AI BASED MACHINE LEARNING DESIGN FOR GENOTYPE PREDICTION AGAINST COVID 19 VIRUS

N.Jagadeeswari,
Assistant Professor, Department of Computer Science and Engineering ,Thanthai Periyar Government Institute of Technology, Vellore-02. Email id: jagadeeswarirajesh01@gmail.com

Abstract:

It has become more crucial to use computational methods in the discovery and development of new drug-like molecules for the treatment of diseases. People who have recently become infected are at a significant risk for mortality or serious sequel from these infections. Researchers have access to a vast array of chemical, biological, and clinical data. As a result, AI based machine learning algorithms can be used to discover connections between specific chemical data properties physiological activity of the molecules. These data may also be used to develop models predicting how a particular virus genotype relates to a patient clinical response to treatment. In this paper, AI based machine learning techniques have been used in antiviral research. Virus-host interactions, medication resistance prediction, and the development of new antiviral medicines and vaccines can all benefit from machine learning technologies. Coronaviruses are the most important consideration in the present machine learning evaluation against antiviral research that shows an accurate analysis of 98% that is higher than other methods.

Keywords: Antiviral compound discovery, Machine learning, covid 19 virus

1. Introduction

When it comes to COVID-19 drug discovery, it is all about trial and error. However, virtual screening (VS) is becoming increasingly popular due to the inefficiencies of laboratory throughput screening [1]. Reasoning-based drug discovery is an approach for restricting cell growth and/or activity that is based on the computational targeting of certain macromolecules [2]- [5]. Because both computational and empirically validated viral protein structures have access to, virtual screening (VS) is an efficient and cost-effective technique for identification of the candidates [6] [7].

It has been established that conventional vaccine discovery procedures are expensive, and it may take years for vaccine development that is effective against a specific virus. An approach to vaccine design known as Reverse Vaccinology (RV) [8] was introduced in the early 1990s that eliminated the need for bacterial culturing in order to discover vaccine targets. This approach revolutionized the discipline by doing away with the need for bacterial culture to discover vaccine targets [9]-[11]. In addition, it is possible to identify the protein antigens than only those isolated from bacterial cultures [12], which is a significant improvement over previous methods. Scientists were able to build RV prediction programs as a result of the combination of these benefits.

Artificial intelligence models have revolutionized drug discovery research (as in Figure 1) during the previous decade [13]– [15]. As a result of artificial intelligence, rule-based filtering methods [16] and RV virtual frameworks [17] have been established. Through AI based machine learning (ML), it is feasible to develop models to generalize from patterns found in current data while also making inferences from patterns that have not yet been observed. Because of advancements in deep learning, automated feature extraction from raw data can now be used as part of the learning process in deep learning (DL). Deep learning feature extraction has also recently been discovered to produce superior performance than other computer-aided models [18]–[20], which is a significant breakthrough.

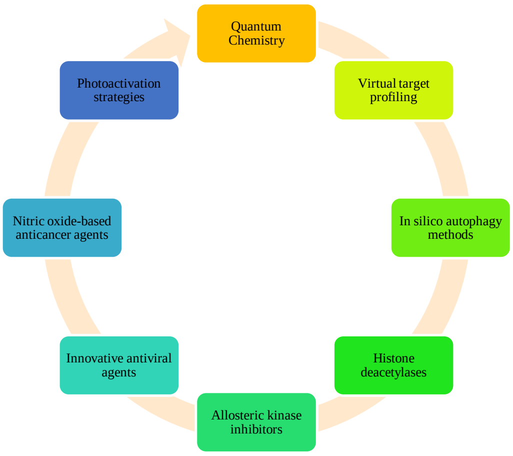
Quantum Chemistry
Virtual target profiling
In silico autophagy methods
Histone deacetylases
Allosteric kinase inhibitors
Innovative antiviral agents
Nitric oxide-based anticancer agents
Photoactivation strategies

Figure 1: Schematic of Antiviral Discovery strategies

In this work, an AI based machine learning technologies have been used to aid in the discovery of new antiviral agents. Computer-aided drug discovery by the prediction of treatment resistance, are all made possible through the application of Artificial Intelligence. Coronaviruses are the most important element in the current machine learning appraisal of antiviral research, and they are also the most prevalent.

2. Background

Machine learning has had a significant impact on a wide range of scientific and engineering areas over the last several years. Speech and facial recognition [21], as well as customized targeted marketing [22], are just a few examples of how artificial intelligence has influenced our daily lives. In large part, the success of AI may be due to the tremendous amount of data available and the capacity to perform autonomous feature learning [23]. A considerable impact on drug and vaccine development has been achieved [24], with predictions of drug and vaccine activity , chemical characteristics 25], all being made possible by the method.

Artificial Intelligence has also had a favorable impact on vaccine design according to the authors. When it comes to antigen prediction, VaxiJen was the first RV approach to include ML, and it has shown promising results. Employing mathematical machine learning to RV, an application that employs ML to predict antigens, provides further evidence that using mathematical machine learning to RV has proven to be a successful technique. Overall, these pipelines contain features for extraction, selection, and data augmentation and cross-validation. These pipelines are used to anticipate vaccine candidates against a variety of bacteria and viruses that are known to cause infectious diseases, including influenza and hepatitis.

For the first time, antibodies can now be represented by graph-based properties rather than expert-designed qualities, which is a significant advancement. In [27], mean pooling is used to categorize data after it has been extracted from a graph using graph feature extraction, pooling, and classification using deep models. DL have revolutionized the area of vaccination, allowing for accurate neoantigens prediction and binding affinity [28]. In the past few years, auto encoders for the human leukocyte antigen (HLA-A) have made significant progress in extracting the characteristics, which could be employed in vaccine discovery.

In the field of AI, the ability of a computer to learn from data is a critical component of success. Bioactivity and toxicity of pharmaceutical substances can be anticipated using a number of computational modelling methods that use AI. Aside from these applications, AI can be utilized in drug discovery to predict protein-protein interactions and ADMET characteristics, as well as protein folding and QSAR.

It is possible to predict LogP and solubility properties with high accuracy using classic QSAR techniques. In spite of the fact that artificial intelligence is an excellent tool for identifying preclinical candidates more cost-and time-efficiently than ever before, accurately predicting the binding affinity of therapeutic compounds to receptors is still problematic for a variety of reasons. First and foremost, artificial intelligence (AI) is a data mining technology whose efficacy is greatly dependent on the quantity and quality of data provided as input. There is a scarcity of high-quality data from public databases, as well as a diverse range of data sources, such as different biological investigations, which makes AI learning more difficult.

3. Proposed Method

The training and testing datasets were constructed using information gathered from these sources. The ChEMBL target ID 613731 was carefully selected for cell-based studies in order to exclude tests that did not contain any chemicals, standardize molarity units, and check for chemical structural errors. In order to further curate these data sets and develop the ANN, we used the automated approaches described in Figure 2.

3.1. Artificial Neural Network

The use of artificial neural networks (ANNs) is becoming increasingly used across a wide range of industries to address a wide range of difficulties. Complex and detailed data can be understood, patterns can be created, and trends can be forecasted with the help of neural networks. Feedforward neural networks (FNNs) are the neural networks that are most frequently employed in real-world applications because they are particularly well-suited to handling classification and regression problems. Traditional neural networks typically contain an architectural structure, activation

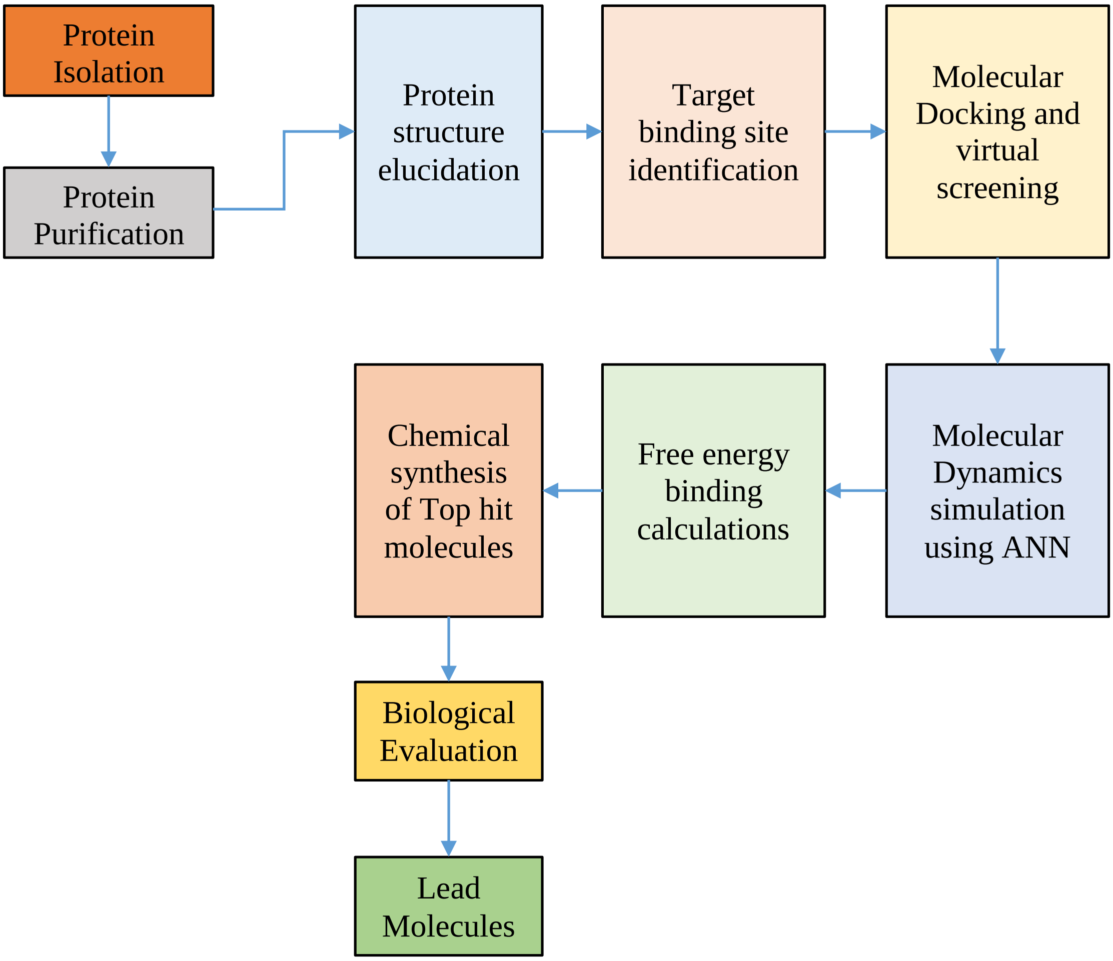
Protein Isolation
Protein Purification
Protein structure elucidation
Target binding site identification
Molecular Docking and virtual screening
Molecular Dynamics simulation using ANN
Free energy binding calculations
Chemical synthesis of Top hit molecules
Biological Evaluation
Lead Molecules

Figure 2: Schematic Layout of the Proposed Workflow

functions for each layer, and a training technique, all of which are well-known concepts. The performance of neural networks is directly influenced by the components listed above.

In order to use FNNs effectively, a large investment in time and resources must be made, both in terms of training and in terms of implementing the results. The FNN training algorithm is responsible for identifying and implementing the optimal weight and inclination combinations for the FNN structures in the FNN structures. Gradients and meta-heuristics are commonly used to train neural networks, and this is a frequent practice.

The gradient technique converges to the optimal solution in a short period of time. The gradient approach, on the other hand, frequently fails to identify the global optimum. Meta-heuristic algorithms, such as crossover operators, allow for random mutations to occur, whereas the gradient process steers each point in the desired direction. A single point is reached via the gradient approach, and it becomes stuck at that location. As a result of their randomness, meta-heuristic algorithms, on the other hand, are able to avoid reaching local optimums. In locations where the gradient is low, deterrent techniques such as gradient descent does not work very well.

It is the most common type of neural network architecture in which information or signals are only formed in one direction, from input to output, and are not created in any other manner. The back propagation algorithm is used to train artificial neural networks. This type of artificial neural network (ANN) contains a layer of neurons that is not visible to the user.

This artificial neural network (ANN) is capable of estimating nonlinear and difficult properties. During training, the MLFNN weights and biases are fine-tuned to achieve the best results. For better or worse, it is founded on the disparity between what was anticipated and what really occurred.

Input of the feed-forward network

It is necessary to establish the input variables of the FNN before it can be used to construct an artificial network. The artificial network for the reliable diagnosis of adenomyosis is constructed by combining a variety of different input variables. The activation function is defined as below:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

This is defined explicitly with linear input combinations as below:

$$\sigma(w_1 x_1 + w_2 x_2 + \cdots + w_N x_N)$$

Adding bias to the unit function, we can modify the equation as below:

$$\sigma(\theta + w_1 x_1 + w_2 x_2 + \cdots + w_N x_N)$$

Preprocessing

The order in which the input variables are presented is crucial for neural network modelling. Standardization of input variables is important because input variables encompass a wide variety of physical units and ranges that are difficult to compare. Therefore, prior to constructing the neural network, it is necessary to standardize the input and output parameters between 0 and 1.

Training procedure

Once the input variables have been discovered, it is possible to find the appropriate neural network architecture to use. This ANN is capable of estimating nonlinear and difficult properties. Optimization of the recommended artificial network is achieved by the use of an algorithm based on repeating procedures. The number of neurons in a FNN is an important performance indicator for this type of neural network. The number of neurons in the network hidden layers is set to the maximum possible number. When the number of neurons in hidden layers is more than their ideal value, it is possible for ANNs to over fit. This may have a negative impact on the FNN overall performance.

During the training procedure, the back propagation method makes use of a Bayesian regularization technique, which is described in detail below. Some feel that the sigmoid function of Tan reflects the tangent hyperbolic features of tangent hyperbolic functions. When constructing an ANN, the Bayesian regularization technique is utilized to establish the weights and biases that should be employed in order to minimize the average square error and build the right ANN. This procedure is known as Bayesian regularization. It is considered a learning step when the number of MSE repetitions remains constant.

4. Results and Discussions

Based on the AI based machine learning technique that was utilized, the results of the training were different. For instance, there is a difference of more than 0.1 across techniques in terms of accuracy, precision, and recall. The compounds considered for simulation is given in Table.1.

Table 1: Compounds Considered

Compound	Antiviral Screening Drug
1	11626003
2	Paroxetine
3	Ezetimibe
4	Cinoxacin
5	SB202190

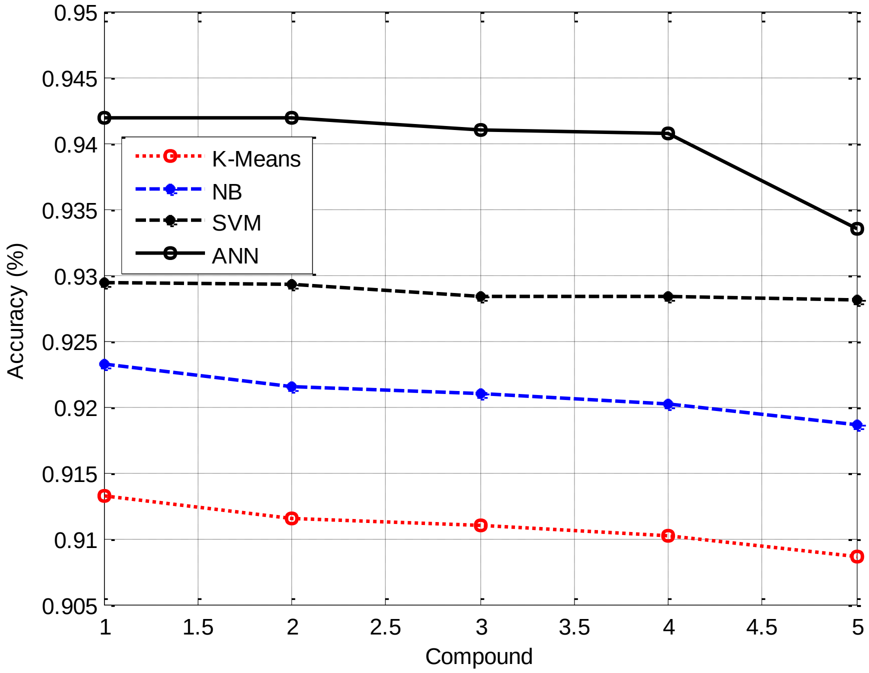

Figure 3: Accuracy

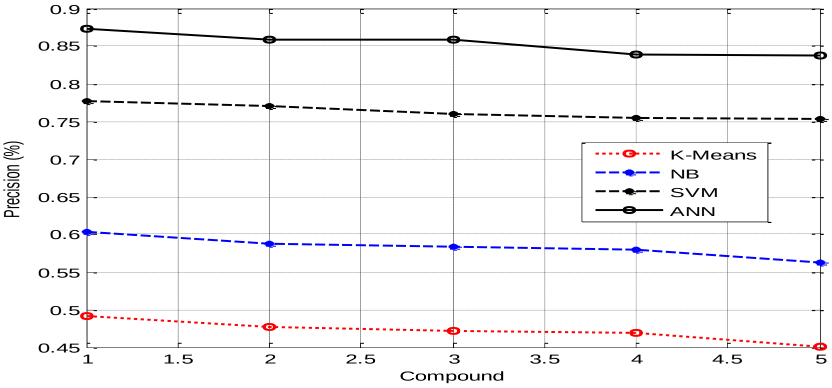

Figure 4: Precision

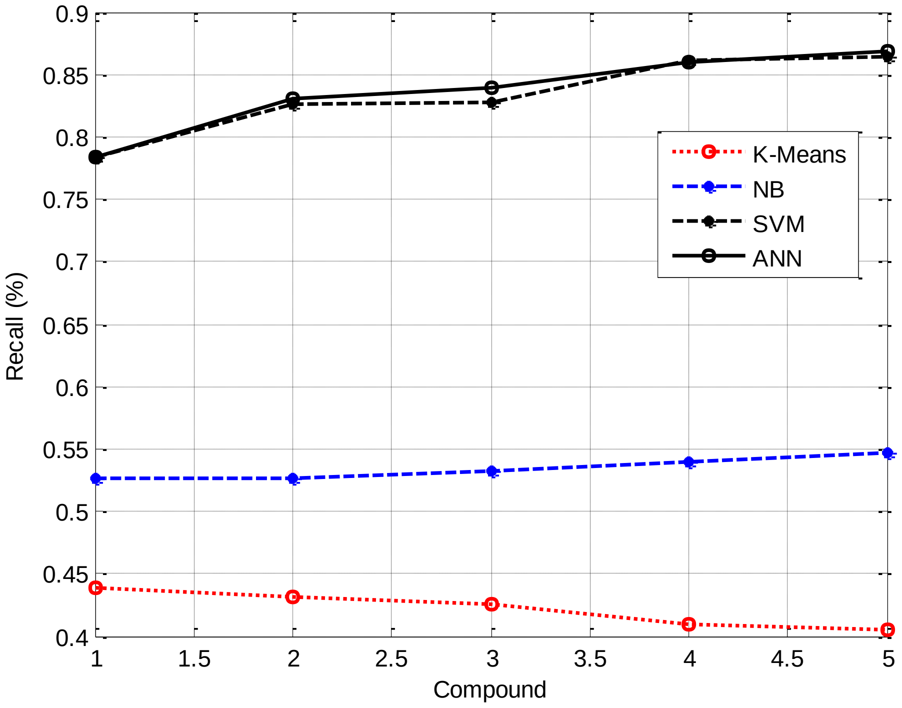

Figure 5: Recall

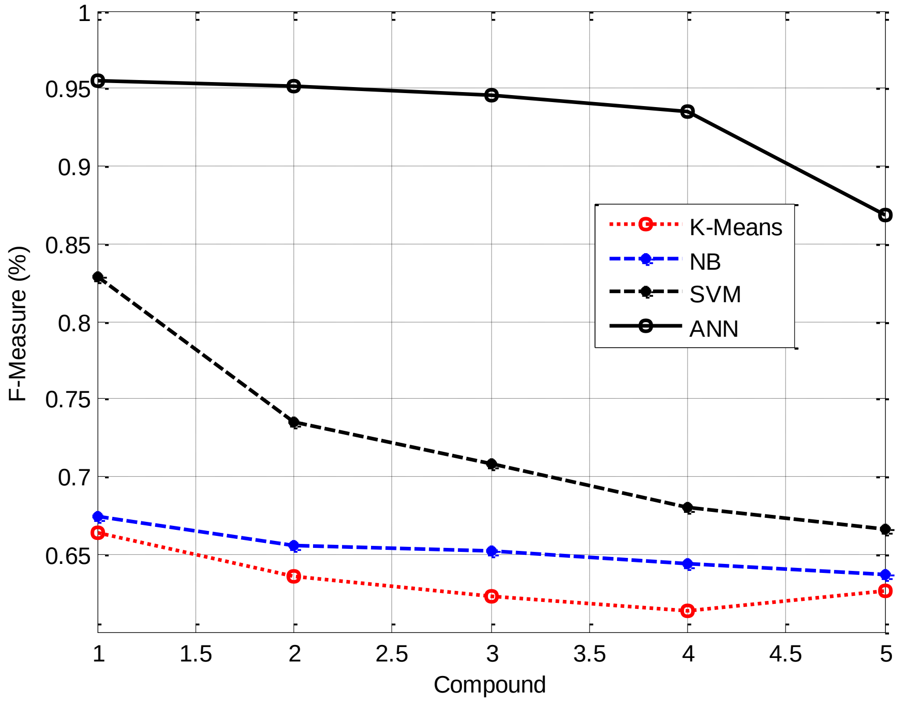

Figure 6: F-measure

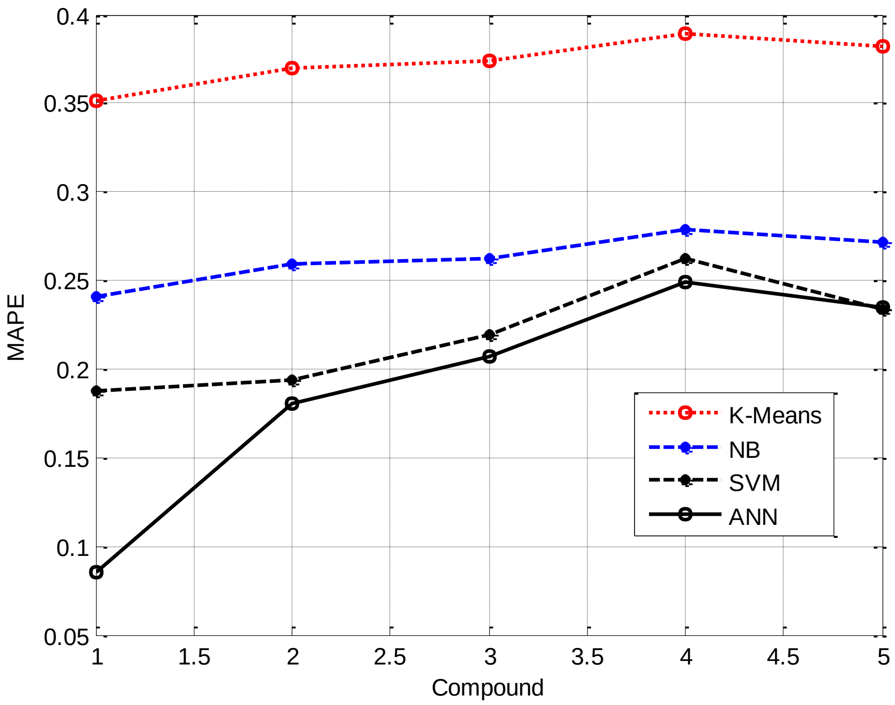

Figure 7: MAPE

Despite the fact that the performance (Figure 3 – Figure 7) on the training set was similar, external validation revealed a significant difference. When all criteria were taken into consideration, DNN emerged as the clear winner in terms of external validation, followed by all other machine learning algorithms.

Therefore, even though the technique was successful in properly classifying molecules in training data, it failed miserably in the external evaluation of the technique. Because machine learning methods cannot distinguish between molecules that are identical but belong to distinct classes, they are unable to determine the crucial features that distinguish them as active or inactive.

According to this example, when dealing with a small data set, ANN performance can be comparable to, or even worse than, that of other methodologies. In general, the findings are consistent with previous research on machine learning and drug development, which has found that RF, SVM, and Bayesian algorithms are the most successful methods of classification.

An external examination revealed that the top three algorithms to use were SVM, RF, and NB. The accuracy, sensitivity, and recall levels obtained from these techniques show that they were able to accurately and reliably predict and distinguish between active and inactive classes of compounds when not using the training data as input. Thus, despite their disparities in precision and recall, the harmonic mean of the two measures for the best three models (with an accuracy of 99% is achieved) is near to one another.

5. Conclusions

In this work, machine learning technologies have been used to aid in the discovery of new antiviral agents. Computer-aided drug discovery and prediction of treatment resistance, are all made possible through the application of machine learning. Every year, an increasing number of people are diagnosed with and affected by COVID-19; the absence of approved pharmaceuticals for treatment and immunizations is a worldwide public health crisis that necessitates urgent research into novel treatments and vaccines to combat the disease. Considering that this RNA virus has the ability to evolve and acquire drug resistance, it is vital to research targeting several therapeutic targets in order to maximize the effectiveness of treatment while also reducing the likelihood that the virus will evolve and develop drug resistance.

References

[1] Kandeel, M., & Al-Nazawi, M. (2020). Virtual screening and repurposing of FDA approved drugs against COVID-19 main protease. *Life sciences*, *251*, 117627.

[2] Lionta, E., Spyrou, G., K Vassilatis, D., & Cournia, Z. (2014). Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Current topics in medicinal chemistry*, *14*(16), 1923-1938.

[3] Yu, W., & MacKerell, A. D. (2017). Computer-aided drug design methods. In *Antibiotics* (pp. 85-106). Humana Press, New York, NY.

[4] Arshadi, A. K., Salem, M., Collins, J., Yuan, J. S., and Chakrabarti, D. (2020). Deepmalaria: artificial intelligence driven discovery of potent antiplasmodials. Front. Pharmacol. 10:1526.

[5] Broom, A., Rakotoharisoa, R. V., Thompson, M. C., Zarifi, N., Nguyen, E., Mukhametzhanov, N., ... & Chica, R. A. (2020). Evolution of an enzyme conformational ensemble guides design of an efficient biocatalyst. *bioRxiv*.

[6] Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706-710.

[7] Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., ... & Hilgenfeld, R. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, *368*(6489), 409-412.

[8] Bullock, J., Luccioni, A., Pham, K. H., Lam, C. S. N., & Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against COVID-19. *Journal of Artificial Intelligence Research*, *69*, 807-845.

[9] Bruno, L., Cortese, M., Rappuoli, R., & Merola, M. (2015). Lessons from Reverse Vaccinology for viral vaccine design. *Current opinion in virology*, *11*, 89-97.

[10] Heinson, A. I., Gunawardana, Y., Moesker, B., Hume, C. C. D., Vataga, E., Hall, Y., ... & Woelk, C. H. (2017). Enhancing the biological relevance of machine learning classifiers for reverse vaccinology. *International journal of molecular sciences*, *18*(2), 312.

[11] Khadidos, A., Khadidos, A. O., Kannan, S., Natarajan, Y.,

Mohanty, S. N., & Tsaramirsis, G. (2020). Analysis of COVID-19 Infections on a CT Image Using DeepSense Model. Frontiers in Public Health, 8.

[12] Bowman, B. N., McAdam, P. R., Vivona, S., Zhang, J. X., Luong, T., Belew, R. K., ... & Woelk, C. H. (2011). Improving reverse vaccinology with a machine learning approach. *Vaccine*, *29*(45), 8156-8164.

[13] Zhong, F., Xing, J., Li, X., Liu, X., Fu, Z., Xiong, Z., ... & Jiang, H. (2018). Artificial intelligence in drug design. *Science China Life Sciences*, *61*(10), 1191-1204.

[14] Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63-71.

[15] Lavecchia, A. (2019). Deep learning in drug discovery: opportunities, challenges and future prospects. *Drug discovery today*, *24*(10), 2017-2032.

[16] Naz, K., Naz, A., Ashraf, S. T., Rizwan, M., Ahmad, J., Baumbach, J., & Ali, A. (2019). PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. *BMC bioinformatics*, *20*(1), 1-10.

[17] Ong, E., Wang, H., Wong, M. U., Seetharaman, M., Valdez, N., & He, Y. (2020). Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. *Bioinformatics*, *36*(10), 3185-3191.

[18] Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, *55*(2), 263-274.

[19] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug discovery today*, *23*(6), 1241-1250.

[20] Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., ... & Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology*, *37*(9), 1038-1040.

[21] Alaghband, M., Yousefi, N., & Garibay, I. (2020). FePh: an annotated facial expression dataset for the RWTH-PHOENIX-weather 2014 Dataset. *arXiv preprint arXiv:2003.08759*.

[22] Zhai, S., Chang, K. H., Zhang, R., & Zhang, Z. M. (2016, August). Deepintent: Learning attentions for online advertising with recurrent neural networks. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1295-1304).

[23] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

[24] Mojjada, R. K., Yadav, A., Prabhu, A. V., & Natarajan, Y. (2020). Machine learning models for covid-

19 future forecasting. Materials Today: Proceedings.

[25] Fooshee, D., Mood, A., Gutman, E., Tavakoli, M., Urban, G., Liu, F., ... & Baldi, P. (2018). Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, *3*(3), 442-452.

[26] Miyake, J., Kaneshita, Y., Asatani, S., Tagawa, S., Niioka, H., & Hirano, T. (2018). Graphical classification of DNA sequences of HLA alleles by deep learning. *Human cell*, *31*(2), 102-105.

[27] Abbasi, B. A., Saraf, D., Sharma, T., Sinha, R., Singh, S., Gupta, P., ... & Gupta, A. (2020). Identification of vaccine targets & design of vaccine against SARS-CoV-2 coronavirus using computational and deep learning-based approaches. preprints

[28] Wu, J., Wang, W., Zhang, J., Zhou, B., Zhao, W., Su, Z., ... & Chen, S. (2019). DeepHLApan: a deep learning approach for neoantigen prediction considering both HLA-peptide binding and immunogenicity. *Frontiers in immunology*, *10*, 2559.