International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Genome-Based Drug Repurposing: Identifying Potential Targets Using FASTA Sequences and Machine Learning Ahmed Al

Shahab[1], Dr. Vaishali A. Chavan[2]

[1,2]Dr. G. Y. Pathrikar College of Computer Science and IT, MGM University, Chhatrapati Sambhajinagar

Abstract

This study uses genomic data to explore machine learning techniques for drug repurposing in viral diseases. This research aims to develop classification models that utilize FASTA protein sequence data to find similar genomes and, based on identified similar genome could help in drug repurposing for viral diseases. To develop this model, we explored three machine learning models Decision Tree, Random Forest , and K-Nearest Neighbours. These models were implemented and assessed.

The study used genetic information from seven viral disease families obtained from the National Center for Biotechnology Information (NCBI). Data preprocessing involved cleaning and encoding the FASTA protein sequences.

These three models were implemented to identify similar genomes for targeted viral disease, and tested targeted HMPV Viral disease on all three models and found Rotavirus as the closest match.

The Random Forest model shows the best performance with an accuracy of 98.79% and F1 Score of 0.988.

Keywords: Machine Learning, Genome Predictor, FASTA sequence, Drug Repurposing

1. Introduction

Drug Repurposing is the most powerful method for handling new diseases with existing and approved drugs by the concerned authorities.

This practice dates back to 1897 when Felix Hoffman a german chemist who modified the salicylic acid to develop acetylsalicylic acid, popularly known today as aspirin, which was originally developed to reduce pain but later aspirin proved to help prevent heart attacks and strokes [1].

Another best example is ”Thalidomide” , Thalidomide initially developed in the 1950s to fight for morningsickness, then it was subsequently known for its effectiveness in treating leprosy and multiple myeloma [2].

A recent example of drug repurposing is Remdesivir and Hydroxychloroquine, Both drugs were originally developed for different purposes but were successfully adapted for the treatment of COVID-19, this demonstrates the tremendous potential of drug repurposing, especially in pan-demic situations.

The following are some examples of drug repurposing.

Drug	Drug Developed For	Drug Repurposed For
Aspirin	Pain relief	Cardiovascular disease
Thalidomide	Morning sickness	Multiple myeloma
Remdesivir	Ebola	COVID-19
Sildenafil	Hypertension	Erectile dysfunction
Dapsone	Leprosy	Malaria
Minoxidil	Hypertension	Hair loss
osmidomycin	Urinary tract infections	Antimalarial

Table 1: Examples of Drug Repurposing

Drug discovery could be time- and cost-consuming and involves many clinical trials, according to study new drug development cost estimates ranging from 314*millionto*2.8 billion [3] and estimates the duration of 12–15 years [4].

This study investigates different supervised machine learning techniques, focusing on Random Forest, Decision Tree, and K-Nearest Neighbors (K-NN), to identify the most accurate predictive ML model for identifying the most similar viral genomes to a targeted genome.

We anaylised each model using performance metrics such as accuracy, precision, recall, and F1-score to find out the best-performing algorithm.

We utilized viral genome data in FASTA format acquire from the National Center for Biotech-nology Information (NCBI), generally used public repository that gives access to biomedical and genomic information.

The FASTA format is a widely used text format that represents sequences of nucleotides or proteins.

To prepare the FASTA dataset for machine learning, we performed data preprocessing steps that included converting nucleotide/protein sequences into numerical format using one hot encod-ing and for removing duplicate entries used custom python program to ensure data high quality and stuitabilty with selected machine learning models. After preprocessing, the cleaned dataset was used to train and test the selected machine learning models. Additionally, we compared our find-ings with results from some most related studies to validate our method and mentioned potential improvements.

Related Works

As drug repurposing is popular and ancient method and less expensive and fastest way to treat diseases there are some existing work has been done with computational and machine learning methods, this method becoming popular nowadays because its easy to perform experiment and large authentic database available in desire format or even with slightingly data process can achieve desire results, specially after COVID-19 & ai popularity researchers around the world using computational & machine learning approach.

A rare orphan disease which is a rare medical condition that affects a small percentage of people and has limited research and funding, [5] used . Interactions between drugs, genes, enzymes, and targets are analyzed to identify potential candidate treatments for rare orphan diseases. The paper shows remarkable findings regarding the use of machine learning (ML) techniques for drug repurposing, specifically for rare orphan diseases. The study analyzed various machine learning models to find out their effectiveness in

predicting potential drug candidates.

Out of all tested models Random Forest came up as the most effective, and outperforming all other models over four evaluation metrics, also with the highest Area Under the Receiver Operating Characteristic (AUROC) value of 0.827, which shows strong predictive performance [5].

The Random Forest & the K-Nearest-Neighbors (KNN) model also showed promising results, by achieving higher AUROC values as compared to the other tested models, which had AUROC values close to 0.5. This shows that Random Forest and KNN are the most reliable models for particular this application [5].

The method taken in this paper involved combining heterogeneous information from multiple sources such as drug-target interactions, drug-side effects, drug-enzymes, and drug-gene interactions.

This thorough data integration is very important for find out the potential drug candidates for rare diseases [5].

The authors of this research suggested that combining extra data sources, such as protein-protein interactions and 3-D chemical structures of drugs, could improve the accuracy of their predictions.

They also proposed exploring multiple clustering methods and increasing the dataset used in their experiments to improve the robustness of their findings [5].

Overall, the results shows the potential of machine learning in drug repurposing, specifically for rare orphan diseases as title suggest.

The successful finding of existing drugs that could help & aid in treatments for these diseases could remarkable accelerate the drug discovery process, which making it more cost-effective and efficient.

The findings shows the effectiveness of machine learning models, specially the Random Forest, in this application of drug repurposing, also outlining path for future research to improve predictive capabilities of the ML models [5].

The Ranking-based k-Nearest Neighbor (Re-KNN) [6] approach for drug repurposing in cardio-vascular diseases integrates the traditional KNN algorithm with the Ranking SVM (Support Vector Machine) algorithm to obtain neighbors that provide more precise and dependable outcomes. [6].

The Re-KNN approach achieved an AUC of 0.928 and an AUPR of 0.588, shows the heighest predictive capability for drug repositioning specifically cardiovascular diseases if compared to other methods like WNN-GIP and Super-Target [6]. Findings indicated that it recognized 300 new rela-tionships between 117 drugs, including 35 cardiovascular disorders, demonstrating its potential for discovering novel drug uses [6].

This approach prioritized accuracy and sensitivity, by achieving a sensitivity of 0.388 and accu-racy of 0.983, which shows its effectiveness in managing imbalanced datasets [6].

The 10-fold cross-validation method was used to ensure the reliability of the results, then vali-dating the robustness of the Re-KNN method [6].

Some predicted drug disease associations were authenticate by existing literature, which sug-gesting the method's practical capability in drug discovery [6].

TTwo-stage prediction with machine learning , in first stage diseases are clustered by gene expression based on pattern and similarities and in second stage assessed drug efficacy by the reversibility of abnormal gene expression [7].

The total of 171 genes were identified which directly affected by nine compounds with 123 of these genes common to disease-specific genes for inclusion body myositis (IBM) acquire through GEO2R analysis [7].

According to IBM and polymyositis (PM), nine medications have been identified; furthermore four drugs are linked to IBM and dermatomyositis (DM) Note worth that this study UMAP for dimensionality reduction of gene expression data , along with followed by k-means clustering which successfully classified 31 diseases into distinct groups [7].
The DRIAD approach was highlighted for its effectiveness in correlating drug-induced changes with molecular markers of Alzheimer's disease, despite the fact that it has limitations in wider applications [7]. These findings underscore the potential of a two-stage prediction approach in drug repurposing, which may inform future drug discovery and clinical trials across various human diseases.

Data

The pivotal stage of any research, especially in the machine learning approach, is data selection. In this study, genetic information on viral diseases was obtained from NCBI database, which is funded by the United States government and houses a vast database of genomes, families, and variations [8]. Specifically, data from seven families of viral diseases was utilized for model training.

SR No.	Viral Disease Family	Total Variations
1.	Checkenpox(aka varicella-zoster)	21,987
2.	Chikungunya	10,022
3.	Dengue	48,477
4.	Diphtheria	1
5.	Ebola	30,532
6.	Rotavirus	1,33,390
7.	Zika	2,241

Table 2: Viral Diseases Families and their variations

1.1 Data Type

There are couple of formats available in NCBI portal, including csv, json, and xml,and the sequence type we used is protein sequence in FASTA format [9].
FASTA is a format based on text utilized for the storage and comparison of nucleotide or amino acid sequences [9], it employs a heuristic word method to align sequences, and these sequences can be stored as single or multiple sequences, where each letter represents an amino acid or nucleic acid [9].

Amino Acid	3-letter Code	1-letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G

Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
Asparagine or Aspartic acid	Asx	B
Glutamine or Glutamic acid	Glx	Z

Table 3: Standard amino acid codes used in protein sequence representations. Compiled and adapted from public-domain sources such as NCBI and UniProt [9].

Example FASTA Sequence for Human Hepatitis:
LPWYSYLYAVSGALDGLGDKTDSTFGLVSIQIANYNHSDEYLSFSCYLSVTEQSEFYFPR
APLNSNAMLSTESMMSRIAAGDLESSVDDPRSEEDRRFESHIECRKPYKELRLEVGKSRL
KYAQEELSNEVFPPPKMKGLFSQAKISLFYTEEHEIMKFSWRGVTADTRALRRFGFSMA
AGRSVWTLEMDAGVLTGRLVRLNDEK

Data Pre-processing

The most machine learning models can only handle comfortably numerical data, and characters such as ”A”, ”R”, or ”N” must be transformed into numerical values such as 0 or 1 prior to their utilization in the models. Consequently, the obtained FASTA protein sequence data is in textual form, necessitating additional data preprocessing and encoding.

1.2 Data Cleaning

The collected viral genome data including all variations related to the viral genome family needed further cleaning and processing as there are duplication of sequences present in the obtained FASTA data from [8]. To address this challege, a Python script was employed to remove the duplicates before moving on to the encoding stage. This preprocessing step not only reduced the size of the fasta squences data but also will led to a notable reduction in both data encoding and training times, the intensive data processing involved demands substantial computational resources, thereby increasing the overall cost of the experiment, note the collected data is formatted in CSV, which is a lightweight and universally supported format across various programming languages.

1.3 Data Encoding

After cleaning the data, data encoding is required as most machine learning models only process numeric and binary data.

One-Hot Encoding: One-Hot Encoding is a specific type of encoding used for categorical data, especially when there is no ordinal relationship among categories. It creates a binary matrix representation of the data (e.g 10100100, 10100101, 10100111).

Following are example of encoding of FASTA letter into binary

Amino Acid	Name	One-Hot Encoding
A	Alanine	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
C	Cysteine	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
D	Aspartic Acid	[0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
E	Glutamic Acid	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
F	Phenylalanine	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
G	Glycine	[0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
H	Histidine	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
I	Isoleucine	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
K	Lysine	[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
L	Leucine	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
M	Methionine	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
N	Asparagine	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0]

Table 4: One-Hot Encoding of Amino Acids

2. Experimental Results

The classification performance of different machine learning models, including K-Nearest Neighbors (KNN), Decision Tree, and Random Forest, using the evaluation metrics such as Precision, Re-call, F1-Score, and Support, and these models were implemented to the dataset to evaluate their effectiveness in predicting the targeted classes.

1.4 K-Nearest Neighbors (KNN)

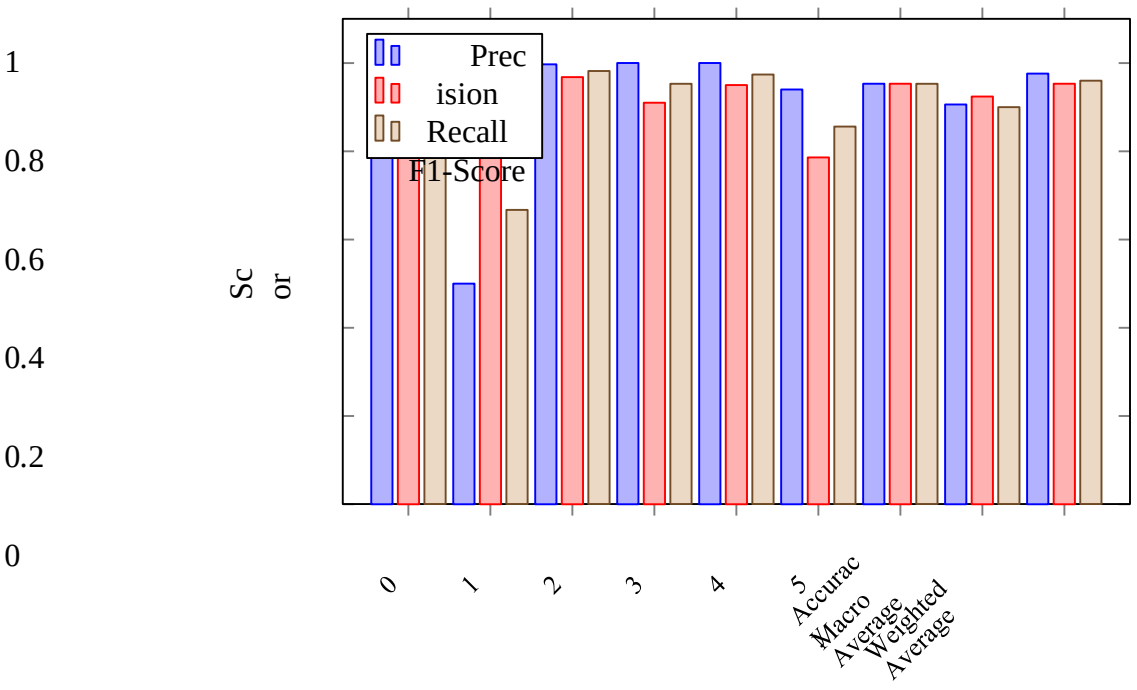
The K-Nearest Neighbors classifier shows moderate performance with an overall accuracy of 95.28%, and It performed well for most classes, by achieving high precision and recall values, well however, class 1 showed lower precision (0.500), which indicating mis-classification issues due to overlapping features, but despite this the model maintained a weighted average F1-score of 0.959, which shows its robustness for balanced datasets.

1.5 K-Nearest Neighbors (KNN) Evaluation Metrics

The K-Nearest Neighbors (KNN) classifier was asses based on precision, recall, F1-Score, and sup-port for each and every class. The classification performance is summarized in Table 5.

Label	Precision	Recall	F1-Score	Support
Class 0	1.000	0.933	0.965	400
Class 1	0.500	1.000	0.667	956
Class 2	0.997	0.968	0.982	4989
Class 3	1.000	0.910	0.953	681
Class 4	1.000	0.950	0.974	13505
Class 5	0.940	0.786	0.856	280
Accuracy	0.953			
Macro Average	0.906	0.924	0.900	20811
Weighted Average	0.976	0.953	0.960	20811

Table 5: Evaluation metrics for the K-Nearest Neighbors (KNN) classifier, generated using Scikit-learn's classification_report() function. Metrics include precision, recall, F1-score, and support for each class.



Metric

Figure 1: K-Nearest Neighbors (KNN) Classification Performance

1.6 Decision Tree

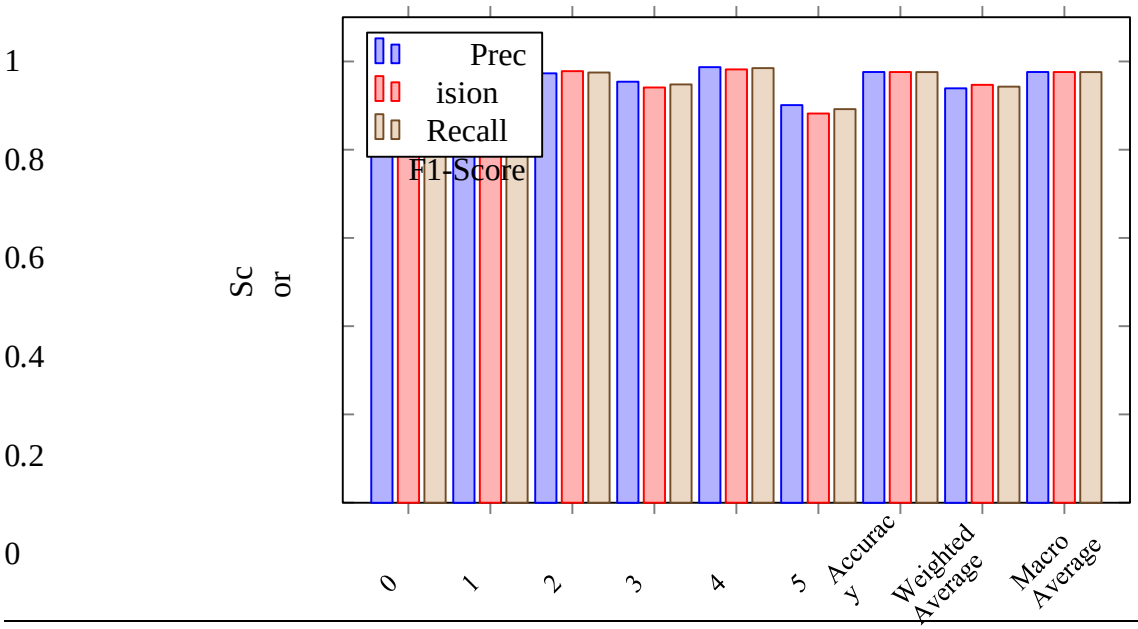
The Decision Tree classifier out-performed KNN with an accuracy of 97.61%, and tt exhibited balanced performance across all classes, with maintaining high precision and recall scores, macro average precision (0.939) and recall (0.947) indicate that the model effectively distinguished between different classes. The Decision Tree's ability to capture complex patterns contributed to its improved results compared to KNN.

1.7 Decision Tree Evaluation Metrics

The Decision Tree classifier were evaluated based on precision, recall, F1 score, and support for each class. The classification performance is summarized in Table **??**.

Class Label	**Precision Score**	**Recall Rate**	**F1 Measure**	**Instances**
Class 0	0.9059	0.9625	0.9333	400
Class 1	0.9134	0.9372	0.9251	956
Class 2	0.9727	0.9776	0.9751	4989
Class 3	0.9539	0.9413	0.9475	681
Class 4	0.9868	0.9825	0.9846	13505
Class 5	0.9015	0.8821	0.8917	280
Overall Accuracy	0.9761			
Macro Average	0.9390	0.9472	0.9429	20811
Weighted Average	0.9763	0.9761	0.9762	20811

Table 6: Evaluation metrics for the Decision Tree classifier, generated using Scikit-learn's classification_report() function. Metrics include precision, recall, F1-score, and support for each class.



Metric

Figure 2: Decision Tree Classification Performance

1.8 Random Forest

The Random Forest classifier got the best performance among all models, with an accuracy of 98.79%, and it provided stable and high precision, recall, and F1-scores across all classe, macro average precision (0.995) and weighted average F1-score (0.988) which shows it's robustness and capability to generalize well all over different classes, and the ensemble nature of Random Forest helped in reducing overfitting and improving overall classification performance.

1.9 Random Forest Evaluation Metrics

The Random Forest classifier was evaluated based on precision, recall, F1 score, and support for each class. The classification performance is summarized in Table 7.

Class Label	**Precision**	**Recall**	**F1-Score**	**Support**
Class 0	1.000	0.963	0.981	400
Class 1	1.000	0.932	0.965	956
Class 2	1.000	0.980	0.990	4989
Class 3	1.000	0.947	0.973	681
Class 4	0.982	1.000	0.991	13505
Class 5	0.988	0.882	0.932	280
Accuracy	0.988			
Macro Average	0.995	0.951	0.972	20811
Weighted Average	0.988	0.988	0.988	20811

Table 7: Performance metrics for the Random Forest classifier, computed using Scikit-learn's classification_report() function. The table summarizes precision, recall, F1-score, and sup-port for each class label.

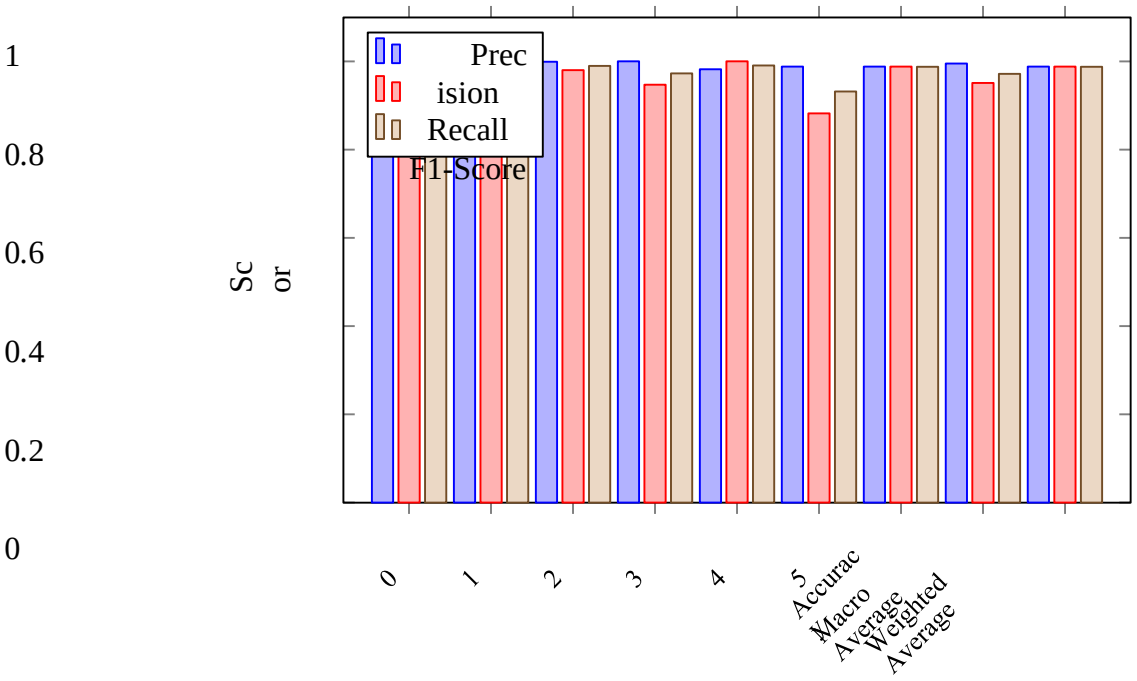


Figure 3: Random Forest Classification Performance

3. Prediction

The HMPV genome FASTA protein sequence was used as input to identify similar genomes for potential drug repurposing. All three predictive models consistently identified Rotavirus as the closest match.
Human metapneumovirus (HMPV) symptoms include headache, exhaustion, coughing, runny nose or nasal congestion, sore throat, and fever [10] [11].
Rotavirus symptoms include fever, loss of appetite, runny nose, watery diarrhoea, vomiting, stomach pain, and coughing [12] [13].
When the COVID-19 genome FASTA protein sequence was examined, the models predicted Rotavirus as the closest match, to further validate the model's reliability a Dengue FASTA sequence was given as input, and the model correctly identified Dengue, a virus already present in the training dataset.
It should be noted that this study does not offer any medical advice or advocate the direct use of any particular medication, and the application of these findings is contingent upon validation by regulatory bodies and medical professionals. These findings imply that antiviral medications used to treat Rotavirus may have potential for repurposing in the treatment of HMPV.

4. Discussion

Classifier	Precision	Recall	F1-Score
KNN (K-Nearest Neighbors)	0.906	0.924	0.899

Decision Tree	0.939	0.947	0.943
Random Forest	0.995	0.951	0.972

Table 8: Aggregated performance comparison across classification models based on macro-averaged metrics.

The Random Forest shows the highest performance in all metrics as showed in Table 8 , that includes accuracy, precision, recall, and F1-score, also Its aggregation learning approach effectively take complex patterns in the data, which resulted in higher generalization and robustness, as a result, it is the most trustworthy classification model in this research.

Although the Decision Tree model's precision, recall, and F1-scores were all comparatively high, it did not perform as well as Random Forest. Nevertheless, due to its interpretability and speed of computation, the Decision Tree remains a valuable model.

On the other hand, the K-Nearest Neighbors (KNN) algorithm came-out the weakest perfor-mance among the three models. While KNN can be very effective for some other classification tasks, but it struggled with class imbalance and computational efficiency, specially as the dataset size increased. Also Its dependency on distance based similarity made it ineffective in handling complex data distributions.

Despite slight differences in the evaluation metrics of all three models gave same prediction results. KNN & Decision Tree although slightly behind in comparison with Random Forest,but still gives consistent and accurate predictions. This suggests that while metric variations exist, all models effectively classify the data with high accuracy.

5. Conclusion and Future Work

Drug repurposing using machine learning presents a promising approach to accelerating drug dis-covery, especially in pandemic-like situations, by identifying genomic similarities with the genome of new viral diseases. This study explores a machine learning-based genomic similarity prediction method to assist medical professionals in identifying existing drugs for targeted viral diseases. After evaluating all three models, Random Forest came-out as top performer and could be utilized in future similar studies.

However, it is important to emphasize that this study does not provide medical advice or endorse the direct use of any specific drug. These findings is subject to validation by medical professionals and regulatory authorities.

We believe that this work will open up new avenues for drug repurposing by predicting similar viral diseases for a targeted new viral genome using existing viral disease genomic data. This study is limited to protein sequences, but combining nucleotide and protein sequences could yield more accurate and promising results.

References

1. J. Miner and A. Hoffhines, "The discovery of aspirin's antithrombotic effects," Tex Heart Inst J, vol. 34, no. 2, pp. 179–186, 2007.
2. J. H. Kim and A. R. Scialli, "Thalidomide: the tragedy of birth defects and the effective treatment of disease," Toxicol Sci, vol. 122, no. 1, pp. 1–6, Jul 2011, erratum in: Toxicol Sci. 2012

Feb;125(2):613. Epub 2011 Apr 19.
3. O. J. Wouters, M. Mckee, and J. Luyten, “Estimated research and development investment needed to bring a new medicine to market, 2009-2018,” JAMA, vol. 323, no. 9, p. 844, 2020.
4. J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, “Principles of early drug discovery,”
5. British Journal of Pharmacology, vol. 162, no. 6, pp. 1239–1249, March 2011.
6. R. Manicavasaga, P. B. Lamichhane, P. Kandel, and D. A. Talbert, “Drug repurposing for rare orphan diseases using machine learning techniques,” The International FLAIRS Conference Proceedings, vol. 35, 2022.
7. X. Tian, M. Xin, J. Luo, and Z. Jiang, “Using the ranking-based knn approach for drug repositioning based on multiple information,” in Proceedings of the Conference, vol. 9771, Aug 2016, pp. 317–327.
8. Y. Cong, M. Shintani, F. Imanari, N. Osada, and T. Endo, “A new approach to drug repurpos-ing with two-stage prediction, machine learning, and unsupervised clustering of gene expres-sion,” OMICS, vol. 26, no. 6, pp. 339–347, Jun 2022, erratum in: OMICS. 2023 Aug;27(8):402-
9. 405. doi: 10.1089/omi.2023.29093.correx. Epub 2022 Jun 3.
10. National Center for Biotechnology Information, “Ncbi - national center for biotechnology information,” 2024, accessed: 2024-03-06. [Online]. Available: https://www.ncbi.nlm.nih.gov/
11. ——, “Fasta format - genbank overview,” 2024, accessed: 2024-03-06. [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/fastaformat/
12. W. H. Organization, “Who advisory on trends of acute respiratory infection, including human metapneumovirus,” 2025. [Online]. Available: https://www.who.int
13. M. Binnicker, “What you need to know about hmpv,” Mayo Clinic News Network, 2025. [Online]. Available: https://newsnetwork.mayoclinic.org/discussion/video-what-you-need-to-know-about-hmpv/
14. World Health Organization, “Rotavirus infections,” WHO, 2024. [Online]. Available: https://www.who.int/health-topics/rotavirus
15. ——, “Rotavirus: What you need to know,” WHO Factsheet, 2024. [Online]. Available: https://www.who.int/docs/librariesprovider2/default-document-library/factsheet$_r$otavirus.pdf ?sfvrsn = 53480def$_2$download = true