the form of a confusion matrix graph. The graph shows the *true label* and *predicted label* data according to the *confusion matrix* size in Table 1. The confusion matrix graph can be seen in Figure 6.

Fig. 6. Confusion Matrix

5. CONCLUSION

Viruses and Bacteria continue to evolve alongside humans. Viruses are spreading too fast and causing a huge loss of life in the world. Viruses play an important role as dangerous pathogens that continue to spread various infectious diseases. Metegenomics is the application of large sequencing technology to genetic material obtained directly from one or more environmental samples, resulting in at least 50Mb random samples and multiple long sequences. It is important to identify the origin of the virus to prevent the spread of outbreaks. Understanding the biological side of these viruses and how they affect their ecosystems depends on knowing which hosts they infect. We can use metagenomic features derived from the viral genome to find out the virus host type. The traditional way of predicting virus hosts consumes a lot of time and effort in the process. Technology can be one of the solutions that can be used to predict the type of virus host. Popular technologies such as *machine learning*, which is one of the fields of *artificial intelligence*, can be applied to this. SVM algorithm is offered in this research as a virus host prediction model. The SVM model before optimisation produced an accuracy value of 80%. The initial stage of model optimisation by first determining the type of kernel that will be used by the SVM model. The *grid-shaped* parameters used include *linear*, *rbf*, *poly* kernels with degrees 2 and 3. K-CV is run on the parameters n_splits = 10 with n_repeats = 10 in 43.7 s. The result of this optimisation is an increase in the accuracy value of the SVM model resulting in an accuracy value of 80%. This SVM model still needs to be studied further in an effort to predict the virus host. This model can be compared with several other types of machine learning algorithms. The results of the comparison can be a benchmark for which algorithms can really be applied to virus host prediction.

REFERENCES

[1] Iis Setiawan Mangkunegara and P. Purwono, "Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV," in *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 2022, pp. 427–432, doi: 10.1109/CyberneticsCom55287.2022.9865624.

[2] L. Lu, S. Su, H. Yang, and S. Jiang, "Antivirals with common targets against highly pathogenic viruses," *Cell*, vol. 184, no. 6, pp. 1604–1620, 2021, doi: 10.1016/j.cell.2021.02.013.

[3] G. Abbas *et al.*, "Synthesis and investigation of anti-COVID19 ability of ferrocene Schiff base derivatives by quantum chemical and molecular docking," *J. Mol. Struct.*, vol. 1253, 2022, doi: 10.1016/j.molstruc.2021.132242.

[4] E. Sobhanie *et al.*, "Recent trends and advancements in electrochemiluminescence biosensors for human virus detection," *TrAC - Trends Anal. Chem.*, vol. 157, p. 116727, 2022, doi: 10.1016/j.trac.2022.116727.

[5] M. Sevvana, T. Klose, and M. G. Rossmann, "Principles of Virus Structure," *Encycl. Virol. (Fourth Ed.*, vol. 1, pp. 257–277, 2021, doi: https://doi.org/10.1016/B978-0-12-814515-9.00033-3.

[6] M. Ramazzotti and G. Bacci, "16S rRNA-Based Taxonomy Profiling in the Metagenomics Era," in *Metagenomics: Perspectives, Methods, and Applications*, Academic Press, 2018, pp. 103–119. DOI: https://doi.org/10.1016/B978-0-323-91631-8.00013-5

[7] M. E. Walker, J. B. Simpson, and M. R. Redinbo, "A structural metagenomics pipeline for examining the gut microbiome," *Curr. Opin. Struct. Biol.*, vol. 75, p. 102416, Aug. 2022, doi: 10.1016/J.SBI.2022.102416.

[8] Q. Hou, F. Pucci, F. Pan, F. Xue, M. Rooman, and Q. Feng, "Using metagenomic data to boost protein structure prediction and discovery," *Comput. Struct. Biotechnol. J.*, vol. 20, pp. 434–442, Jan. 2022, doi: 10.1016/J.CSBJ.2021.12.030.

[9] Y. Xu and D. Wojtczak, "Dive into machine learning algorithms for influenza virus host prediction with hemagglutinin sequences," *Biosystems*, vol. 220, p. 104740, Oct. 2022, doi: 10.1016/J.BIOSYSTEMS.2022.104740.

[10] F. H. Coutinho *et al.*, "RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content," *Patterns*, vol. 2, no. 7, p. 100274, Jul. 2021, doi: 10.1016/J.PATTER.2021.100274.

[11] Y. Yang *et al.*, "Reservoir hosts prediction for COVID-19 by hybrid transfer learning model," *J. Biomed. Inform.*, vol. 117, p. 103736, May 2021, doi: 10.1016/J.JBI.2021.103736.

[12] M. F. Drummond *et al.*, "Challenges of Health Technology Assessment in Pluralistic Healthcare Systems: An ISPOR Council Report," *Value Heal.*, vol. 25, no. 8, pp. 1257–1267, Aug. 2022, doi: 10.1016/J.JVAL.2022.02.006.

[13] L. Dey, S. Chakraborty, and A. Mukhopadhyay, "Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins," *Biomed. J.*, vol. 43, no. 5, pp. 438–450, Oct. 2020, doi: 10.1016/J.BJ.2020.08.003.

[14] W. Zhang, X. Gu, L. Tang, Y. Yin, D. Liu, and Y. Zhang, "Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge," *Gondwana Res.*, vol. 109, pp. 1–17, Sep. 2022, doi: 10.1016/J.GR.2022.03.015.

[15] H. Hassan *et al.*, "Supervised and weakly supervised deep learning models for COVID-19 CT diagnosis: A systematic review," *Comput. Methods Programs Biomed.*, vol. 218, p. 106731, May 2022, doi: 10.1016/J.CMPB.2022.106731.

[16] S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases," *Comput. Biol. Chem.*, vol. 97, p. 107619, Apr. 2022, doi: 10.1016/J.COMPBIOLCHEM.2021.107619.

[17] T. Yan, S. L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1292–1303, Aug. 2022, doi: 10.1016/J.JRMGE.2022.03.002.

[18] T. B. Alakus and I. Turkoglu, "Prediction of viral-host interactions of COVID-19 by computational methods," *Chemom. Intell. Lab. Syst.*, vol. 228, p. 104622, Sep. 2022, doi: 10.1016/J.CHEMOLAB.2022.104622.

[19] D. D. Holcomb *et al.*, "Protocol to identify host-viral protein interactions between coagulation-related proteins and their genetic variants with SARS-CoV-2 proteins," *STAR Protoc.*, vol. 3, no. 3, p. 101648, Sep. 2022, doi: 10.1016/J.XPRO.2022.101648.

[20] J. K. Das, S. Chakraborty, and S. Roy, "A scheme for inferring viral-host associations based on codon usage patterns identifies the most affected signaling pathways during COVID-19," *J. Biomed. Inform.*, vol. 118, p. 103801, Jun. 2021, doi: 10.1016/J.JBI.2021.103801.

[21] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Comput. Methods Programs Biomed.*, vol. 220, p. 106773, Jun. 2022, doi: 10.1016/J.CMPB.2022.106773.

[22] L. M. Matos, J. Azevedo, A. Matta, A. Pilastri, P. Cortez, and R. Mendes, "Categorical Attribute traNsformation Environment (CANE): A python module for categorical to numeric data preprocessing," *Softw. Impacts*, vol. 13, p. 100359, Aug. 2022, doi: 10.1016/J.SIMPA.2022.100359.

[23] J. Alcaraz, M. Labbé, and M. Landete, "Support Vector Machine with feature selection: A multiobjective approach," *Expert Syst. Appl.*, vol. 204, no. April, p. 117485, 2022, doi: 10.1016/j.eswa.2022.117485.

[24] M. Marchetti, L. Fongaro, A. Bulgheroni, M. Wallenius, and K. Mayer, "Classification of uranium ore concentrates applying support vector machine to spectrophotometric and textural features," *Appl. Geochemistry*, vol. 146, p. 105443, 2022, doi: https://doi.org/10.1016/j.apgeochem.2022.105443.

[25] K. R. Singh, K. P. Neethu, K. Madhurekaa, A. Harita, and P. Mohan, "Parallel SVM model for forest fire prediction," *Soft Comput. Lett.*, vol. 3, no. June, p. 100014, 2021, doi: 10.1016/j.socl.2021.100014.

[26] R. Umar, I. Riadi, and Purwono, "Perbandingan Metode SVM, RF dan SGD untuk Penentuan Model Klasifikasi Kinerja Programmer pada Aktivitas Media Sosial," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 329–335, 2020. doi: https://doi.org/10.29207/resti.v4i2.1770

[27] D. Maulina and R. Sagara, "Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document Frequency," *J. Mantik Penusa*, vol. 2, no. 1, pp. 35–40, 2018. Web: https://repository.amikom.ac.id/files/2017/Publikasi_14.11.8030.pdf

[28] T. Yan, S. L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1292–1303, 2022, doi: 10.1016/j.jrmge.2022.03.002.

[29] G. S. K. Ranjan, A. Kumar Verma, and S. Radhika, "K-Nearest Neighbors and Grid Search CV Based Real Time Fault Monitoring System for Industries," in *2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019*, 2019, no. March, doi: 10.1109/I2CT45611.2019.9033691.

[30] T. Yan, S. L. Shen, A. Zhou, and X.-S. Chen, "Prediction of geological characteristics from shield operational parameters using integrating grid search and K-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotech. Eng.*, p. 100310, 2022, doi: https://doi.org/10.1016/j.jrmge.2022.03.002.

[31] X. Xiong, S. Hu, D. Sun, S. Hao, H. Li, and G. Lin, "Detection of false data injection attack in power information physical system based on SVM–GAB algorithm," *Energy Reports*, vol. 8, pp. 1156–1164, 2022, doi: 10.1016/j.egyr.2022.02.290.

[32] S. Katoch, V. Singh, and U. S. Tiwary, "Indian Sign Language Recognition System using SURF with SVM and CNN," *Array*, p. 100141, 2022, doi: https://doi.org/10.1016/j.array.2022.100141.

[33] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

[34] P. Purwono, A. Ma'arif, I. S. Mangku Negara, W. Rahmaniar, and J. Rahmawan, "Linkage Detection of Features that Cause Stroke using Feyn Qlattice Machine Learning Model," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 7, no. 3, p. 423, 2021, doi: 10.26555/jiteki.v7i3.22237.