EasyChair Preprint

№ 10213

Covid19 Prediction Using Machine Learning Algorithms: a Comparative Study

Imane Aitouhanni

May 18, 2023

Covid19 Prediction using Machine Learning Algorithms: A Comparative study

AITOUHANNI Imane
SSL Lab ENSIAS
Mohammed V University in Rabat
Rabat, Morocco
imane_aitouhanni@um5.ac.ma

Abstract

Artificial intelligence has been used in many ways to combat the COVID-19 pandemic caused by the SARS-CoV-2 virus. One such approach is to use machine learning algorithms to predict different virus variants. By analyzing large volumes of genomic data, machine learning algorithms can identify patterns and make predictions about the behavior and characteristics of different viral strains.

This article presents a comparative study aimed at identifying the most effective machine learning algorithm for developing an approach based on artificial intelligence and machine learning methods to combat the virus.

The study evaluated a total of forty algorithms, including Decision Tree, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, and Artificial Neural Network (ANN), among others. we analyzed each algorithm for its performance using criteria such as RMSE, R-Squared and time taken, using data from the Chembl database.

To do this, we using the latest information from biological publications and medical reports in order to carefully select inputs and targets.

Keywords—Covid-19, Machine learning, Artificial intelligence, Pandemic

I. INTRODUCTION

The SARSCov-2 virus and its concerning variants have caused significant restrictions on human mobility for more than three years. This virus, which has undergone numerous mutations, is highly contagious and can be easily transmitted from person to person. Coronaviruses are a diverse group of viruses that can cause severe respiratory illness in humans. While many individuals infected with the virus experience mild to moderate symptoms and recover without requiring medical intervention, others, especially the elderly or those with pre-existing health conditions like chronic illnesses or cancer, may become severely ill and require hospitalization.

It sounds like we describing a typical workflow for developing a machine learning model to predict the pIC50 values of compounds based on their molecular descriptors. Here's a brief overview of the steps involved:
Data collection and pre-processing: This involves gathering the relevant data from the ChEMBL database [1] and cleaning it up as needed. This could involve removing duplicates, filtering out irrelevant data, and converting the data into a format that is suitable for analysis.
Exploratory analysis: This involves using descriptive statistics and data visualization techniques to gain a better understanding of the data and identify any patterns or trends that may be present. Box and scatter plots are commonly used to compare the distribution of molecular descriptors for active and inactive compounds.

Descriptor calculation using PADEL-Descriptor: PADEL-Descriptor is a software package that can be used to calculate a large number of molecular descriptors for a given set of compounds.

Dataset preparation: The X and Y dataframes are prepared using the calculated molecular descriptors as the X features and the pIC50 values as the Y variable.

Machine learning model development: A variety of machine learning algorithms can be used to develop regression models that predict pIC50 values based on the molecular descriptors. The performance of different models can be compared using metrics such as mean squared error or R-squared.

Model validation and optimization: The trained models are evaluated using cross-validation techniques to ensure that they generalize well to new data. Hyperparameter optimization can be used to fine-tune the models for optimal performance.

Overall, this workflow involves a combination of data cleaning, feature engineering, exploratory analysis, and machine learning to compare the predictives models for compound activity.

The remaining parts of the article are structured as follows: Section II provides a concise summary of articles that tackled the same issue. In Section III, a background study is

presented. Section IV details the data collection process, including clustering and data preprocessing to ensure the quality of the selected data. Section V features the COMPARING REGRESSORS section, where a comparison of regressors is conducted based on three criteria: R-squared, RMSE, and time consumption. Lastly, the article concludes with Section VI.

II. Related work

The analysis of prior works in an article serves several purposes. Firstly, it helps to identify the existing research in the field and to understand the current state of knowledge on the topic. This is important for situating the new study within the broader context of the field and for identifying any gaps or inconsistencies in the current literature. Secondly, it allows to build on the work of others and to incorporate the best practices and techniques from prior research into their own study. Finally, it helps to provide readers with a comprehensive understanding of the topic, including the historical development of the field and the current state of knowledge.

The article [2] proposes a taxonomy tree to investigate disease-confronting methods and their positive and negative effects, and presents a case study and systematic literature review to evaluate the effectiveness of the proposed methods against the COVID-19 outbreak from December 2019 to July 2020. The experimental results and observations demonstrate the impact of the proposed medical, prevention, detection, prediction, and social methods for controlling the spread of the virus. The article concludes that the case study can provide people with more information about the disease, its impact on human health, and effective self-care methods and therapies.

The aim of this article [3] was to create a comprehensive machine learning framework to evaluate the predictive value of these symptoms, along with others, in identifying COVID-19 infections. To achieve this aim, they conducted a multicenter case-control study, wherein suspected cases were tested with real-time reverse transcription polymerase chain reaction and asked to report the presence and severity of their symptoms using visual analog scales. The collected data was then fed into machine learning algorithms, which were trained and tested using a 50-fold cross-validation approach. This study involved a total of 777 patients, and we found that loss of smell and taste had the highest odds ratios of 6.21 and 2.42, respectively, for COVID-19 positivity. the machine learning models achieved an average accuracy of 80%, with a sensitivity of 82% and a specificity of 78% when using VAS to predict a COVID-19 diagnosis. This study demonstrates the effectiveness of machine learning models in predicting COVID-19 infections, and highlights the importance of loss of smell and taste as key indicators of the disease.

A study was conducted in [4] comparing five standard machine learning models, including Linear Regression (LR), Decision Tree, Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest, and Support Vector Machine (SVM), to predict the potentially harmful variables of COVID-19. Each of these models made forecasts in three categories: total active cases, total deaths, and total recoveries for the upcoming five days.

III. BACKGROUND STUDY

A. *SARS-CoV-2 Genome*

The virus known as SARS-CoV-2 represents an acronym for "severe acute respiratory syndrome coronavirus 2" and is the cause of Covid-19 disease. Classified as a positive-sense single-stranded RNA virus of group IV under the Baltimore classification, this virus belongs to the betacoronavirus genus, which also encompasses viruses like SARS-CoV-1 and MERS-CoV. SARS-CoV-2 is a novel strain of the SARSr-CoV coronavirus.

The source of SARS-CoV-2 remains unresolved, but it is believed that the animal reservoir for sarbecoviruses is found in Asian bats of the Rhinolophus genus, commonly known as horseshoe bats. It is possible that the virus adapted to humans through direct transmission from bats, but it is also plausible that an intermediate host was involved, which has yet to be identified. Alternatively, there is speculation that the virus may have resulted from laboratory gain-of-function experiments.

The SARS-CoV-2 genome comprises a single-stranded RNA consisting of 29,903 nucleotides. In terms of nucleotide homology, SARS-CoV-2 shares 79.5% similarity with SARS-CoV and 50% with MERS-CoV.

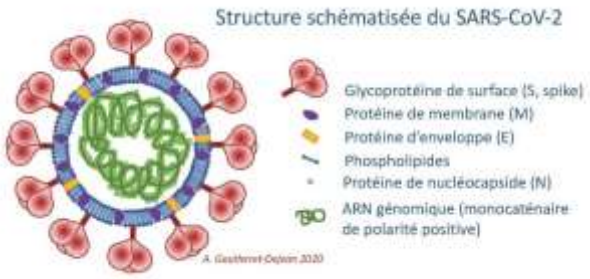

Figure 1 : Structure du SARS-CoV-2

Source : https://www.eurofins-biomnis.com/wp-content/uploads/2021/03/SARSCOV2-1-870x447.png

Various variants of the SARS-CoV-2 virus have emerged since late 2020, and they are named after the Greek alphabet. One of these variants, known as the Alpha variant or B.1.1.7, was first detected in the United States and rapidly spread throughout France since its emergence in December 2020, becoming the predominant strain by March 2021. Another variant, the Beta variant or B.1.351, was first reported in South Africa in early 2021 and has a distinct profile from the Alpha variant. Similarly, the Gamma variant or P.1, which surfaced for the first time in Brazil, was also circulating during the first half of 2021, although to a lesser degree than the Beta variant. The Delta variant emerged in May 2021 and quickly replaced previous variants, becoming the dominant strain in France by July 2021 and accounting for over 99% of circulating variants from August 2021. The most recent variant, the Omicron variant, appeared at the end of November 2021 and its distribution is currently on the rise.[5] [6] [7]

B. *Machine learning algorithm*

Machine learning algorithms are computational models that enable machines to learn from data and make predictions or decisions without being explicitly programmed. There are various types of machine learning algorithms, each designed to tackle different types of problems and data: [8] [9] [10] [11]

- Supervised Learning Algorithms: These algorithms learn from labeled training data, where each input data point is associated with a corresponding target label. Examples include:
  - Decision Trees: These algorithms create a tree-like model to make decisions based on feature values.
  - Support Vector Machines (SVM): SVM finds a hyperplane that best separates data points into different classes.
  - Random Forest: It combines multiple decision trees to make predictions.
  - Naive Bayes: This algorithm applies Bayes' theorem with the assumption of independence between features.
  - Neural Networks: Neural networks consist of interconnected nodes (neurons) that mimic the human brain's structure.
  - Unsupervised Learning Algorithms: These algorithms learn patterns and relationships from unlabeled data without explicit target labels. Examples include:
  - Clustering Algorithms: K-means, DBSCAN, and hierarchical clustering algorithms group similar data points together.
  - Dimensionality Reduction Algorithms: Principal Component Analysis (PCA) and t-SNE reduce the dimensionality of data while retaining its meaningful information.
- Reinforcement Learning Algorithms: These algorithms learn through interactions with an environment, receiving feedback in the form of rewards or punishments. Examples include:
  - Q-Learning: This algorithm learns an optimal policy by estimating the values of state-action pairs.
  - Deep Q-Networks (DQN): DQN combines reinforcement learning with deep neural networks.
  - Semi-Supervised Learning Algorithms: These algorithms utilize both labeled and unlabeled data to learn patterns and make predictions. They are useful when labeled data is scarce.

  - Transfer Learning Algorithms: These algorithms leverage knowledge learned from one task to improve performance on a different but related task.

Choosing the right machine learning algorithm is of utmost importance. Hence, this article provides a comparison to aid researchers in selecting the most suitable one.

K-Means Clustering: An unsupervised learning algorithm that partitions the data into k clusters based on their similarity in the feature space.

IV. METHODOLOGY

We follow the below process to compare machine learning algorithms. (i) Data collection, (ii) Clustering, (iii) Data Pre-processingand (iv) Data testing

A. *Dataset Collection*

The ChEMBL bioactivity data [1] provides us with a database consisting of over 2 million compounds with curated bioactivity data. This data is compiled from over 76,000 documents and includes 1.2 million assays across 13,000 targets and 1,800 cells, spanning 33,000 indications.To properly collect data from the Chembl database[1], we follow these steps:

1. Install and import the libraries needed to download the data, including panda library which allows manipulation and analysis of the data. It gives data structures and operations for manipulating numerical arrays and time series[12].
2. Search for the dataset corresponding to the coronovirus disease
3. Select and retrieve from the downloaded dataset the most compatible with our needs in terms of size and characteristics, we choose the fourth entry from list of Table 1, then we select just the data that has the IC50 Value.
4. Handling missing data; delete incomplete data, using his id.

B. *Clustering*

We have structured our data into three categories and according to the IC50 value of each:

- If the value of IC50 is less than 1000, the data is labeled "active"
- If the value of IC50 is between 1000 and 10000, the data takes the label of "intermediate"
- And if the IC50 value is greater than 10000, the data will be structured in the "inactive" zone

Figure 2: Description of the features considered

Table 2 shows the features chosen in this work, we have combined between the canonical_mile and the bioactivity_class (active, inactive or intermediate) and the standrd_values in a single dataframe and save it as a CSV file.

C. *Data Pre-processing*

As a first step in the data preparation phase, we have made all values greater than 10^8 in the standard value column to have 10^8 as the upper value.

Before calculating the Lipinski discriptors, we converted the IC50 value to the negative logarithmic scale in order to allow the IC50 data to be more evenly distributed.
To achieve this goal, we have customized the function which first converts nM to M by multiplying the IC50 value by (10^-9), then it calculates the new value following the calculation model:

pIC50 = -log(IC50 * 10^(-9))

The custom function removes the old value from IC50 and replaces it in a new column with the pIC50 value
Figure 2 shows the difference between the two IC50 and pIC50 values; also removing values greater than 10^8

In order to create a more evenly distributed set of IC50 data, we will perform a conversion of IC50 values to the negative logarithmic scale (-log10(IC50)). This will be done through the implementation of a custom function called pIC50(), which will take in a DataFrame as input and perform the following operations:

Extract IC50 values from the standard_value column and convert them from nanomolar (nM) units to molar (M) units by multiplying the value by 10^-9.
Apply the negative logarithmic scale (-log10) to the molar IC50 value.
Remove the original standard_value column and replace it with a new pIC50 column containing the converted values.

```
[14] df_combined.standard_value.describe()

count    1.330000e+02
mean     8.017926e+12
std      3.344888e+13
min      5.000000e+01
25%      1.070000e+04
50%      2.350000e+04
75%      3.000000e+05
max      3.311311e+14
Name: standard_value, dtype: float64
```



```
df_norm.standard_value_norm.describe()

count    1.330000e+02
mean     2.110164e+07
std      4.089714e+07
min      5.000000e+01
25%      1.070000e+04
50%      2.350000e+04
75%      3.000000e+05
max      1.000000e+08
Name: standard_value_norm, dtype: float64
```

```
df_final.pIC50.describe()

count    133.000000
mean       4.060148
std        1.783762
min        1.000000
25%        3.522879
50%        4.628932
75%        4.970616
max        7.301030
Name: pIC50, dtype: float64
```

Figure 3 : calculates the pIC50 value

We then removed the intermediate class, in order to have a dataset which contains two classes "active" and "inactive"

We plotted the diagram of the pIC50 value to check the distribution of the two classes of bioactivity that we defined in the data processing.

The result of the diagram is the following:

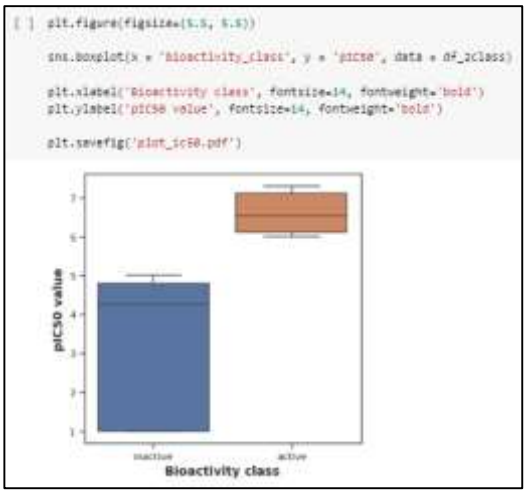

Figure 7 : Diagram of the pIC50 value

The distribution as we expected because we use as a threshold to label the "active" IC50 value 1000 and after the processing we have done to the data, the threshold becomes 6, because -log (1000*10^6) = 6 and -log (10000*10^-9) = 6, then values greater than 6 are "active" and less than 5 are incative

D. *Data testing*

In applied machine learning, we often need to determine whether two data samples have the same or different distributions.

The Mann-Whitney U test [13] is a nonparametric test of statistical significance to determine whether two independent samples have been drawn from a population with the same distribution. It give as :

- H0 rejection failed: sample distributions are equal.

- Reject H0: The sample distributions are not equal.

For this work we will perform the mann-whitney test to test the statistical significance of the difference whether they are different or not different, the code to perform for the test is downloaded from [13] and modified. the mann-whitney function compares the active class and the inactive class to determine if there is a difference or not.

The results after the mann-whitney test are as follows :

```
[ ] mannwhitney('pIC50')
```

	Descriptor	Statistics	p	alpha	Interpretation
0	pIC50	972.0	0.122713	0.05	Same distribution (fail to reject H0)

Figure 12: Test Data

Looking at pIC50 values, active and inactive show no statistically significant difference, which is to be expected since the cut-off values (IC50 < 1,000 nM = Active while IC50 > 10,000 nM = Inactive, corresponding to pIC50 > 6 = Active and pIC50 < 5 = Inactive) were used to define active and inactive.

V. Comparing Regressors : Results and Discussion

In this part, we will compare several ML algorithms to choose the most efficient model in order to build the regression model of the Covid 19 disease

A. *Evaluation Criteria*

This part gives a definition of the evaluation criteria used then we evaluate the model built using Scatter plot of the experimental and predicted values of pIC50, then we make the comparisons between the other machine learning algorithms

- RMSE

A metric that tells us how far predicted values are from observed values in a data set, on average. The lower the RMSE, the better a model fits a data set. [14] [15]

Formula 1:

$$RMSE = \sqrt{\Sigma(P_i - O_i)^2 / n}$$

- R-squared

The actual calculation of R-squared requires several steps. This includes taking the data points (observations) of dependent and independent variables and finding the line of best fit, often from a regression model. From there you would calculate predicted values, subtract actual values and square the results. This yields a list of errors squared, which is then summed and equals the unexplained variance. [16] [17] [18]

Formula 2:

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

B. *Comparative Study*

In this section, we will conduct a comparative analysis of multiple machine learning algorithms, employing the evaluation criteria outlined in the preceding paragraph.

```
# Performance table of the training set (80% subset)
predictions_train
```

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
Lars	3165510097627003736111413068562751268957601032...	-2864032945472050699382247323461487710853944530...	29724548102203554831867631119827173422603750604...	0.11
KernelRidge	7.39	-4.78	4.22	0.03
QuantileRegressor	2.22	-0.10	1.85	0.27
LassoLars	2.11	0.00	1.76	0.02
DummyRegressor	2.11	0.00	1.76	0.04
PassiveAggressiveRegressor	1.98	0.12	1.65	0.02
Lasso	1.97	0.12	1.65	0.02
ElasticNet	1.69	0.37	1.39	0.04
LarsCV	1.58	0.48	1.27	0.51
OrthogonalMatchingPursuitCV	1.51	0.54	1.19	0.04
LassoLarsCV	1.50	0.55	1.18	0.12
ElasticNetCV	1.49	0.56	1.17	6.18
LassoCV	1.48	0.56	1.16	5.00
SVR	1.47	0.57	1.15	0.05
LinearSVR	1.47	0.58	1.14	0.11
HuberRegressor	1.45	0.59	1.13	0.07
BayesianRidge	1.45	0.59	1.12	0.06
KNeighborsRegressor	1.45	0.59	1.12	0.02
NuSVR	1.44	0.61	1.10	0.02
GammaRegressor	1.43	0.61	1.10	0.03
SGDRegressor	1.43	0.61	1.10	0.05
TweedieRegressor	1.42	0.62	1.08	0.03
AdaBoostRegressor	1.41	0.63	1.06	0.05
PoissonRegressor	1.39	0.64	1.05	0.03
HistGradientBoostingRegressor	1.39	0.65	1.04	0.20
LGBMRegressor	1.39	0.65	1.04	0.08
OrthogonalMatchingPursuit	1.39	0.65	1.04	0.02
RidgeCV	1.36	0.68	1.00	0.05
BaggingRegressor	1.35	0.69	0.98	0.04
RandomForestRegressor	1.33	0.70	0.96	0.37

Figure 15: comparative study between different ML algorithms

C. Evaluation

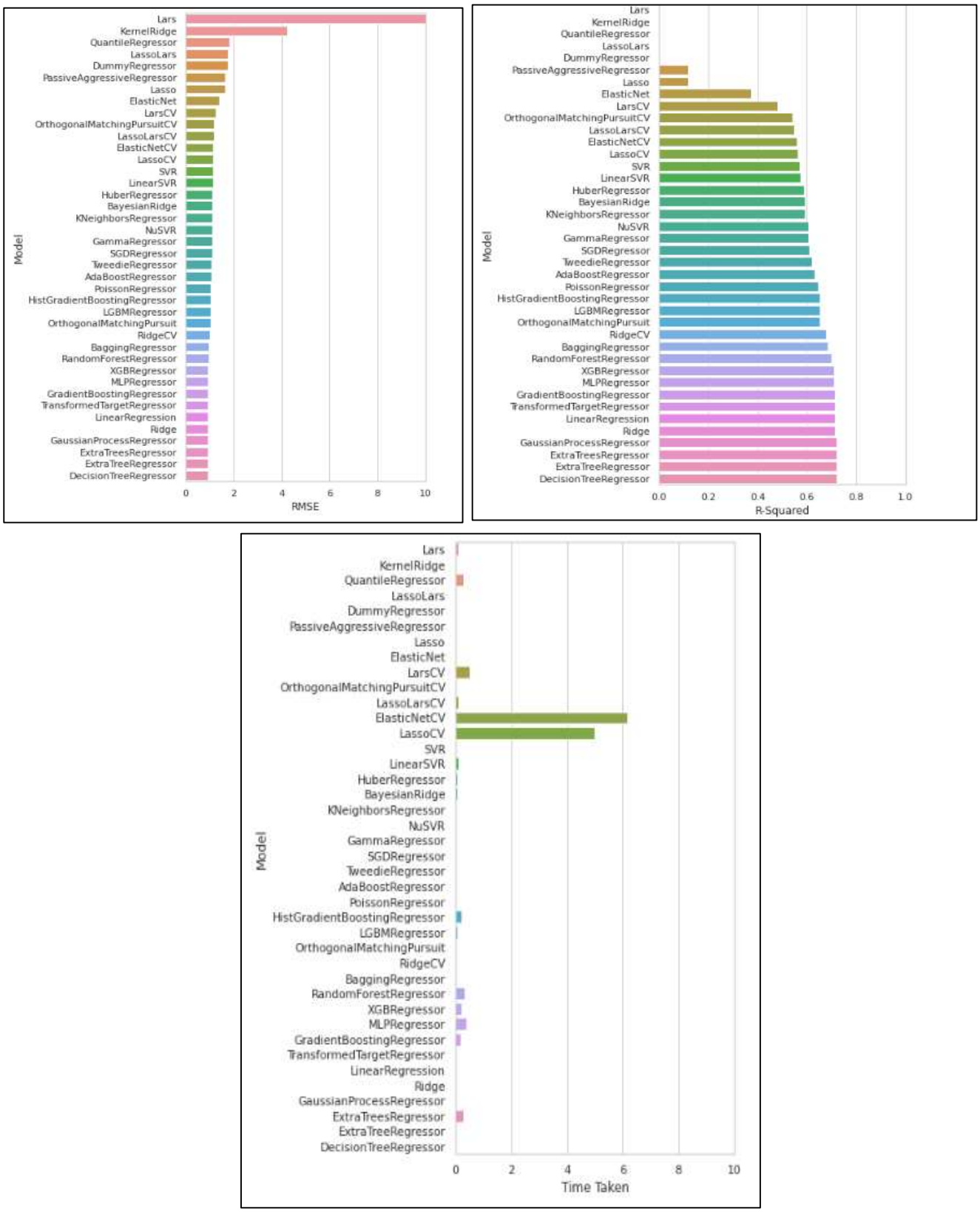

Figure 16 : Evaluation of machine learning algorithms according to RMSE, R-Squared and time taken criteria

Based on the diagrams presented above, it can be concluded that the Lars algorithm demonstrates superior performance in terms of the RMSE, R-Squared, and time taken criteria. The lower the RMSE value, the better the algorithm's predictive accuracy, while a higher R-squared value indicates a better fit of the model to the data. Additionally, a lower time taken suggests that the algorithm is more computationally efficient.

Therefore, the findings suggest that the Lars algorithm outperforms the other algorithms considered in the study in terms of these evaluation criteria. Researchers and practitioners may consider utilizing the Lars algorithm for their predictive modeling tasks, as it demonstrates promising performance in terms of accuracy, model fit, and computational efficiency.

VI. CONCLUSION AND FUTURE WORK

The emergence of SARS-CoV-2 in late 2019 posed a major public health challenge to health systems in all countries around the world to control viral infections and their spread.

To fight against this pandemic, the use of artificial intelligence and machine learning will be an obligation to accelerate biological processes and interpret the results of in vivo experiments.

Our study aimed to compare and evaluate various machine learning algorithms using biological data related to COVID-19. The objective was to assist researchers in selecting the most appropriate algorithm for their specific needs. To collect the necessary data, we utilized the freely available Chembl database and employed SMILE rotation to extract the chemical structures of the compounds.

Furthermore, we conducted a comprehensive comparative analysis of several standard machine learning models, including linear regression (LR), decision trees, and least absolute deviation, among others. As previously discussed in a dedicated section, these models were evaluated based on their performance using the selected dataset.

The study we conducted surpasses previous works as it compares more than thirty machine learning algorithms. Our study has a broader scope, as it encompasses nearly all existing machine learning algorithms.

It is worth noting that our study solely utilized a single database. We acknowledge that the size and characteristics of the data can significantly impact the performance of machine learning algorithms. As a result, we believe that a future investigation into the effects of different databases and data criteria on our model would be valuable and should be considered.

REFERENCES

[1] ChEMBL data base, [Online]. Available: https://www.ebi.ac.uk/chembl/, accessed on Decembre 25, 2022.

[2] Amir Masoud Rahmani, Seyedeh Yasaman Hosseini Mirmahaleh, Coronavirus disease (COVID-19) prevention and treatment methods and effective parameters: A systematic literature review,

[3] Sustainable Cities and Society,Callejon-Leblic, M.A.; Moreno-Luna, R.; Del Cuvillo, A.; Reyes-Tejero, I.M.; Garcia-Villaran, M.A.; Santos-Peña, M.; Maza-Solano, J.M.; Martín-Jimenez, D.I.; Palacios-Garcia, J.M.; Fernandez-Velez, C.; et al. Loss of Smell and Taste Can Accurately Predict COVID-19 Infection: A Machine-Learning Approach. J. Clin. Med. 2021, 10, 570. https://doi.org/ 10.3390/jcm10040570 *Office québécois de la langue française, 2020, « apprentissage automatique »*

[4] V. Bhadana, A. S. Jalal and P. Pathak, "A Comparative Study of Machine Learning Models for COVID-19 prediction in India," *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, Chennai, India, 2020, pp. 1-7, doi: 10.1109/CICT51604.2020.9312112.

[5] Fondation pour la Recherche Médicale, “Tout savoir sur le Sars-Cov-2 et la Covid-19” 54 rue de Varenne - 75007 Paris, 11/02/2022.

[6] Medicine, vol. 15, pp. 195-200, Mar. 2017.

[7] Martin Koppe, “Vaccination : quelle stratégie face aux variants ?” le 22.02.2022

[8] "A Few Useful Things to Know About Machine Learning" by Pedro Domingos. doi:10.1145/2347736.2347755

[9] Christian Robert (2014) Machine Learning, a Probabilistic Perspective, CHANCE, 27:2, 62-63, DOI: 10.1080/09332480.2014.914768
To link to this article: https://doi.org/10.1080/09332480.2014.914768 "Pattern Recognition and Machine Learning" by Christopher M. Bishop.

[10] "Machine Learning: The Art and Science of Algorithms that Make Sense of Data" by Peter Flach.

[11] "Foundations of Machine Learning" by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar.

[12] Librarie Pandas, Available :https://pandas.pydata.org/

[13] Test U de Mann-Whitney https://datatab.fr/tutorial/mann-whitney-u-test

[14] Jason Brownlee on May 16, 2018 in Statistics, “How to Calculate Nonparametric Statistical Hypothesis Tests in Python”, Available : https://machinelearningmastery.com/nonparametric-statistical-significance-tests-in-python/

[15] BY ZACH, “RMSE vs. R-Squared: Which Metric Should You Use? “, JUNE 22,2021 available https://www.statology.org/rmse-vs-r-squared/

[16] Gouda W, Almurafeh M, Humayun M, Jhanjhi NZ. Detection of COVID-19 Based on Chest X-rays Using Deep Learning. Healthcare (Basel). 2022 Feb 10;10(2):343. doi: 10.3390/healthcare10020343. PMID: 35206957; PMCID: PMC8872326.

[17] Zhang Y, Sun F, Huang Z, Song L, Jin S, Chen L. Predicting the impact of the COVID-19 pandemic on globalization. J Clean

Prod. 2023 Jul 10;409:137173. doi: 10.1016/j.jclepro.2023.137173. Epub 2023 Apr 21. PMID: 37101511; PMCID: PMC10119637.

[18] Alzahrani SI, Aljamaan IA, Al-Fakih EA. Forecasting the spread of the COVID-19 pandemic in Saudi Arabia using ARIMA prediction model under current public health interventions. J Infect Public Health. 2020 Jul;13(7):914-919. doi: 10.1016/j.jiph.2020.06.001. Epub 2020 Jun 8. PMID: 32546438; PMCID: PMC7837129.