Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Applied Soft Computing 113 (2021) 107945

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A deep learning ensemble approach to prioritize antiviral drugs against novel coronavirus SARS-CoV-2 for COVID-19 drug repurposing

Deepthi K. [a,b,*], Jereesh A.S. [b], Yuansheng Liu [c]

[a] Department of Computer Science, College of Engineering, Vadakara (CAPE, Govt. of Kerala), Kozhikkode 673104, Kerala, India
[b] Bioinformatics Lab, Department of Computer Science, Cochin University of Science and Technology, Kochi 682022, Kerala, India
[c] College of Information Science and Engineering, Hunan University, 2 Lushan S Rd, Yuelu District, 410086, Changsha, China

ARTICLE INFO

Article history:
Received 10 July 2021
Received in revised form 22 August 2021
Accepted 23 September 2021
Available online 6 October 2021

Keywords:
COVID-19 drug repurposing
Deep learning
Convolutional neural network
XGBoost
SARS-coV-2
Antiviral drugs

ABSTRACT

The alarming pandemic situation of Coronavirus infectious disease COVID-19, caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), has become a critical threat to public health. The unexpected outbreak and unrealistic progression of COVID-19 have generated an utmost need to realize promising therapeutic strategies to fight the pandemic. Drug repurposing-an efficient drug discovery technique from approved drugs is an emerging tactic to face the immediate global challenge. It offers a time-efficient and cost-effective way to find potential therapeutic agents for the disease. Artificial Intelligence-empowered deep learning models enable the rapid identification of potentially repurposable drug candidates against diseases. This study presents a deep learning ensemble model to prioritize clinically validated anti-viral drugs for their potential efficacy against SARS-CoV-2. The method integrates the similarities of drug chemical structures and virus genome sequences to generate feature vectors. The best combination of features is retrieved by the convolutional neural network in a deep learning manner. The extracted deep features are classified by the extreme gradient boosting classifier to infer potential virus–drug associations. The method could achieve an AUC of 0.8897 with 0.8571 prediction accuracy and 0.8394 sensitivity under the fivefold cross-validation. The experimental results and case studies demonstrate the suggested deep learning ensemble system yields competitive results compared with the state-of-the-art approaches. The top-ranked drugs are released for further wet-lab researches.

1. Introduction

Coronavirus disease 2019 (COVID-19) is a highly contagious and pathogenic respiratory illness that created a dreadful situation worldwide, affecting people's lives and causing many deaths. The causative agent for the disease, severe acute respiratory syndrome coronavirus-2, SARS-CoV-2 (previously 2019-novel coronavirus, 2019-nCoV), is an enveloped positive-strand RNA virus with mammalian hosts [1,2]. They are classified under the Coronaviridae family and Betacoronavirus genus. SARS-CoV-2 is the seventh coronavirus known to infect humans, and they are genetically very similar to SARS (severe acute respiratory syndrome) and MERS (Middle East respiratory syndrome) coronaviruses [3,4]. With a genome of 80–160 nm in length and 27–32 kb in size, coronaviruses are the biggest among known RNA viruses [5,6]. The ongoing COVID-19 pandemic has highlighted the urgency to develop, test, and deploy new drugs and therapeutics. However, designing a novel drug from scratch is very tough and tedious, and thus impractical to combat the global challenge of the SARS-CoV-2 pandemic. One efficient way to face this challenge is to effectively screen clinically approved drugs for their anti-viral activity against SARS-CoV-2 for repurposing.

Drug repurposing (DR) is the deployment of already approved drugs for different indications other than the drug's original therapeutic application. Since the de novo drug discovery is high-cost, lengthy, and laborious, DR has become a promising strategy to combat newly emerging diseases [7,8]. This effective drug discovery technique can quickly identify potential therapeutic agents for difficult to treat diseases like COVID-19 [9]. DR based on biological experiments is generally a high-risk and high-investment process [10]. The availability of huge biological and structural databases and high-performance computing empowered computational DR as an alternative to experimental approaches to identify the most efficacious drugs against specific diseases in a short time [11–13].

In the present study, we propose a deep learning ensemble approach, DLEVDA, capable of identifying novel virus–drug associations to combat the rapidly evolving pandemic of COVID-19.

* Corresponding author at: Department of Computer Science, College of Engineering, Vadakara (CAPE, Govt. of Kerala), Kozhikkode 673104, Kerala, India.
E-mail address: deepthi523@gmail.com (Deepthi K.).

https://doi.org/10.1016/j.asoc.2021.107945
1568-4946/

Ensemble methods integrate the potential of multiple classifiers, thus yield models with enhanced predictive power and credibility. The proposed approach combines the pairwise similarities of drug chemical structures and virus genome sequences to build classification features. The generated samples were fed to the convolutional neural network for learning intricate input patterns. The learned abstract features were used to train the extreme gradient boosting classifier, which infers promising candidate drugs against SARS-CoV-2 infections. We conducted fivefold cross-validation (CV) to assess the efficacy of DLEVDA; The model achieved an AUC of 0.8897 with 0.8571 prediction accuracy and 0.8394 sensitivity. The comparison results with the state-of-the-art methods and competing classifiers reveal the predictive power of the approach. Experiments were conducted with distinct datasets. We could confirm the majority of the predicted results with existing literature which indicates the robustness of the model in prediction.

The rest of the paper is organized as follows: Section 2 discusses a review of recent works in drug repositioning to identify potential therapeutic drugs targeting covid-19. Section 3 describes the data preparation and architectural details of DLEVDA. Section 4 presents and analyzes the results of the proposed method. Section 5 discusses the results and Section 6 concludes the paper with future directions.

2. Related works

Based on the method, computational DR applied to infer antiviral drugs against COVID-19 falls mainly into two categories: network-based and machine learning-based. Network-based DR methods represent drugs, diseases, and other biological entities like proteins, genes, etc., as network nodes and their associations as edges between nodes. Zhou et al. [14] suggested a method that identifies new anti-viral drugs against covid-19 by analyzing the association networks related to the Human coronaviruses (HCoVs) and drugs with their target proteins in the human protein–protein interaction (PPI) network. They collected the target proteins related to the HCoVs and generated the HCoV-host protein subnetwork. By computing the network proximity among the drug's target proteins and HCoV-related proteins, potential anti-SARS-CoV-2 drugs were identified. Fiscon et al. [15] proposed a method that identifies new anti-viral drugs against covid-19 from already approved drugs. They quantified the relationships between drug targets and disease-associated genes using a network similarity-based approach to identify repurposable anti-SARS-CoV-2 drugs. The method utilized a drug–target network based on the drugs and their target proteins and a disease–gene network based on the diseases and their associated genes for prediction. The idea behind the algorithm is that a drug will be effective for a specific disease if the drug targets and the disease genes are nearby in the constructed network. Their algorithm calculated the network proximity value between each drug and disease to prioritize appropriate drug candidates against SARS-CoV-2. Meng et al. [16] proposed a method that predicts new anti-viral drugs against SARS-CoV-2 using similarity constrained probabilistic matrix factorization. They utilized drug chemical structure and virus genomic sequence-based similarities and known drug–virus relationships for prediction. They applied probabilistic matrix factorization on the drug–virus relationship matrix by introducing similarity constraints for drugs and viruses in the factorization process.

Adhami et al. [17] suggested a method that infers novel therapeutic drugs against COVID-19 by identifying the causal genes behind it. They retrieved the PPI network corresponding to the human proteins interacting with SARS-CoV-2 from the String database [18]. They identified 7 clusters of proteins that are deeply linked to SARS-CoV-2 and retrieved the genes and associated miRNAs related to the identified protein clusters. Finally, they acquired the drug targeting gene modules from the DGIdb [19] and then rebuilt the drug–gene network for the obtained protein modules. Next, they implemented a network-oriented drug repositioning method using computational bioinformatics tools to identify novel anti-viral drugs to fight COIVD-19. Peng et al. [20] presented an approach that predicts novel anti-viral drugs against COVID from FDA-approved drugs by utilizing drug chemical structure similarities, virus genome sequence-based similarities, and known drug–virus relationships. By integrating this data, they built a heterogeneous network and applied the random walk with restart algorithm [21] on the built network to identify new anti-viral drugs against SARS-CoV-2. The algorithm predicts the association score between SARS-CoV-2 and each drug in the dataset.

Though artificial intelligence-based researches are very active in tackling the Covid-19 epidemic, few articles are concerned with DR. Beck et al. [22] suggested a method that identifies currently available drugs that can interact with the proteins of SARS-CoV-2. They trained a pre-trained drug–target interaction prediction deep learning model [23], with the samples made of drug SMILES (Simplified Molecular Input Line Entry System) strings and amino acid sequences to infer new drug–virus associations. Ke et al. [24] implemented a deep learning method that prioritizes known drugs for their efficacy in fighting against SARS-CoV-2. They trained the model with two datasets; one with the approved drugs against viruses like SARS-CoV, influenza virus, etc., and the other with the confirmed protease inhibitors. They tested the efficacy of the identified drugs using in vitro cell-based assays. With the obtained results, they retrained their model and finally built a model that could identify efficacious drugs against COVID-19. Both of these studies did not evaluate their models quantitatively, and so they do not have exactly comparable results. Systematic deep learning or ensemble machine learning techniques are not applied in the field of virus–drug association prediction. Based on the status of the studies mentioned above, we are proposing a deep learning ensemble approach to infer novel virus–drug associations for identifying promising drug candidates against SARS-CoV-2.

3. Materials and methods

3.1. Datasets

This section describes the preparation of data used in the study:

Drug–virus associations: Experimentally verified virus–drug relationships are obtained from various literature through text mining technology. The dataset contains 455 human drug–virus associations between 219 drugs and 34 viruses. A binary matrix R is to represent the drug–virus associations. $R(r(i), v(j))$ is set as 1 if the drug $r(i)$ has an association with virus $v(j)$; otherwise $R(r(i), v(j)) = 0$.

Intra-drug similarities: The intra-drug similarities were quantified based on the chemical structures of drugs. The drug chemical structures were acquired from DrugBank [25] by adopting the SMILES format [26]. The Molecular Access System (MACCS) fingerprints of drugs were computed using Open Babel v2.3.1 [27]. The similarity between the two drugs was measured using the Tanimoto index [28] based on their MACCS fingerprints. The Tanimoto index between two drugs can be defined as:

$$T = \frac{n_c}{n_a + n_b - n_c}$$

where n_a and n_b represent the number of bits set in the corresponding drug fingerprints, and n_c represents the number of bits that are set in both the fingerprints.

Intra-virus similarities: The intra-virus similarities were measured based on the virus genome sequences. The virus genome nucleotide sequences belonging to the human hosts were acquired from National Center for Biotechnology Information, NCBI [29]. Their pairwise sequence similarities were computed with Multiple Alignment using Fast Fourier Transform, MAFFT version 7 [30], a multiple sequence alignment tool. The known drug–virus relationships and the pairwise similarities among drugs and viruses were acquired from [16]; We consider it the benchmark dataset for this study.

3.2. Methods

In this research, we presented an ensemble machine learning method that identifies new virus–drug relationships utilizing the drug, virus pairwise similarities, and known drug–virus interactions. The overall idea of DLEVDA is narrated in Fig. 1. The proposed approach includes two main segments: convolutional neural network (CNN) and Extreme Gradient Boosting (XGBoost). DLEVDA first generated feature vectors for each drug–virus pair in the dataset by considering the drug and virus pairwise similarities. The feature vector $FV\ (r_i, v_j)$ for the drug–virus pair (r_i, v_j) can be represented by

$$FV(r_i, v_j) = [R_{sim}\ (r_i)\ \boldsymbol{\Lambda} V_{sim}(v_j)]$$

where $R_{sim}\ (r_i)$ denotes the chemical structure-based similarities of the ith drug to all other drugs, $V_{sim}(v_j)$ denotes the genomic sequence-based similarities of the jth virus to all other viruses, and $\boldsymbol{\Lambda}$ denotes the concatenation operation. In more detail, we defined $R_{sim}\ (r_i) = [\ w_{i1}, w_{i2}, w_{i3},\ldots\ , w_{ix}\ldots\ldots\ , w_{iN_r}\ ]$ and $V_{sim}(v_j) = [z_{j1}, z_{j2}, z_{j3},\ldots\ldots\ldots\ , z_{jx}\ldots\ldots\ldots\ldots\ , z_{jN_v}\ ]$, where w_{ix} denotes the pairwise similarity between the ith and xth drugs, z_{jx} is the pairwise similarity between the jth and xth viruses, and N_r, N_v the total number of drugs and viruses respectively in the dataset. Altogether, there were $N_r \times N_v$ samples of length $N_r + N_v$; each corresponds to a drug–virus relationship. The associated labels were picked from the relationship matrix R. The label was set to one if there was a confirmed interaction between the corresponding drug and virus; otherwise, to zero. The samples with label one formed the positive set. Next, random samples from unconfirmed interactions were selected and created the negative set such that the ratio of positive and negative samples was 1:1. There is a chance for unknown positive interactions among the chosen negative samples, but the probability, $455 \div (219 \times 34 - 455) \approx 0.065$ %, is very less when compared to the total unknown interactions in the dataset. Finally, the positive and negative sets were integrated to generate the training set, which comprised 910 samples. With CNN, the intricate patterns of the samples were extracted and fed to the XGBoost classifier to identify novel drug candidates against SARS-CoV-2 and other viruses in the dataset.

3.2.1. Convolutional neural network for feature extraction

CNN is a deep learning algorithm proposed by Lecun et al. [31] that consists of three essential layers — convolutional layer, subsampling layer, and fully connected layer. Convolutional layers are the primary building blocks of CNN which are capable of capturing hidden patterns from the raw input data. CNN comprises two fundamental sections: feature extraction and classification. Feature extraction is achieved by multiple convolutions and subsampling layers and classification by fully connected layers. CNNs were effectively utilized for feature learning and classification in prediction problems for identifying the relationships between diseases, drugs, microRNAs, circular RNAs, etc. [32–34]. In DLEVDA, we employed CNN for extracting the sophisticated input patterns from the concatenated drug–virus feature vectors in a deep learning manner. We performed multiple convolution operations on the input samples using different kernels to generate the activation map. The activation map Q_k at layer k can be described as:

$$Q_k = \vartheta(Q_{k-1} \odot W_k + b_k),$$

where $\vartheta(p)$ denotes the activation function, W_k the convolution kernel at layer k, b_k the offset vector, and $\odot$ the convolution operation. To compress data and minimize overfitting, the subsampling layer is used. The sampling formula at the subsampling layer Q_k can be expressed as:

$$Q_k = Subsampling\ (Q_{k-1})\ .$$

We employed max-pooling at the subsampling layer, which retains the most prominent feature at each filter area. The CNN was trained to decrease the loss function of the network. The training samples were sent to the CNN to capture the significant features. To get our best model, we tuned the CNN hyper-parameters through several experiments. We implemented the convolution operation by using 16 filters of 1×16 size. At the subsampling layer, we set the filter size to 1×2. We used rectified linear unit, Relu [35], as the activation function at the convolution and fully connected layers and the sigmoid function at the output layer. The model was implemented using binary cross-entropy as the error function and Adam as the optimizer. To prevent overfitting, dropout layers [36] are added with convolution and hidden layers. Finally, the learned latent representations after numerous convolution and pooling operations are retrieved for identifying the potential virus–drug relationships.

3.2.2. Extreme gradient boosting based classification

XGBoost is a classification algorithm founded by Chen and Guestrin [37] which works under the framework of gradient boosting. In XGBoost, classification, and regression trees, CART, is created in sequential form. The basic idea is to continuously reduce the residual of the prior model in the gradient direction to get a new model. The algorithm employs multiple regularization parameters, including LASSO (L1) and Ridge (L2), which help prevent overfitting and improve performance.

XGBoost has been successfully applied for binary classification problems such as microRNA/lncRNA-disease association predictions [38,39], prediction of hot spots in protein–DNA binding interfaces [40], protein submitochondrial localization prediction [41], etc. This study established a deep learning model in which XGBoost was employed to perform the task of classification. The feature vectors obtained after convolution and pooling operations from the CNN were a dense, high-level representation of the original samples. We trained the XGBoost classifier with the training set features learned by CNN; the trained model could predict the correlation score for each unverified drug–virus pair in the dataset. The samples with scores above the threshold were considered as potential virus–drug associations, and they were released for future biological tests. We optimized XGBoost hyperparameters through grid search and set the values for parameters such as n_estimators, max_depth, and learning_rate to 150,8 and 0.1, respectively. Fig. 2. depicts the basic structure of the CNN-XGBoost model.

4. Results

4.1. Performance evaluation

We conducted the fivefold CV to evaluate the predictive performance of DLEVDA in identifying new virus–drug relationships.

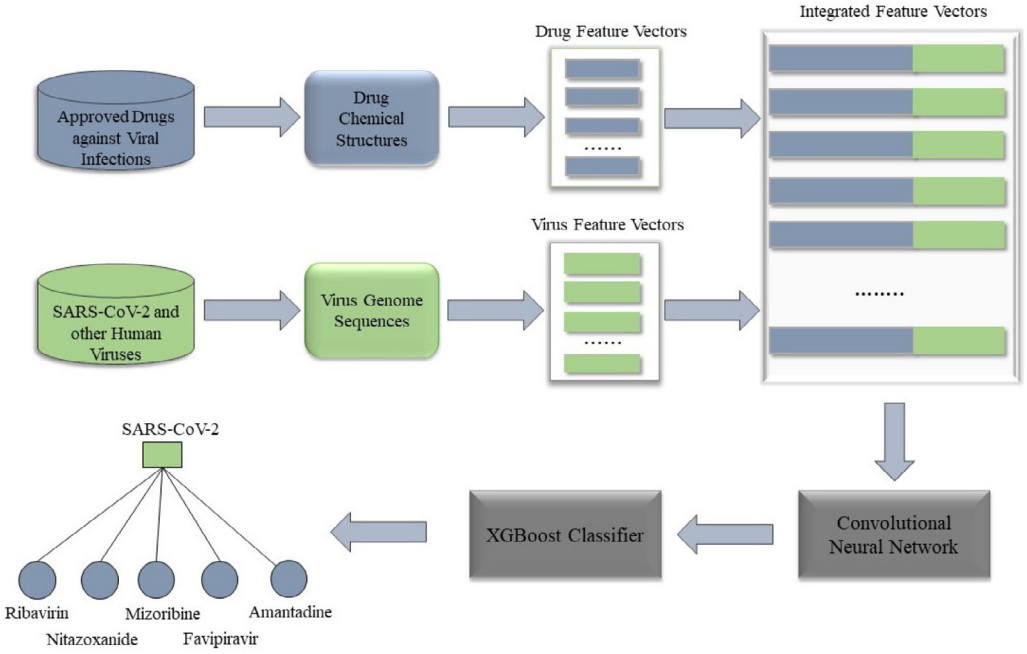

Fig. 1. The workflow of DLEVDA to prioritize potential virus–drug associations. With the integrated features of viruses and drugs as input, DLEVDA comprises two essential components: CNN-based feature learning and XGBoost-based classification.

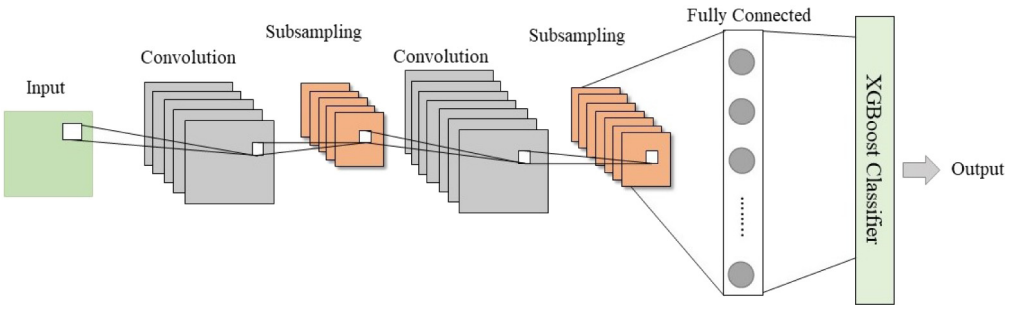

Fig. 2. The basic architecture of the CNN-XGBoost model.

In k-fold CV, the training set is separated into random subsets of uniform size. At each fold, the model was trained with k-1 subsets and validated with the remaining subset. The process was repeated k times until each subset was validated once, and the mean was taken as the end result. In the predicted results, known virus–drug relationships with correlation scores beyond the threshold were treated as true positives (TP), and lower than the threshold were treated as false negatives (FN). Likewise, unknown relationships with correlation scores lower than the threshold were treated as true negatives (TN), and beyond the threshold were treated as false positives (FP). We plotted the Receiver Operating Characteristic curve (ROC) [42,43] by measuring the true positive rate and the false positive rate at different cut-offs, and the model obtained an Area Under the ROC curve (AUC) [44] of 0.8897. We further assessed the predictive capability of DLEVDA by quantifying other statistical parameters such as accuracy, sensitivity, specificity, F1-score, PPV (positive predictive value), NPV (negative predictive value), and Matthews's correlation coefficient (MCC), and the result is shown in Table 1. In addition, we computed the Area Under the Precision–Recall curve (AUPR) [45] as another kind of evaluation metric. The ROC and Precision–Recall (PR) curves based on the fivefold CV are shown in Fig. 3. To diminish the deviations from randomly partitioned samples, we implemented fivefold CV twenty times, and the performances were averaged.

Additionally, we evaluated the model performance by implementing it on an independent test set. We generated the independent test set by randomly choosing 20% samples from the training set such that it contains equal positive and negative samples. The remaining training set samples were partitioned into five random subsets of roughly equal size, which were used as the training and validation sets for the fivefold CV. Next, the validated model is trained with the whole samples in the training set, excluding the independent test set. The trained model is used to predict the correlation score for the samples in the independent test set. The experimental results based on the independent test set are summarized in Table 2.

4.2. Comparison with previous studies

We evaluated the predictive performance of DLEVDA by comparing it to related approaches. Existing researches for identifying repurposable anti-viral drugs against COVID-19 was rare as the COVID-19 researches were mainly focused on sequence data of viruses. We compared DLEVDA with other methods predicting virus–drug relationships such as the similarity constrained probabilistic matrix factorization, SCPMF [16], and virus–drug association prediction based on random walk with restart, VDA-RWR [20], using the same dataset we used. We further evaluated

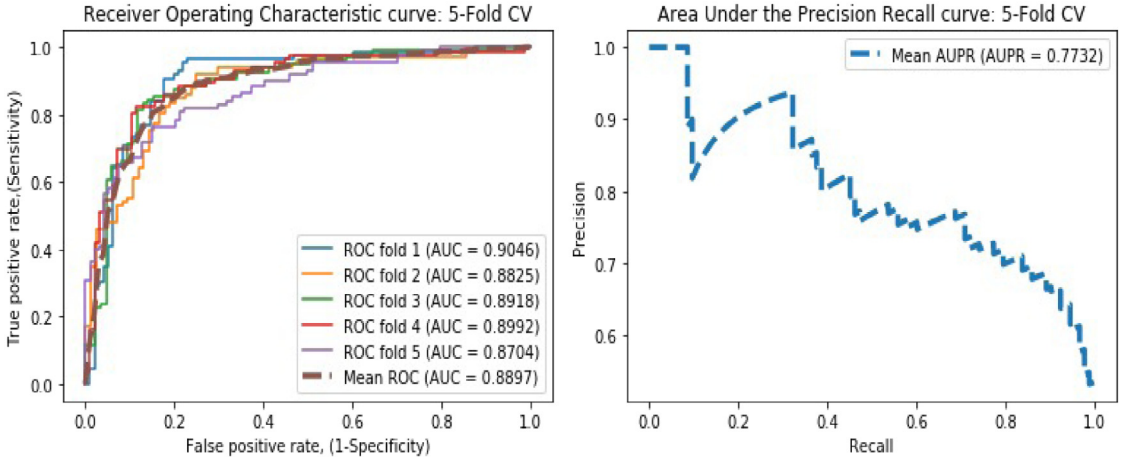

Fig. 3. ROC and PR curves obtained by DLEVDA under the fivefold CV experiment.

Table 1
Cross-validation results of DLEVDA based on fivefold CV experiment.

Method	Accuracy	Sensitivity	Specificity	F1-score	PPV	NPV	MCC	AUC	AUPR
Fivefold	0.8571	0.8394	**0.8624**	0.8432	**0.8563**	**0.8667**	0.7337	0.8897	0.7732

Table 2
Cross-validation results of DLEVDA based on the independent test set.

Method	Accuracy	Sensitivity	Specificity	F1-score	PPV	NPV	MCC	AUC	AUPR
Independent test set	0.8635	0.8418	**0.8692**	0.8316	**0.8602**	**0.8701**	0.7432	0.8926	0.7624

our model by comparing it to other association prediction approaches such as IMCMDA [46], NCPMDA [47], and SAEROF [48]. These three models achieved robust performances in their respective applications. IMCMDA was applied to identify new miRNA-disease associations based on the inductive matrix completion algorithm. NCPMDA identified novel diseases associated with miRNAs based on Network Consistency Projection. SAEROF was applied to predict novel drug–disease relationships utilizing sparse autoencoder and rotation forest. We compared DLEVDA with these models using the same dataset used in our study. The performance was evaluated based on the fivefold CV, and the results are depicted in Table 3. From the table, it is evident that DLEVDA outperformed other methods with high robustness.

4.3. Comparison with different classifiers

To further assess the efficacy of DLEVDA, we compared the model performance with other state-of-the-art classifiers such as random forest (RF), support vector machine (SVM), and decision tree under fivefold CV. In order to assure the fairness of the experiment, we adopted the same feature construction and feature extraction methods during comparison. We could achieve AUCs of 0.8897, 0.8634, 0.8217, and 0.7242 for DLEVDA, RF, SVM, and decision tree classifiers, respectively. Fig. 4 plots a comparison of ROC curves generated by these classifiers. Next, we compared the model performance by implementing these classifiers without employing CNN for feature extraction. The experiments yielded 0.8169, 0.8201, 0.7628, and 0.6581 for XGBoost, RF, SVM, and decision tree classifiers, respectively. From the results, it can be seen that deep learning-based feature retrieval improved the classification results significantly. These experimental results with both the raw and learned high-level features demonstrate the predictive power of DLEVDA in identifying novel virus–drug relationships.

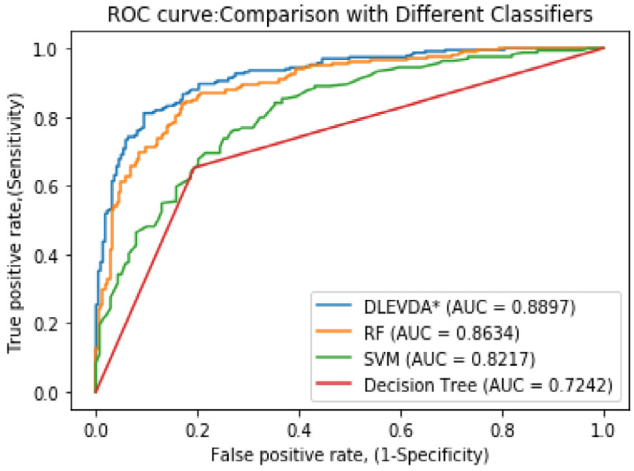

Fig. 4. Comparison of DLEVDA with the competing classifiers under the fivefold CV experiment.

4.4. Comparison with other datasets

To test the influence of various datasets in DLEVDA, we implemented it with another dataset that consists of 96 drug–virus associations between 78 drugs and 11 viruses. In this dataset, 12 viruses akin to SARS-CoV-2 were considered, and their genome sequence-based information was obtained from the NCBI database. Their pairwise similarities were computed using MAFFT version 7. The drugs associated with these viruses were acquired from Drugbank, NCBI, and PubMed databases, and their structural similarities were computed. We conducted fivefold CV, and DLEVDA yielded mean values of 0.8039, 0.7554, **0.7783**, 0.7017, **0.7647, 0.7805**, 0.6505, 0.8420, and 0.6982 for accuracy, sensitivity, **specificity**, F1-score, **PPV**, **NPV**, MCC, AUC, and AUPR,

Table 3
Comparison of DLEVDA with related approaches, through the AUC scores under fivefold CV experiment.

Fivefold	DLEVDA	SCPMF	VDA-RWR	IMCMDA	NCPMDA	SAEROF
AUC score	**0.8897**	0.8631	0.8501	0.6423	0.6711	0.7935

Table 4
The top twelve predicted anti-viral drugs against SARS-CoV-2 with their evidence in the literature.

Rank	Drug	Evidence (PMID)
1	Ribavirin	32227493, 32149772, 33689451
2	Nitazoxanide	32020029, 32568620, 33031085
3	Mizoribine	Unconfirmed
4	Favipiravir	32346491, 32246834, 33176367
5	Amantadine	32361028, 32571606
6	N4-Hydroxycytidine	Unconfirmed
7	Quinacrine	33477376
8	Zanamivir	32511320
9	Maribavir	32147628
10	Chloroquine	32145363, 32074550, 32203437
11	Clevudine	Unconfirmed
12	EIDD-2801	33561864

respectively. The corresponding ROC and PR curves are plotted in Fig. 5. The model performance is slightly low compared to the benchmark dataset used in this study as the number of training data is significantly less. We downloaded this dataset from the supplementary material associated with the paper [20].

4.5. Case studies

To further validate the predictive capability of DLEVDA, we conducted case studies on the top-predicted results. For this, we trained the model with all known drug–virus relationships and predicted correlation scores for all unknown drug–virus pairs in the dataset. The predicted correlation scores were sorted in descending order with the corresponding virus–drug relationships. Specifically, we ranked the top predicted drugs associated with the SARS-CoV-2, Tab. 4. Out of twelve top-predicted drugs, 9 of them could be validated by recent literature. For example, ribavirin, the top-ranked drug against COVID-19, is an anti-viral drug used to treat Hepatitis C and some viral hemorrhagic fevers. It inhibits the replication of RNA viruses and has been applied for treating COVID-19 patients [49–51]. Nitazoxanide, the second top-predicted candidate drug, boosts the host's anti-viral response by upregulating the host interferon and impedes virus replication [52]. It has been proved that Nitazoxanide can prevent SARS-CoV-2 infections at a reduced micromolar concentration and has been recommended for clinical trials to treat COVID-19 [53–55]. The fourth-ranked drug favipiravir is one of the anti-viral agents considered in numerous clinical trials to combat COVID-19. Favipiravir is a purine nucleic acid analog with a broad spectrum of anti-RNA virus activities [56–59]. In addition, among the top twelve predicted drugs against SARS-CoV-2, many are undergoing clinical trials [60,61]. These results reveal the efficacy and credibility of DLEVDA in identifying repurposable ant-viral drugs against COVID-19 and other emerging infectious diseases (see Table 4).

5. Discussion

Prioritization of clinically validated drugs for their anti-viral efficacy is urgent for the rapid clinical trials against COVID-19. This research proposed an ensemble deep learning architecture to infer promising preclinical drug candidates to treat SARS-CoV-2 infections. The proposed architecture comprised two essential segments of CNN-based feature learning and XGBoost-based classification. The CNN was trained with the feature vectors constructed based on drug chemical structures and virus genomic sequences. Then, the learned high-level features were classified with the XGBoost classifier to identify novel candidate drugs to combat COVID-19 and other viral infectious diseases.

There are many factors attributed to the efficient performance of DLEVDA. Ensemble approaches yield excellent results by integrating the potency of multiple classifiers. CNN is a powerful feature extractor that automatically learns high-quality features from the raw input data. However, a large amount of training data is required by CNN to prevent overfitting [62]. The available training data for this research is limited. So, we employed XGBoost for the task of classification. XGBoost utilizes multicore CPU parallel computing to enhance performance. It combines software and hardware optimization strategies to produce more accurate results with lesser computing resources. The incorporation of a regularized model makes the classifier unique [37,63]. However, we examine that XGBoost is still unclear for feature extraction. In addition, the potential of a single classifier may not be sufficient to meet the perfection required for many biological problems. In DLEVDA, the integration of CNN and XGBoost classifiers produced more accurate and robust results.

The network-oriented DR techniques have the drawback that the network needs to be reconstructed whenever a new drug or disease is added to the dataset [14–17,20]. Many of the network-oriented DR strategies cannot be applied to drugs with no confirmed disease interactions or diseases with no confirmed drug interactions in the dataset. DLEVDA can be applied to drugs (diseases) for which no confirmed disease(drug) associations; Hence, we can apply the model to predict potential drugs for emerging diseases like COVID-19. Similar to other machine learning methods, DLEVDA can quickly adapt to changes. When newly discovered drugs, diseases, or drug–virus associations are identified, they can be easily included in the dataset after similarity computation. In addition, DLEVDA incorporated multiple biological data, including the complete genome sequences of viruses and chemical structures of drugs for feature construction. Above all, artificial intelligence-empowered DR is low-cost, fast, and effective and can minimize failures in clinical trials. The limitation is that DLEVDA requires positive and negative samples for training. But it is tough to acquire the actual negative samples. We built the negative set by picking samples from unconfirmed drug–virus relationships at random. There is a possibility for unconfirmed positive interactions in the constructed negative set, even if the probability is low. Besides, the known drug–virus associations available for the study are limited. We believe the performance of DLEVDA can be improved further as more drug–virus associations are discovered. Since similarity scores play a crucial part in predictive performance, it is required to further investigate the kinds of features bundled up for similarity computation.

The overall time complexity of DLEVDA can be expressed as $max(O\left(\tau * m * \left(\sum_{j=1}^{r}(x * k_s^2 * n_k * f_s^2)\right)\right), O(n_t * d * q * log m))$, where the first and second components represent the time complexities of CNN-based feature learning and XGBoost-based classification, respectively, with m training samples [37,64]. In the first component, r denotes the number of convolution layers, τ the number of epochs, x the number of input channels, k_s the spatial size of the kernel, n_k the number of kernels, and f_s the spatial size of the output feature map of the jth layer. In the second component, n_t represents the total number of trees,

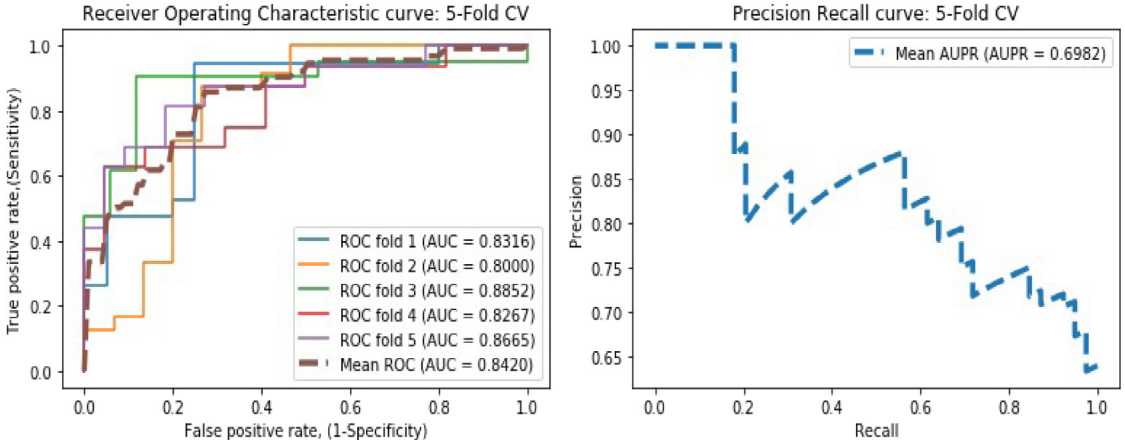

Fig. 5. ROC and PR curves obtained by DLEVDA on a different drug–virus association dataset under the fivefold CV experiment.

d the maximum depth of the tree, and q the number of non-missing entries in the training data. From the equation, it can be inferred that the complexity of DLEVDA depends on the complexity $O\left(\tau * m * \left(\sum_{j=1}^{r}(x * k_s^2 * n_k * f_s^2)\right)\right)$ of CNN-based feature learning. The complexity of pooling and fully connected layers are not involved in this formulation. These layers may take 5%–10% computational time.

6. Conclusion

In summary, this study proposed an efficacious deep learning ensemble model for rapid identification of candidate repurposable drugs to fight against SARS-CoV-2 infections. We carried out extensive experiments and case studies to measure the efficacy of the developed system. The comparison results with the state-of-the-art methods demonstrated an improvement over the existing techniques evaluated under the same condition. Experiments performed with different machine learning classifiers using both the raw and deep features reveal the robustness of the model. The case studies could identify many drugs under clinical trials, which indicate the promising performance of DLEVDA to identify highly credible candidates for experimental analysis. However, the use of randomly chosen negative samples and the limited number of experimentally confirmed virus–drug associations are some of the limitations of the model. All the top predicted drugs against COVID-19 are released for further researches. We believe these drug candidates provide a meaningful reference to support clinicians. In the future, the proposed model can be extended to the next level for predicting the collective effect of a set of drugs against SARS-CoV-2 and other viruses.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code and dataset of DLEVDA are available at https://github.com/Deepthi-K523/DLEVDA.

Implementation details

We implemented DLEVDA with the python toolkits Scikit-learn and Keras library [65,66].

Informed consent

Informed consent has been derived from all the participants.

Funding

No funding received.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] D.S. Hui, E.I. Azhar, T.A. Madani, F. Ntoumi, R. Kock, O. Dar, et al., The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China, Int. J. Infect. Dis. 91 (2020) 264–266.
[2] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nat. Microbiol. 5 (2020) 536–544.
[3] B. Oberfeld, A. Achanta, K. Carpenter, P. Chen, N.M. Gilette, P. Langat, et al., SnapShot: Covid-19, Cell 181 (4) (2020) 954.
[4] M. Pal, G. Berhanu, C. Desalegn, V. Kandi, Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): an update, Cureus 12 (3) (2020).
[5] A.R. Sahin, A. Erdogan, P.M. Agaoglu, Y. Dineri, A.Y. Cakirci, M.E. Senel, et al., 2019 novel coronavirus (COVID-19) outbreak: a review of the current literature, EJMO 4 (1) (2020) 1–7.
[6] C.C. Lai, T.P. Shih, W.C. Ko, H.J. Tang, P.R. Hsueh, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges, Int. J. Antimicrob. Ag. 55 (3) (2020) 105924.
[7] S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, et al., Drug repurposing: progress, challenges and recommendations, Nat. Rev. Drug Discov. 18 (1) (2019) 41–58.
[8] J. Li, S. Zheng, B. Chen, A.J. Butte, S.J. Swamidass, Z. Lu, A survey of current trends in computational drug repositioning, Brief. Bioinform. 17 (1) (2016) 2–12.
[9] S. Dotolo, A. Marabotti, A. Facchiano, R. Tagliaferri, A review on drug repurposing applicable to COVID-19, Brief. Bioinform. 22 (2) (2021) 726–741.
[10] J. Avorn, The $2.6 billion pill–methodologic and policy considerations, N. Engl. J. Med. 372 (20) (2015) 1877–1879.
[11] H.S. Gns, G.R. Saraswathy, M. Murahari, M. Krishnamurthy, An update on Drug Repurposing: Re-written saga of the drug's fate, Biomed. Pharmacotherapy 110 (2019) 700–716.
[12] C. Lippmann, D. Kringel, A. Ultsch, J. Loetsch, Computational functional genomics-based approaches in analgesic drug discovery and repurposing, Pharmacogenomics 19 (9) (2018) 783–797.
[13] M.A. Ahsan, Y. Liu, C. Feng, Y. Zhou, G. Ma, Y. Bai, M. Chen, Bioinformatics resources facilitate understanding and harnessing clinical research of SARS-CoV-2, Brief. Bioinform. (2021).

[14] Y. Zhou, Y. Hou, J. Shen, et al., Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2, Cell Discov. 6 (2020) 14.
[15] G. Fiscon, F. Conte, L. Farina, P. Paci, SaveRUNNER: A network-based algorithm for drug repurposing and its application to COVID-19, PLoS Comput. Biol. 17 (2) (2021) e1008686.
[16] Y. Meng, M. Jin, X. Tang, J. Xu, Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study, Appl. Soft Comput. 103 (2021) 107135.
[17] M. Adhami, B. Sadeghi, A. Rezapour, A.A. Haghdoost, H. MotieGhader, Repurposing novel therapeutic candidate drugs for coronavirus disease-19 based on protein-protein interaction network analysis, BMC Biotechnol. 21 (1) (2021) 1–11.
[18] D. Szklarczyk, et al., STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic Acids Res. 47 (D1) (2018) D607–13.
[19] A.H. Wagner, et al., DGidb 2.0: mining clinically relevant drug–gene interactions, Nucleic Acids Res. 44 (D1) (2016) D1036–44.
[20] L. Peng, L. Shen, J. Xu, X. Tian, F. Liu, J. Wang, et al., Prioritizing anti-viral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures, Sci. Rep. 11 (1) (2021) 1–11.
[21] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, et al., Random walk with restart on multiplex and heterogeneous biological networks, Bioinformatics 35 (3) (2019) 497–505.
[22] B.R. Beck, B. Shin, Y. Choi, S. Park, K. Kang, Predicting commercially available anti-viral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model, Comput. Struct. Biotechnol. J. 18 (2020) 784–790.
[23] B. Shin, S. Park, K. Kang, J.C. Ho, Self-attention based molecule representation for predicting drug-target interaction, in: Machine Learning for Healthcare Conference, PMLR, 2019, pp. 230–248.
[24] Y.Y. Ke, T.T. Peng, T.K. Yeh, W.Z. Huang, S.E. Chang, S.H. Wu, et al., Artificial intelligence approach fighting COVID-19 with repurposing drugs, Biomed. J. 43 (4) (2020) 355–362.
[25] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, et al., DrugBank 4.0: shedding new light on drug mefigolism, Nucleic Acids Res. 42 (D1) (2014) D1091–D1097.
[26] H. Öztürk, E. Ozkirimli, A. Özgür, A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction, BMC Bioinformatics 17 (1) (2016) 1–11.
[27] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, J. Cheminf. 3 (1) (2011) 1–14.
[28] D. Bajusz, A. Rácz, K. Héberger, Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminf. 7 (1) (2015) 1–13.
[29] D.L. Wheeler, C. Chappey, A.E. Lash, D.D. Leipe, T.L. Madden, G.D. Schuler, et al., Database resources of the national center for biotechnology information, Nucleic Acids Res. 28 (1) (2000) 10–14.
[30] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (4) (2013) 772–780.
[31] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.
[32] J. Peng, W. Hui, Q. Li, B. Chen, J. Hao, Q. Jiang, et al., A learning-based framework for miRNA-disease association identification using neural networks, Bioinformatics 35 (21) (2019) 4364–4371.
[33] L. Wang, Z.H. You, Y.A. Huang, D.S. Huang, K.C. Chan, An efficient approach based on multi-sources information to predict circRNA–disease associations using deep convolutional neural network, Bioinformatics 36 (13) (2020) 4038–4046.
[34] K. Deepthi, A.S. Jereesh, An ensemble approach based on multi-source information to predict drug-MiRNA associations via convolutional neural networks, IEEE Access 9 (2021) 38331–38341.
[35] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Icml, 2010.
[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
[37] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[38] X. Chen, L. Huang, D. Xie, Q. Zhao, EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction, Cell Death Dis. 9 (1) (2018) 1–16.
[39] Y. Zhang, F. Ye, D. Xiong, X. Gao, LDNFSGB: prediction of long non-coding rna and disease association using network feature similarity and gradient boosting, BMC Bioinformatics 21 (1) (2020) 1–27.
[40] K. Li, S. Zhang, D. Yan, Y. Bin, J. Xia, Prediction of hot spots in protein–DNA binding interfaces based on supervised isometric feature mapping and extreme gradient boosting, BMC Bioinformatics 21 (13) (2020) 1–10.
[41] B. Yu, W. Qiu, C. Chen, A. Ma, J. Jiang, H. Zhou, Q. Ma, SubMito-Xgboost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics 36 (4) (2020) 1074–1081.
[42] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thoracic Oncol. 5 (9) (2010) 1315–1316.
[43] R. Kumar, A. Indrayan, Receiver operating characteristic (ROC) curve for medical researchers, Indian Pediatr. 48 (4) (2011) 277–287.
[44] A.P. Bradley, The use of the Area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (7) (1997) 1145–1159.
[45] K. Boyd, K.H. Eng, C.D. Page, Area under the precision–recall curve: point estimates and confidence intervals, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, 2013, pp. 451–466.
[46] X. Chen, L. Wang, J. Qu, N.N. Guan, J.Q. Li, Predicting miRNA-disease association based on inductive matrix completion, Bioinformatics 34 (24) (2018) 4256–4265.
[47] C. Gu, B. Liao, X. Li, K. Li, Network consistency projection for human miRNA-disease associations inference, Sci. Rep. 6 (2016) 36054.
[48] H.J. Jiang, Y.A. Huang, Z.H. You, SAEROF: an ensemble approach for large-scale drug-disease association prediction by incorporating rotation forest and sparse autoencoder deep neural network, Sci. Rep. 10 (1) (2020) 1–11.
[49] J.S. Khalili, H. Zhu, N.S.A. Mak, Y. Yan, Y. Zhu, Novel coronavirus treatment with ribavirin: Groundwork for an evaluation concerning COVID-19, J. Med. Virol. 92 (7) (2020) 740–746.
[50] Y.M. Zeng, X.L. Xu, X.Q. He, S.Q. Tang, Y. Li, Y.Q. Huang, et al., Comparative effectiveness and safety of ribavirin plus interferon-alpha, lopinavir/ritonavir plus interferon-alpha, and ribavirin plus lopinavir/ritonavir plus interferon-alpha in patients with mild to moderate novel coronavirus disease 2019: study protocol, Chinese Med. J. 133 (9) (2020) 1132–1134.
[51] M.A. Unal, C.V. Bitirim, G.Y. Summak, S. Bereketoglu, I. Cevher Zeytin, O. Besbinar, et al., Ribavirin shows anti-viral activity against SARS-CoV-2 and downregulates the activity of TMPRSS2 and the expression of ACE2 in vitro, Can. J. Physiol. Pharmacol. 99 (5) (2021) 449–460.
[52] L.D. Jasenosky, C. Cadena, C.E. Mire, V. Borisevich, V. Haridas, S. Ranjbar, et al., The FDA-approved oral drug nitazoxanide amplifies host anti-viral responses and inhibits Ebola virus, Iscience 19 (2019) 1279–1290.
[53] M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, M. Xu, et al., Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro, Cell Res. 30 (3) (2020) 269–271.
[54] V.R. Naik, M. Munikumar, U. Ramakrishna, M. Srujana, G. Goudar, P. Naresh, et al., Remdesivir (GS-5734) as a therapeutic option of 2019-nCOV main protease–in silico approach, J. Biomol. Struct. Dyn. (2020) 1–14.
[55] J.M. Calderón, M.D.R.F Flores, L.P. Coria, J.C.B. Garduño, J.M. Figueroa, M.J.V. Contretas, et al., Nitazoxanide against COVID-19 in three explorative scenarios, J. Infect. Dev. Countries 14 (09) (2020) 982–986.
[56] Y.X. Du, X.P. Chen, Favipiravir: pharmacokinetics and concerns about clinical trials for 2019-nCoV infection, Clin. Pharmacol. Ther. 108 (2) (2020) 242–247.
[57] Q. Cai, M. Yang, D. Liu, J. Chen, D. Shu, J. Xia, et al., Experimental treatment with favipiravir for COVID-19: an open-label control study, Engineering 6 (10) (2020) 1192–1198.
[58] C. Chen, J. Huang, Z. Cheng, J. Wu, S. Chen, Y. Zhang, et al., Favipiravir versus arbidol for COVID-19: a randomized clinical trial, 2020, MedRxiv.
[59] M. Ghasemnejad-Berenji, S. Pashapour, Favipiravir and COVID-19: a simplified summary, Drug Res. (2020).
[60] L. Dong, S. Hu, J. Gao, Discovering drugs to treat coronavirus disease 2019 (COVID-19), Drug Discov. Therapeutics 14 (1) (2020) 58–60.
[61] P. Tarighi, S. Eftekhari, M. Chizari, M. Sabernavaei, D. Jafari, P. Mirzabeigi, A review of potential suggested drugs for coronavirus disease (COVID-19) treatment, Eur. J. Pharmacol. (2021) 173890.
[62] K. Pasupa, W. Sunhem, A comparison between shallow and deep architecture classifiers on small dataset, in: 2016 8th International Conference on Information Technology and Electrical Engineering, ICITEE, IEEE, 2016, pp. 1–6.
[63] J. Ma, Z. Yu, Y. Qu, J. Xu, Y. Cao, Application of the XGBoost machine learning method in PM2. 5 prediction: A case study of shanghai, Aerosol Air Qual. Res. 20 (1) (2020) 128–138.
[64] K. He, J. Sun, Convolutional neural networks at constrained time cost, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5353–5360.
[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[66] F. Chollet, et al., Keras, GitHub, 2015, https://github.com/fchollet/keras.