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.

Computers in Biology and Medicine 132 (2021) 104317

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Deep learning model for virtual screening of novel 3C-like protease enzyme inhibitors against SARS coronavirus diseases

Madhulata Kumari, Naidu Subbarao *

School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

ARTICLE INFO

Keywords:
Deep learning
CNN Model
Convolutional neural network
COVID-19
SARS-CoV
3CLpro
Phytochemical compounds
Virtual screening
FDA-approved drugs

ABSTRACT

In the context of the recently emerging COVID-19 pandemic, we developed a deep learning model that can be used to predict the inhibitory activity of 3CLpro in severe acute respiratory syndrome coronavirus (SARS-CoV) for unknown compounds during the virtual screening process. This paper proposes a novel deep learning-based method to implement virtual screening with convolutional neural network (CNN) architecture. The descriptors represent chemical molecules, and these descriptors are input into the CNN framework to train a model and predict active compounds. When compared to other machine learning methods, including random forest, naive Bayes, decision tree, and support vector machine, the proposed CNN model's evaluation of the test set showed an accuracy of 0.86, a sensitivity of 0.45, a specificity of 0.96, a precision of 0.73, a recall of 0.45, an F-measure of 0.55, and a ROC of 0.71. The CNN model screened 17 out of 918 phytochemical compounds; 60 out of 423 from the natural product NCI divset IV; 17,831 out of 112,267 from the ZINC natural product database; and 315 out of 1556 FDA-approved drugs as anti-SARS-CoV agents. Further, to prioritize drug-like compounds, Lipinski's rule of five was applied to screen anti-SARS-CoV compounds (excluding FDA-approved drugs), resulting in 10, 59, and 14,025 hit molecules. Out of 10 phytochemical compounds, 9 anti-SARS-CoV agents belonged to the flavonoid group. In conclusion, the proposed CNN model can prove useful for developing novel target-specific anti-SARS-CoV compounds.

1. Introduction

The novel severe acute respiratory syndrome coronavirus (SARS-CoV) was first identified as an etiologic agent in 8000 individuals, causing 800 deaths, in July 2003 [1,2]. The virus is an enveloped and positive-sense single-stranded RNA [3]. The symptoms of coronavirus can range from cold to fever, lower respiratory tract infections, and diarrhea [4]. The virus that causes COVID-19, which is also known as SARS-CoV-2, was first identified in Wuhan, China, in December 2019 [https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.pdf] [5]. According to the WHO report, all available evidence for COVID-19 suggests that it is caused by a virus transmitted between animals and people. Since then, the virus has been propagated to other countries by infected people and has become an ongoing global health emergency. No known medicines for the effective management of the disease have created an urgent need to develop novel and effective drugs for treatment. One of the most promising protein targets is the 3C-like protease (3CLpro) of SARS-CoV [6]. The protease inhibitors in the 3CLpro of SARS-CoV are chymotrypsin-like cysteine proteases, which are essential to the coronavirus's proteolytic processing of polyproteins. The protease inhibitors are most effective at blocking replication [7–9]. Thus, the 3CLpro enzyme is a promising target for developing effective inhibitors against SARS-CoV. It is critical to identify a novel candidate for drug development to commit to better treatment during the COVID-19 pandemic.

2. Deep learning for drug discovery

Deep learning is a subfield of machine learning that uses artificial neurons to process data in decision making. Deep learning has been applied to numerous fields, such as text mining and image pattern recognition. The method is also used in drug discovery to speed up the drug development process [10]. In its current state, artificial intelligence and computational biology have unlocked significant potential in the design of drug candidates [11]. Deep learning also plays an important role in the drug discovery process, and it is usually implemented in virtual screening, ADMET properties, QSAR models, and/or for lead optimization [12–14]. In addition, deep learning applications are

* Corresponding author.
E-mail address: nsrao@mail.jnu.ac.in (N. Subbarao).

https://doi.org/10.1016/j.compbiomed.2021.104317
Received 23 December 2020; Received in revised form 2 March 2021; Accepted 2 March 2021
Available online 6 March 2021

Table 1
List of pubchem bioassays of 3CLPro of SARS-CoV.

S. No.	BioAssay AID	Total No. of Compounds	Active Compounds	Inactive Compounds	BioAssay Type
1	1879	380	136	244	Confirmational HTS assay
2	1890	101	44	57	Dose response assay
3	488958	14	9	5	Dose response assay
4	488967	32	15	17	Late stage assay
5	488984	103	10	93	Late stage assay
6	488999	4	3	1	Dose response assay
7	493245	6	3	3	Late stage assay
8	588771	10	5	5	Dose response assay
9	588772	28	14	14	Late stage assay
10	588786	10	3	7	Dose response assay

increasing in the field of pharmaceutical research [15].

To illustrate, a deep learning algorithm was developed to predict drug-induced liver injuries [16]. Gianchandani et al. also proposed ensemble deep transfer learning models to diagnose coronavirus infections from radiography images [17]. Singh et al. proposed a densely connected convolutional networks-based automated COVID-19 screening model [18]. Moreover, the deep learning approach has been used to discover new antibacterial molecules [19]. Kumari et al. applied machine learning algorithms, such as random forest (RF), support vector machine (SVM), and decision tree (DT), for the classification of anti-tubercular molecules [20]. Chen et al. built a deep learning-based model to detect novel coronavirus pneumonia from an image [21]. Peng et al. and Hu et al. employed convolutional neural network (CNN) models to predict drug-target interactions [22,23]. Finally, Meyer et al. applied CNN and RF models to learn drug functions from chemical structures [24]. Virtual screening has excellent applications for in-silico screening, can accelerate the drug discovery process, and can reduce the costs and time associated with experimental work. This study aims to develop a deep learning-based model to screen out novel anti-SARS-CoV agents against 3CLpro enzymes to treat COVID-19 infections.

In the present report, we developed and proposed a deep learning model to build a CNN model based on a dataset of 3CLpro enzymes from SARS-CoV and then applied them to predict anti-SARS-CoV activity in unknown compounds. We also compared the proposed model with popular machine learning methods, including RF, Naïve Bayes (NB), DT, and SVM. Before developing the models, we extracted important descriptor vectors for the prediction of bioactivity. The results of the validated model showed acceptable values for various internal and external validations. The model developed with the deep learning algorithm performed virtual screenings of unknown databases to search for novel inhibitors against 3CLpro enzymes to treat COVID-19 infections.

3. Materials and methods

3.1. Data collection and data curation

We collected publicly available experimental datasets of SARS-CoV from the PubChem Bioassay (https://pubchem.ncbi.nlm.nih.gov/bioassay/) shown in Table 1. Here, we took two types of assays for our study: a conformational high throughput screening bioassay and a dose-response bioassay. The total number of active compounds was 198, and the total number of inactive compounds was 446. After that, we went through the data curation process and obtained 423 unique chemical structures, where 80 compounds were active and 343 compounds were inactive. The activity of the compounds had already been classified by experimental bioassays. We then converted two-dimensional structures to three-dimensional structures by adding hydrogen atoms using CORINA software [25].

3.2. Descriptor calculation

In order to convert chemical molecules into machine language, we converted the three-dimensional structure into a one-dimensional vector containing sufficient structural information, including 147 binary vectors of pharmacophore fingerprints, 24 weighted burdens, and 8 molecular properties employed in PowerMV [26]. Pharmacophore fingerprints are popular methods for molecular representation. Each element of the fingerprint vector indicates the presence or absence of a specific feature in a molecule.

3.3. Dataset division

We implemented a deep learning-based CNN model in TensorFlow. The model was trained on 70% of the random split set and then validated on the remaining 30% of the curated dataset. The following hyperparameters could vary to optimize the model's performance: learning rate, hidden layers, number of neurons, activation functions dropout, and batch normalizations. For virtual screening, the output of the CNN model was the probability for a compound to be active.

3.4. Classification models

Our study built four different machine learning models such as NB, RF, DT, and SVM and compared them with the proposed CNN model. Based on the Bayesian theorem, the NB method assumes that each predictor is conditionally independent of the other [27]. Breiman developed the RF method according to multiple DTs [28]. The DT method builds DTs from a labeled training set using each descriptor to make a decision by splitting the dataset into smaller subsets [29].

3.5. CNN model development

We used labeled data to train the model with learning techniques called supervised learning techniques. There are different supervised learning approaches for deep learning, including deep neural networks, CNN, recurrent neural networks, long short-term memory, and gated recurrent units. CNN has several advantages over other neural networks and is effective at learning, extracting abstractions from two-

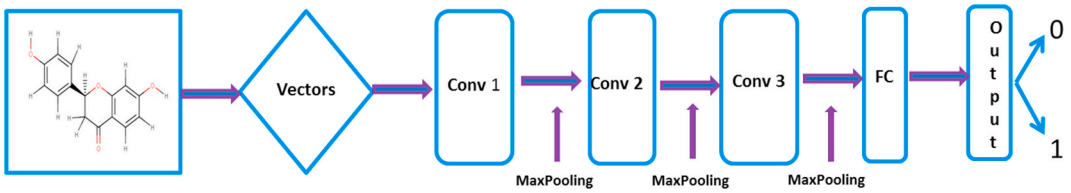


Fig. 1. CNN architecture of our proposed model.

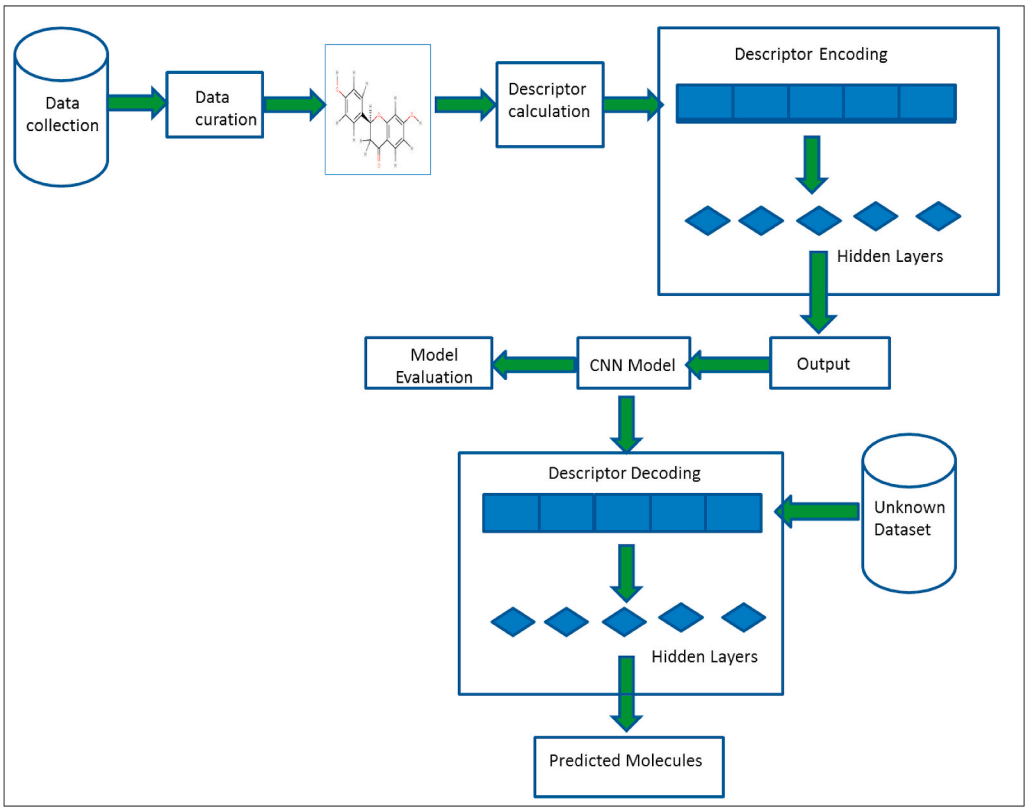


Fig. 2. The illustration of the overall CNN model pipeline for Virtual Screening.

dimensional features, and max-pooling. CNN can produce highly optimized weights. Therefore, among the multitude of available classification methods, we employed CNN because it is highly robust and widely successful in areas outside of drug prediction. Fukushima first proposed this CNN structure in 1988 [30].

3.6. Architecture of CNN

As shown in Fig. 1, this study used the CNN architecture of the proposed model. It consisted of three types of layers: convolutional, pooling, and fully connected (FC) layers.

3.7. Convolutional layers

As the first layer, the convolutional layer is designed to learn the feature map of the input data. It consists of three convolutional layers that work as feature extractors. As shown in Fig. 1, the convolutional layers are composed of several convolution kernels; the kernel sizes of the first, second, and third convolutional layers are 32, 64, and 128, respectively.

3.8. Pooling layers

The second layer is the pooling layer, and it is used to lower the convolutional burden by reducing the number of connections between convolutional layers. Max-pooling is most widely used in CNN architecture to select the most prominent regions of the feature map covered by the filter [31].

3.9. FC layer

The last layer is an FC layer that has a full connection to the neurons [32,33]. The classification layers are FC layers with 655,616 hidden nodes that use sigmoid as their activation function for classification. The sigmoid function is a nonlinear activation function that each neuron in a multilayer neural network uses to predict the probability as an output in the range of 0 and 1. The sigmoid produces a sigmoid curve [34]. The sigmoid function is defined as follows:

$$S(x) = {}^{1}/_{(1+e^{-x})}$$

where x is the input.

3.10. Dropout

Dropout was first introduced by Hindton et al. [35] when they applied it to FC layers. Dropout has since proven to be significantly effective in reducing overfitting. During the training phase, dropout is used as regulation to prevent overfitting and enhance training speed. The dropout neurons have no contribution to the forward or back-propagation during the training phase. The dropouts are set to 0.5 in all layers for classification.

3.11. Rectified linear units (ReLUs)

A ReLU is one of the most non-saturated activation functions [36]. When compared to the sigmoid function, the training time for a ReLU is reduced by fastening the convergence of stochastic gradient descent (SGD). The ReLU activation function is defined as follows:

$$a_{i,j,k} = \max(z_{i,j,k}, 0)$$

where $z_{i,j,k}$ is the input of the activation function at the location (i, j) on the k-th channel.

ReLU is a piecewise linear function that prunes the negative part to zero and retains the positive part. CNN is the most popular neural network and is an effective solution in classification and recognition problems for large datasets [37]. Goh et al. described the CNN model for the prediction of chemical properties in compounds [38].

This study aimed to predict novel drug candidates for COVID-19. We used Python 3.6 for modelling and evaluation. In addition, TensorFlow (https://www.tensorflow.org/), a deep learning library, and Keras (https://keras.io/) were used as architecture for training the CNN model. An illustration of the pipeline is shown in Fig. 2. In general, a backpropagation method is used to optimize the weights between the hidden layers of a CNN, which requires an extreme iteration step to predict output more accurately. Our goal was to develop and train a model so that it can effectively generalize training data, enabling us to measure how well our predicted class matches the actual class. Before the development of a final model, a trial and error method is required. As such, we used a combination of all hyperparameters, including hidden layers, layer type (dense layer), activation function ReLU, output layer function (sigmoid), model optimizer SGD, and epochs, to optimize the model. We used binary cross-entropy as our loss function. Accuracy was used to evaluate the performance of the model. The number of hidden layers was directly proportional to the training time and the increased speed of training. Because training can take several days or weeks to achieve the best performance, we consider training speed to be a valuable property in a cost-effective computing environment that uses Google Collaboratory without GPU.

The calculated descriptors, with 282*179 dimensions for chemical compounds, function as input vectors to feed into CNN and identify active compounds. Here, 179 is the fixed size of the descriptors, and 282 is the hidden state's dimension at each step. Our proposed CNN model consists of three convolutional layers, with max-pooling layers and batch normalization, followed by one fully connected layer and dropout to avoid overfitting problems and improve the performance of prediction.

3.12. Model validation

In order to evaluate the model's performance, various metrics were used for the classification model. The accuracy, sensitivity, specificity, precision, recall, F-measure, ROC, loss, and gain were calculated in SKlearn. Rather than hold onto the last epoch for each target, the best epoch for each target was saved to further screen for unknown molecules. Accuracy, precision, and ROC were applied to assess the proposed model's performance.

The confusion matrix is a specific table that allows for the visualization of the model's performance [39]. While each column of the matrix represents the instances in a predicted class, each row represents the actual class's instances. This makes it easy to see if the system is confusing two classes. A binary classification scheme consists of four sections: true positives (TPs) and true negatives (TNs) represent the correctly predicted active and inactive compounds, respectively; false positives (FPs) indicate that inactive compounds have incorrectly been classified as active compounds; and false negatives (FNs) show that active compounds have incorrectly been classified as inactive compounds.

3.13. Virtual screening and activity prediction of unknown compound databases

The deployment of the unknown dataset on the predictive model is a critical step in identifying novel potential 3CLpro enzyme inhibitors against COVID-19. The developed CNN model was used to virtually screen the phytochemical dataset. It was extracted from a medicinal plant database that contained alkaloid (108), aromatic (81), flavonoid (327), saponin (51), tannin (1), and terpenoid (350) compounds; 423 natural products from the NCI divset IV; 112,267 natural compounds from the ZINC database; and 1556 FDA-approved drugs. Subsequently, Lipinski's rule of five (RO5) was applied during screening to prioritize drug-like compounds. RO5 predicts poor absorption when molecules have more than 5 hydrogen bond donors or more than 10 hydrogen bond acceptors, and when the molecular mass is more than 500 Da, the octanol-water partition coefficient (Log P) is greater than 5, or the rotatable bonds are more than 10 [40]. Finally, the remaining compounds had their anti-COVID-19 activity predicted with the proposed CNN model.

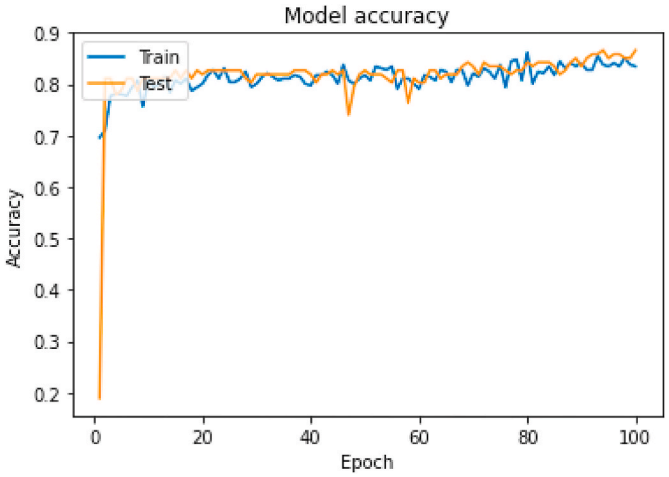


Fig. 3. Accuracy of training and validation set are plotted against the number of training epoch on our CNN model.

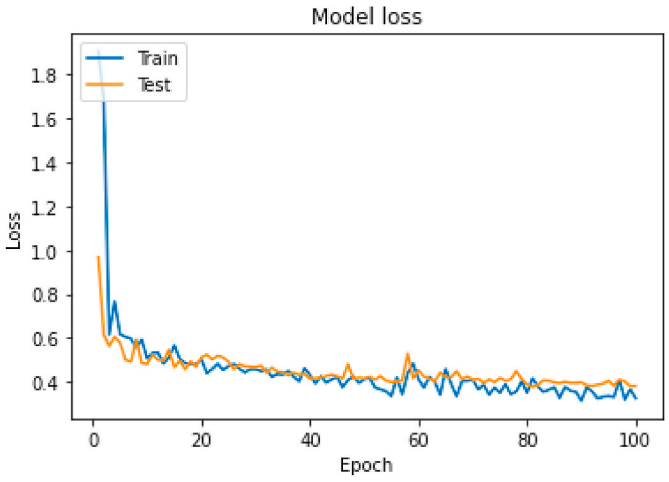


Fig. 4. The loss value of the training and testset are plotted against the number of epoch on our proposed CNN model.

4. Results and discussion

In this investigation, we developed a deep learning-based CNN model to predict the activity of compounds for the inhibition of 3CLpro enzymes in SARS-CoV infections and built a machine learning classifier for comparative analysis. Before the modelling, we collected bioactive molecules from 10 different experimental bioassays and then curated them by removing duplicate compounds, salts, and metal ions. This led us to obtain 423 unique chemical structures for model development. We computed simple, meaningful, and easily interpretable two-dimensional descriptors to develop an easily reproducible model that can be used for the prediction and screening of unknown dataset compounds. The model was designed with 179 descriptors. The descriptors for the molecules describe the structural and functional requirements of the 3CLpro enzyme. Our proposed CNN architecture used three convolution layers: max-pooling, one FC layer with ReLU, and the sigmoid activation function for binary classification. The results showed good predictive ability based on both internal and external validation techniques. The

Table 2
The statistical results of machine learning models of the 3CLPro of SARS-CoV testset.

Classifiers	TP	TN	FP	FN	Accuracy	Specificity	Sensitivity	Precision	Recall	F Measure	ROC
NB	21	25	78	3	0.36	0.24	0.87	0.21	0.87	0.33	0.55
RF	7	100	3	17	0.84	0.96	0.29	0.70	0.29	0.41	0.62
DT	11	92	11	13	0.81	0.89	0.45	0.50	0.45	0.47	0.66
SVM	9	96	7	15	0.82	0.93	0.37	0.56	0.37	0.44	0.65
CNN	11	99	4	13	0.86	0.96	0.45	0.73	0.45	0.55	0.71

NB: Gaussian Naïve Bayes; RF: Random Forest; DT: Decision Tree; SVM: Support Vector Machine; TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative; ROC: Receiver Operating Characteristic.

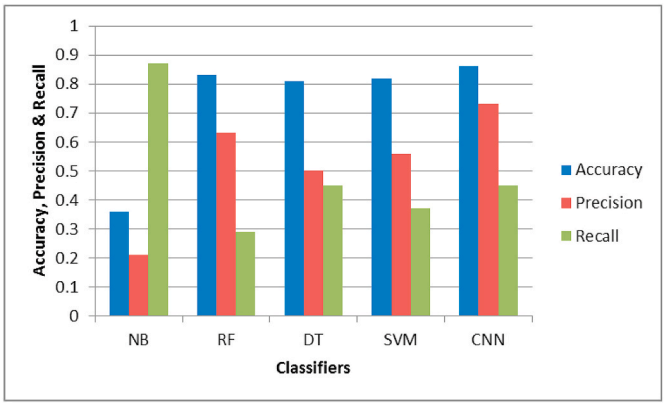


Fig. 5. Bar chart is showing accuracy, precision, and recall for the different models where the CNN model shows maximum accuracy (0.86) and maximum precision (0.73) while NB shows maximum recall (0.87) of the 3CLPro of SARS-CoV testset.

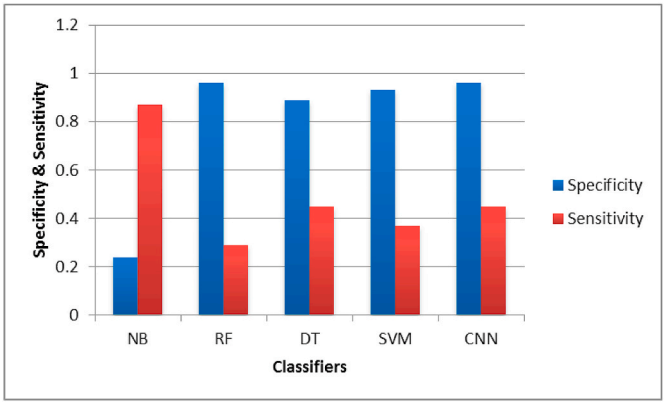


Fig. 6. Bar chart is showing specificity and sensitivity of for the different models where the CNN model and RF show a maximum specificity (0.96) and NB shows a maximum sensitivity (0.87) of the 3CLPro of SARS-CoV testset.

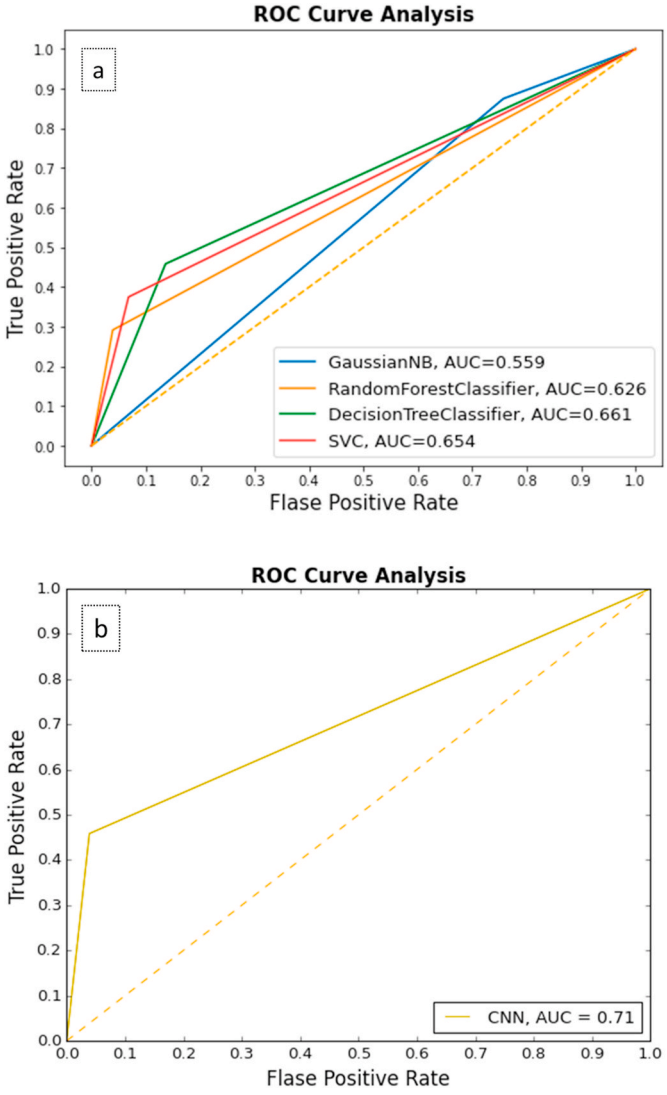


Fig. 7. The ROC plot depicts significant AUC curve values for NB, RF, DT, SVM (a), and CNN model (b) of the 3CLPro of SARS-CoV testset.

CNN framework and six hyperparameters such as learning rate, hidden layers, the number of neurons, activation functions, dropout, and batch normalizations were investigated for quality and performance measures. The varied curves for the accuracy and loss function of our pre-training model on a dataset are illustrated in Figs. 3 and 4, respectively.

5. Loss function of the model

For the optimization algorithm, the loss function was used to evaluate a candidate solution. Fig. 4 presents the two distinct loss curves of a model for the training set and test set of anti-SARS-CoV bioassays named "Train" and "Test," respectively. The training loss curve presents a sharp drop at first, then fluctuates with an increment of epochs, and finally drops slowly. The training loss curves show a faster convergence speed during 2–10 epochs, achieving robust and excellent performance with training. The test set for the loss curves also shows faster convergence at the start and then slowly converges with an increment of epochs. Therefore, a model can take less training time to predict the activity of molecules.

6. Comparative analysis

The best model was chosen by comparing the performance of the

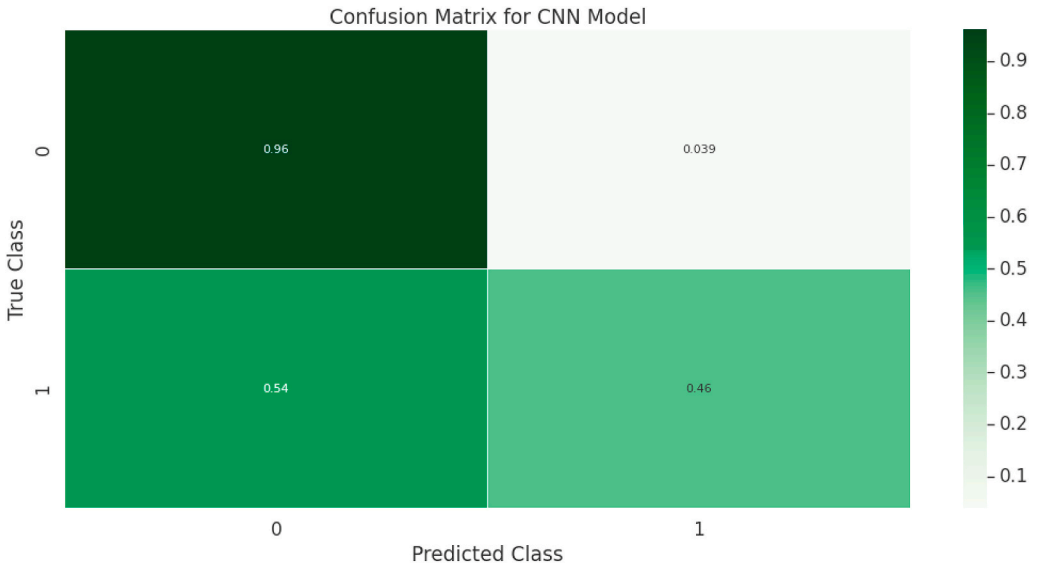


Fig. 8. Heatmap of confusion matrix of CNN model showing the proportion of each predicted class (x-axis) for molecules in each true class (y-axis); 0 represents inactive molecules, and 1 represents active molecules.

models using various statistical parameters. The statistical results based on the test set are reported in Table 2. The accuracy of the models enabled the evaluation of the overall efficiency of the model presented in Fig. 5. The sensitivity is also known as the TP rate and is the proportion of the truly positive parts of the dataset. This means that sensitivity measures the correctly identified active molecules. By comparison, specificity represents the TN rate and measures the proportion of the dataset that is truly negative. Put differently, this means that specificity measures the correctly identified inactive molecules (Fig. 6). The classifier model comparisons revealed that the CNN model achieved 0.86 accuracy, 0.45 sensitivity, 0.96 specificity, 0.73 precision, 0.45 recall, 0.55 F-measure, and 0.71 ROC. The second-best model was RF, which showed an accuracy of 0.84 with a specificity of 0.96 and a sensitivity of 0.29. While the SVM model achieved an accuracy of 0.82 with a specificity of 0.93 and a sensitivity of 0.37, the DT model obtained an accuracy of 0.81 with a specificity of 0.89 and a sensitivity of 0.45. The NB model showed the lowest accuracy of 0.36 with a specificity of 0.24 and the highest sensitivity of 0.87 (Table 2). The sensitivity of the CNN model (0.45) was lower than the sensitivity of the NB model (0.87). This means that the CNN model's prediction of active molecules is genuinely true.

In addition, the ROC was measured to prove the model's robustness. This revealed that the model can be widely used for quick performance assessments of virtual screening approaches. As illustrated in Fig. 7b, the CNN model's AUC curve was 0.71, establishing it as the best model. This was followed by the AUC curves of the DT and SVM models at 0.66 and 0.65, respectively, and the RF and NB models at 0.62 and 0.55, respectively (Fig. 7a and b). The confusion matrix of the CNN model enabled visualization of the percentage of classified compounds. This revealed a TP of 0.46, an FN of 0.54, a TN of 0.96, and an FP of 0.039 (Fig. 8). Hence, the comparative analysis indicated that CNN was the best model, followed by the RF, SVM, NB, and DT models. Based on the above performance, the CNN model was selected as the best among the evaluated models. Therefore, the results suggest that this model can be effective for screening large databases.

7. Deployment of the CNN model for the prediction and virtual screening of activity

The proposed CNN model was used to predict the compounds' activities on various datasets (i.e., the phytochemical database, natural products from the NCI divset IV, natural compounds from the ZINC database, and FDA-approved drugs). We screened molecules based on a trained model. Only 17 out of 918 phytochemical compounds, 60 out of 423 natural products from the NCI divset IV, 17,831 out of 112,267 natural compounds from the ZINC database, and 315 compounds out of 1556 FDA-approved drugs were predicted as anti-SARS-CoV agents. Further, to prioritize drug-like compounds, we applied Lipinski's RO5 on all the screened anti-SARS-CoV compounds except the FDA-approved drugs. This resulted in 10, 59, and 14,025 hit molecules, respectively. Out of the 10 phytochemical compounds shown in Table 3, 9 of the hit molecules belonged to the flavonoid group.

8. Conclusion

This work developed a deep learning-based CNN model that was extremely effective and efficient in its approach to virtual screening. We developed the CNN model to predict anti-SARS-CoV drug candidates and compare them with other classification methods, including RF, NB, DT, and SVM modelling. The model was trained on 282 compounds and predicted an external validation test set of 141 compounds with an accuracy of 0.86, a sensitivity of 0.45, a specificity of 0.96, a precision of 0.73, a recall of 0.45, an F-measure of 0.55, and a ROC of 0.71. The CNN model screened 17 out of 918 phytochemical compounds; 60 out of 423 natural products from the NCI divest IV; 17,831 out of 1,12,267 natural compounds from the ZINC natural product database; and 315 out of 1556 FDA-approved drugs as anti-SARS-CoV agents. Further, to prioritize drug-like compounds, Lipinski's RO5 was applied to all the screened anti-SARS-CoV compounds except the FDA-approved drugs, resulting in 10, 59, and 14,025 hit molecules. Of the 10 phytochemical compounds, 9 anti-SARS-CoV agents belonged to the flavonoid group. To conclude, the proposed CNN model can prove useful for predicting novel target-specific anti-SARS-CoV compounds. The deep learning model can also see widespread use in chemical and drug informatics studies that cover anti-COVID-19 prediction.

Summary

We have developed a deep learning model that may be used to predict the inhibitory activity of 3CLpro of SARS coronavirus for unknown compounds in the virtual screening with the convolutional neural network (CNN) architecture and compared with other classification methods such as RF, NB, DT, and SVM. We extracted the experimental datasets of SARS-CoV from the various PubChem Bioassay. The

Table 3
The screening of active anti-SARS-CoV phytochemical compounds.

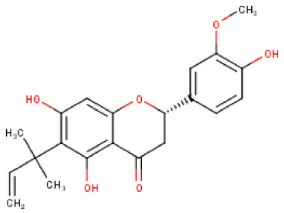
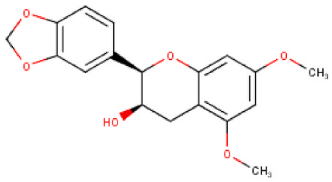
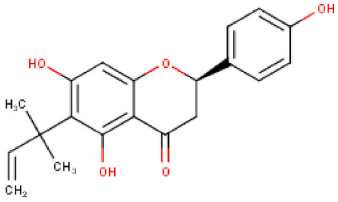
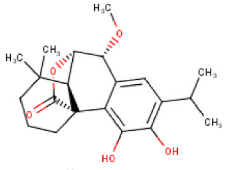
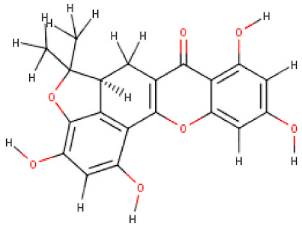
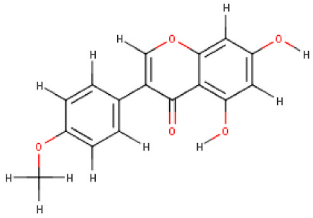
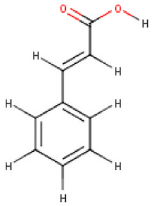
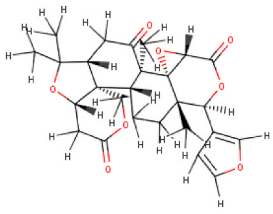
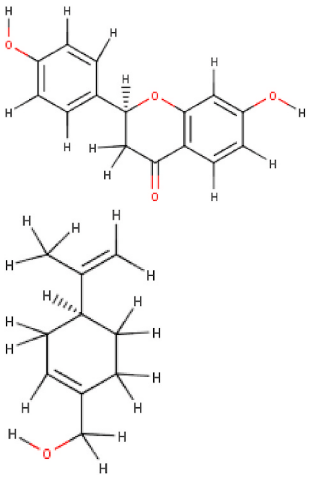
S. No.	Chemical ID	Chemical Structure
1	NPACT00111	
2	NPACT00171	
3	NPACT00182	
4	NPACT00196	
5	NPACT00282	
6	NPACT00335	
7	NPACT00423	
8	NPACT00713	
9	NPACT00716	

Table 3 (*continued*)

S. No.	Chemical ID	Chemical Structure
10	NPACT01038	

total no. of active compounds is 198, and inactive compounds are 446. After that, we have gone through the data curation process and finally got 423 unique chemical structures, where 80 compounds were active, and 343 compounds were inactive. The descriptors represent chemical molecules, and these descriptors are input into the CNN framework to train a model and predict active compounds. When compared to other machine learning methods, including RF, NB, DT, and SVM, the proposed CNN model's evaluation of the testset showed an accuracy of 0.86, a sensitivity of 0.45, a specificity of 0.96, a precision of 0.73, a recall of 0.45, an F-measure of 0.55, and a ROC of 0.71. The CNN model screened 17 out of 918 phytochemical compounds, 60 out of 423 from the natural product NCI divset IV, 17,831 out of 1,12,267 from the ZINC natural product database, and 315 out of 1556 FDA-approved drugs as anti-SARS-CoV agents. Further, to prioritize drug-like compounds, Lipinski's RO5 was applied to screened anti-SARS-CoV compounds (excluding FDA-approved drugs), resulting in 10, 59, and 14,025 hit molecules. Out of 10 phytochemical compounds, 9 anti-SARS-CoV agents belonged to the flavonoid group. The proposed CNN model may see widespread use in chemical and drug informatics studies covering subject anti-COVID-19 prediction.

Declaration of competing interest

None Declared.

Acknowledgements

Centres of Excellence in bioinformatics supported by Department of Biotechnology of Government of India and the Department of Bioinformatics, Indian Council of Medical Research, New Delhi, India.

Abbreviations

3CLpro	3C-like protease
CNN	Convolutional Neural Network
RO5	Lipinsky's rules
ReLU	Rectified Linear Unit
SARS-CoV	Severe Acute Respiratory Syndrome Coronavirus

References

[1] N.S. Zhong, B.J. Zheng, Y.M. Li, Poon, Z.H. Xie, K.H. Chan, P.H. Li, S.Y. Tan, Q. Chang, J.P. Xie, X.Q. Liu, J. Xu, D.X. Li, K.Y. Yuen, Peiris, Y. Guan, Epidemiology and cause of severe acute respiratory syndrome (SARS) in

Guangdong, People's Republic of China, in February, 2003, Lancet 362 (2003) 1353–1358.
[2] J. Ziebuhr, Molecular biology of severe acute respiratory syndrome coronavirus, Curr. Opin. Microbiol. 7 (2004) 412–419.
[3] A. Zumla, J.F.W. Chan, E.I. Chan, D.S.C. Hui, K.Y. Yuen, Coronaviruses -drug discovery and therapeutic options, Nat. Rev. Drug Discov. 15 (2016) 327–347.
[4] S.H. Myint, Human coronavirus infections, in: S.G. Siddell (Ed.), The Coronaviridae. The Viruses, Springer, Boston, MA, 1995, pp. 389–401.
[5] S. Khan, G. Nabi, G. Han, R. Siddique, S. Lian, H. Shi, N. Bashir, A. Ali, M. A. Shereen, Novel coronavirus: how things are in Wuhan, Clin. Microbiol. Infect. 26 (2020) 399–400.
[6] C.W. Lin, F.J. Tsai, C.H. Tsai, C.C. Lai, L. Wan, T.Y. Ho, C.C. Hsieh, P.D. Chao, Anti-SARS coronavirus 3C-like protease effects of Isatis indigotica root and plant-derived phenolic compounds, Antivir. Res. 68 (2005) 36–42.
[7] S. Chen, L.L. Chen, H.B. Luo, T. Sun, J. Chen, F. Ye, J.H. Cai, J.K. Shen, X. Shen, H. L. Jiang, Enzymatic activity characterization of SARS coronavirus 3C-like protease by fluorescence resonance energy transfer technique, Acta Pharmacol. Sin. 26 (2005) 99–106.
[8] R. Ramajayam, K.P. Tan, H.G. Liu, P.H. Liang, Synthesis and evaluation of pyrazolone compounds as SARS-coronavirus 3C-like protease inhibitors, Bioorg. Med. Chem. 18 (2010) 7849–7854.
[9] V. Kumar, K. Roy, Development of a simple, interpretable and easily transferable QSAR model for quick screening antiviral databases in search of novel 3C-like protease (3CLpro) enzyme inhibitors against SARS-CoV diseases, SAR QSAR Environ. Res. 31 (2020) 511–526.
[10] P. Hop, B. Allgood, J. Yu, Geometric deep learning autonomously learns chemical features that outperform those engineered by domain experts, Mol. Pharm. 15 (2018) 4371–4377.
[11] F. Ghasemi, A. Mehridehnavi, A. Perez-Garrido, H. Perez-Sanchez, Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks, Drug Discov. Today 23 (2018) 1784–1790.
[12] K.A. Carpenter, D.S. Cohen, J.T. Jarrell, X. Huang, Deep learning and virtual drug screening, Future Med. Chem. 10 (2018) 2557–2567.
[13] S. Hu, P. Chen, P. Gu, B. Wang, A deep learning-based chemical system for QSAR prediction, IEEE J Biomed Health Inform 24 (2020) 3020–3028.
[14] J. Ma, R.P. Sheridan, A. Liaw, G.E. Dahl, V. Svetnik, Deep neural nets as a method for quantitative structure-activity relationships, J. Chem. Inf. Model. 55 (2015) 263–274.
[15] S. Ekins, The next era: deep learning in pharmaceutical research, Pharm. Res. (N. Y.) 33 (2016) 2594–2603.
[16] Y. Xu, Z. Dai, F. Chen, S. Gao, J. Pei, L. Lai, Deep learning for drug-induced liver injury, J. Chem. Inf. Model. 55 (2015) 2085–2093.
[17] N. Gianchandani, A. Jaiswal, D. Singh, V. Kumar, M. Kaur, Rapid COVID-19 diagnosis using ensemble deep transfer learning models from chest radiographic images, J Ambient Intell Humaniz Comput 16 (2020) 1–13.
[18] D. Singh, V. Kumar, M. Kaur, Densely connected convolutional networks-based COVID-19 screening model, Appl. Intell. (2021), https://doi.org/10.1007/s10489-020-02149-6.
[19] J.M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N.M. Donghia, C. R. MacNair, S. French, L.A. Carfrae, Z. Bloom-Ackermann, V.M. Tran, A. Chiappino-Pepe, A.H. Badran, I.W. Andrews, E.J. Chory, G.M. Church, E. D. Brown, T.S. Jaakkola, R. Barzilay, J. Collins, A deep learning approach to antibiotic discovery, Cell 181 (2020) 475–483.
[20] M. Kumari, N. Tiwari, N. Subbarao, S. Chandra, Evaluation of predictive models based on random forest, decision tree and support vector machine classifiers and virtual screening of anti-mycobacterial compounds, Int. J. Comput. Biol. Drug Des. 10 (2017) 248–263.
[21] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, Q. Chen, S. Huang, M. Yang, X. Yang, S. Hu, Y. Wang, X. Hu, B. Zheng, K. Zhang, H. Wu, Z. Dong, Y. Xu, Y. Zhu, X. Chen, H. Yu, Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography, Sci. Rep. 10 (2020) 19196.
[22] J. Peng, J. Li, X. Shang, A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network, BMC Bioinf. 21 (2020) 1–13.
[23] S. Hu, C.P. Chen, J. Zhang, B. Wang, Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks, BMC Bioinf. 20 (2019) 689.
[24] J.G. Meyer, S. Liu, I.J. Miller, J.J. Coon, A. Gitter, Learning drug functions from chemical structures with convolutional neural networks and random forests, J. Chem. Inf. Model. 59 (2019) 4438–4449.
[25] J. Sadowski, J. Gasteiger, G. Klebe, Comparison of automatic three-dimensional model builders using 639 X-ray structures, J. Chem. Inf. Model. 34 (1994) 4.
[26] K. Liu, J. Feng, S.S. Young, PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation, J. Chem. Inf. Model J Chem Inf Model 45 (2005) 515–522.
[27] N. Friedman, D. Geiger, M. GoldSzmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.
[28] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[29] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106.
[30] K. Fukushima, Neocognitron: a hierarchical neural network capable of visual pattern recognition, Neural Network. 1 (1988) 119–130.
[31] R. Yamashita, M. Nishio, R.K.G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights Imaging 9 (2018) 611–629.
[32] Q. Xu, M. Zhang, Z. Gu, G. Pan, Overfitting remedy by sparsifying regularization on fully-connected layers of cnns, Neurocomputing 328 (2019) 69–74.
[33] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, Recent advances in convolutional neural networks, Pattern Recogn. 77 (2018) 354–377.
[34] Y.A. LeCun, L. Bottou, G.B. Orr, K.-R. Müller, Efficient backprop, in: Neural Networks: Tricks of the Trade, second ed., 2012, pp. 9–48.
[35] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving Neural Networks by Preventing Co-adaptation of Feature Detectors, CoRR abs/1207.0580.
[36] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the International Conference on Machine Learning, ICML, 2010, pp. 807–814.
[37] D. Jimenez-Carretero, V. Abrishami, L. Fernandez-de-Manuel, I. Palacios, A. Quilez-Alvarez, A. Diez-Sanchez, M.A. Del Pozo, M.C. Montoya, Tox_(R)CNN: deep learning-based nuclei profiling tool for drug toxicity screening, PLoS Comput. Biol. 14 (2018), e1006238.
[38] G.B. Goh, C. Siegel, A. Vishnu, N.O. Hodas, N. Baker, Chemception: a Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-Developed QSAR/QSPR Models. arXiv, 2017, p. 1706, 06689.
[39] K.M. Ting, Confusion matrix, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of Machine Learning and Data Mining, Springer, Boston, MA, 2017.
[40] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 46 (2001) 3–26.

Madhulata Kumari is a research associate in the School of Computational & Integrative Sciences, Jawaharlal Nehru University, New Delhi, India. She holds Ph.D. in Information Technology by the Kumaun University, Nainital, Uttarakhand, India. Her main area of research interest is the data mining, machine learning, deep learning, molecular docking, Molecular dynamic simulation, pharmacophore modelling, 3D-QSAR modelling, lead optimization and in silico ADMET prediction and drug design. Her work has been published in various peer-reviewed journals.

Naidu Subbarao is an Associate Professor in the School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India. He received his MSc and PhD from IIT Kanpur. His research interests molecular modelling, molecular docking, Molecular dynamic simulation, pharmacophore modelling, 3D-QSAR modelling, development of drug target databases of *Plasmodium falciparum* and *Mycobacterium tuberculosis*, computational biology, cooperativity in macromolecules, protein-protein interactions, and structure based drug designing. His work has been published in various peer-reviewed journals.