Enhanced ResNet50 Model for Aqueous Solubility Prediction of Drug Compounds Using Deep Learning Techniques

Omprakash Dewangan [1*], Vasani Vaibhav Prakash [1]

Abstract

Background: Aqueous Solubility (AS) is a critical factor in drug discovery (DD), directly influencing a drug's bioavailability and overall efficacy. Accurate prediction of AS remains a challenge despite the advancement in machine learning techniques, which are essential for improving the pharmacokinetics and formulation of new compounds. Methods: This study determines an enhanced ResNet50 deep learning architecture for predicting AS in drug compounds. Deep-net models with 8, 16, and 20-layer ResNet50 Convolutional Neural Network (CNN) architectures were developed. A dataset of 9,532 drug compounds, represented by molecular footprints, was used to train the models. The training process utilized a ten-fold cross-validation technique to optimize the model's predictive performance. Results: The 20-layer ResNet50 model outperformed human experts and shallower models, achieving an R^2 value of 0.423 and an RMSE of 0.678. The model also demonstrated an impressive ASP accuracy rate of 90.6%, significantly surpassing the predictions made by human experts and simpler neural network models. Conclusion: This study demonstrates that deeper-net architectures, particularly the 20-layer ResNet50 model, offer superior performance in predicting AS. These deep learning models provide a reliable and efficient solution for improving solubility predictions, crucial for advancing drug discovery efforts.

Keywords: Aqueous Solubility, Drug Discovery, Prediction, ResNet50, Convolutional Neural Network, Deep Learning.

Significance | This study presents a novel 20-layer ResNet50 model for improving drug solubility predictions, surpassing traditional methods and human expertise.

*Correspondence. Omprakash Dewangan ,Department of CS & IT, Kalinga University, Raipur, India.
E-mail: ku.omprakashdewangan@kalingauniversity.ac.in

Editor Surendar Aravindhan, Ph.D., And accepted by the Editorial Board Sep 01, 2024 (received for review Jul 09, 2024)

1. Introduction

Aqueous solubility (AS) is a crucial physicochemical property of compounds used in anti-cancer drug discovery (DD), significantly influencing their pharmacokinetics and composition. Various Artificial Intelligence (AI) approaches, including machine learning (ML) and deep learning (DL), have been employed to develop tools for AS prediction and evaluation (Gupta et al., 2021, Rutba-Aman et al. 2023, Tanvir et al., 2023).

The drug discovery process is highly complex, and evaluating a compound's ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties is essential. AS plays a vital role in ADMET evaluation (Deore et al., 2019). AS refers to a compound's ability to dissolve in water at a standard temperature of 25°C, which directly impacts its absorption by organisms.

Traditional methods of assessing AS were laborious, as demonstrated by early chemical experiments. In 2000, a new approach to predicting AS using molecular topology was introduced (Huuskonen, 2000). The following year, Tetko et al. (2001) introduced the E-State score method for evaluating water-soluble compounds. The authors also applied ML algorithms, such as multiple-step permutation relevance and Bayesian optimization,

Author Affiliation.
[1] Department of CS & IT, Kalinga University, Raipur, India.

Please cite this article.
Omprakash Dewangan and Vasani Vaibhav Prakash (2024), Enhanced ResNet50 Model for Aqueous Solubility Prediction of Drug Compounds Using Deep Learning Techniques, 8(9), 1-5, 9869

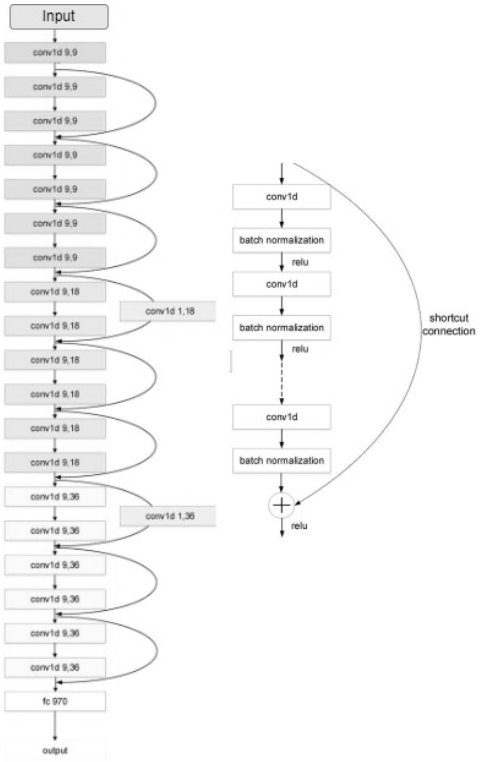

Figure 1. Improved ResNet50 architecture

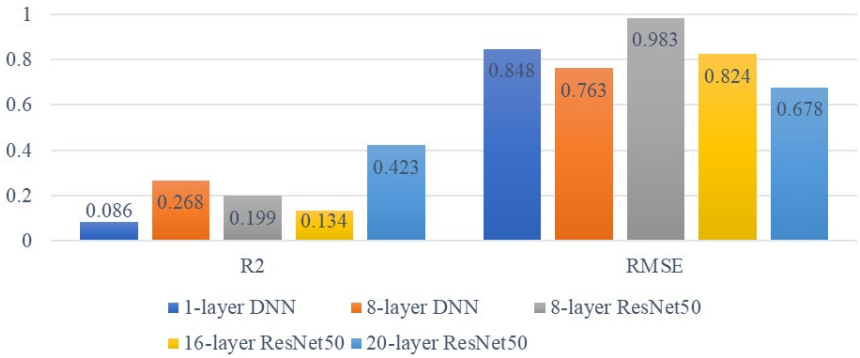

Figure 2. Performance of various DL models for ASP of novel compounds in drugs

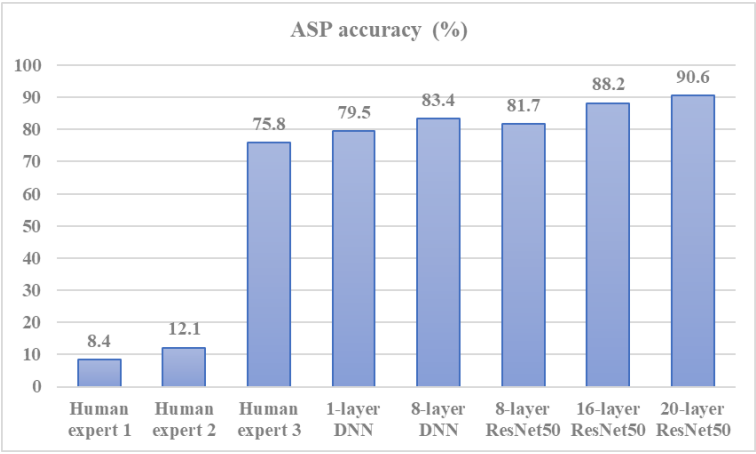

Figure 3. ASP accuracy (%) of various human experts and DL models

to predict AS using chemical structure data (Lovrić et al., 2021).

Recent ML techniques, such as Artificial Neural Networks (ANN), Multi-linear Regression (MR), Support Vector Machines (SVM), and k-nearest neighbors, have been used to study AS (Salman & Banu, 2023; Bennett-Lenane et al., 2022). However, these shallow architectures often struggle with complex functions due to limited samples and computational capacity (Rika et al., 2023). To address issues like overfitting in ANN, researchers have proposed solutions like self-organizing fuzzy neural networks (SOFNN) (Wang & Qiao, 2021). Initial connection weights, which serve as nonlinear mappings from input to output, significantly affect the learning process (Fan & Yang, 2022).

In another notable development, Francoeur and Koes (2021) introduced SolTranNet, a model for predicting AS based on SMILES representations. Despite the common trend of larger models being less effective in this task (Surendar et al., 2024), SolTranNet, with only 3,401 parameters, outperformed linear ML approaches (Sovannarith et al., 2023). Additionally, a new quantitative structure-property relationship (QSPR) framework using a deep neural network (DNN) demonstrated success in predicting the solubility of drug-like molecules in water (P. Vijayakumar et al., 2019).

DL-based AS models generally consist of shallow neural networks with 3 to 7 layers (Gupta et al., 2021). Increasing the depth of these networks often yields better results (Zheng et al., 2021). Convolutional Neural Networks (CNNs), with their ability to learn local features, have shown promise in capturing molecular sub-structures important for AS prediction (Wu et al., 2018). The small number of compounds with empirical AS data constrains the depth of DL, and CNNs provide an effective solution by focusing on local features and substructures (Bobir et al., 2024; Cai et al., 2021).

2. Materials and Methods

2.1 ASP using ResNet50 architecture

The DL models employed the ResNet50 architecture, in which the traditional matrix illustrations of ResNet50 layers, filters, and characteristic maps were substituted with vector forms. The number of levels, represented by N, includes 16, 20 (as depicted in Figure 1), and 26 layers. The network comprises N-1 levels and one fully connected (FC) level. Utilizing vector types was necessary because the inputs consisted of 881-D vectors rather than matrices of image pixels.

The CNN algorithms have been trained using the ten-fold cross-validation technique to create two shallow-net DL ASP models. The cross-validation method randomly partitioned the 9,532 compounds into ten groups of approximately equal sizes. One set was designated a testing database, while the other nine were employed as learning datasets to learn the CNN models. The hyperparameters of the CNN were fine-tuned by evaluating its overall efficacy across ten datasets used for training and testing. The hyperparameters include multiple factors, including loss functions, kernel dimensions, number of filters, stride, number of concealed layers in FC networks, the number of neurons in the FC level, activation operation, optimization, learning rate, normalization, size of batch, and epochs.

An evaluation was conducted on the effectiveness of various activation functions of all forward levels. The weight initialization was performed using a uniform distribution. The implementation involved applying an insignificant quantity of L2 weight decay to achieve L2 normalization.

A DL model, similar to CNN ResNet50, consists of 20 attribute layers. The "$conv1d\ x, y$" refers to a 1D-convolution layer that utilizes x different kernel sizes and y filters. The curvilinear arrows represent the abbreviated links. The shortcut connection, when combined with an attribute layer, rises dimensions. The architectural design of the ResNet50-like DL system includes a shortcut connection. Shortcut connections achieve uniqueness mapping by bypassing one or more levels. Their outputs are incorporated into the levels piled up without additional parameters or are computationally complex.

The effectiveness of the developed DL models in predicting solubility was evaluated using two metrics commonly employed to assess previously developed shallow NN models. One of the values used to assess the asset of a linear association between two variables is the R^2 value, where R represents the correlation coefficient, which is given as:

$$R^2 = 1 - \frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2} \quad (1)$$

The Root Mean Square Error (RMSE) has been given as

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

Where $\hat{y}_i$ is the ASP and y_i is the experimental AS values.

R^2, also known as the determination coefficient, is a symmetrical measure that quantifies the ratio of the variance in the dependent factor that any of the independent variables can explain. The coefficient of determination is a statistical metric employed in a regression model to quantify the degree of fit between the framework and the information. In theory, it signifies a measure of how well something fits, ranging from -∞ to 1. A higher R^2 value specifies a stronger fit between the framework and the information, while a lower R^2 indicates a weaker fit. The alternative metric, $RMSE$, is calculated by taking the square root of the mean of the squared errors. It is a quantity of statistics that quantifies the discrepancies between the ASP values of the archetypal and the actual values. $RMSE$ is a measure of the accuracy of the framework's forecasts. It is always non-negative, meaning it is never less than

zero. A lesser $RMSE$ value shows a better fit to the data, with values closer to zero being the most accurate.

3. Results and Discussion

A total of 9532 compounds were used to create three deep-net models with 8, 16, and 20 layers, using the 10-fold cross-validation approach.

The 60 recently released novel compounds served as a test for our deeper-net models' ASP abilities. As the shallow-net models, we also trained an 8-layer DNN model, a 1-layer DNN model, and an 8-layer ResNet50 framework. Figure 2 contains the testing results for these models. The 20-layer ResNet50 model ($R^2 = 0.423, RMSE = 0.678$) outperformed all other representations, including the well-established tools and the shallow-net structures (R^2 values from $0.268\ to\ 0.199$, $RMSE = 0.763\ to\ 0.983$), according to the R^2 and RMSE values. The bootstrap sampling approach was used to assess the R^2 and RMSE values of four well-known tools, shallow-net and DL models. The 20-layer deeper-net model performs much better than the others. These revealed that DL at the correct depth might significantly improve ASP for novel drug compounds.

Figure 3 depicts the ASP accuracy (%) of various human experts and DL models. The findings indicate that human experts 1 and 2 exhibit shallow levels of ASP accuracy, with rates of 8.4% and 12.1%, respectively. In contrast, human expert 3 demonstrates superior performance, with an accuracy rate of 75.8%. Nevertheless, all DL models surpass the capabilities of human specialists, with even a basic 1-layer DNN having an accuracy of 79.5%. The accuracy of ASP increases as the DL models get more complicated, with the 20-layer ResNet50 model obtaining the greatest accuracy of 90.6%. These findings indicate that DL models, particularly deeper ones such as multi-layer ResNet50, have the potential to surpass human experts in tasks involving the ASP of drugs.

4. Conclusion

In conclusion, this study demonstrates the significant potential of deep learning models, particularly the improved 20-layer ResNet50 architecture, for predicting the aqueous solubility (AS) of drug compounds. By utilizing a deep convolutional neural network trained on 9,532 compounds, the model achieved superior accuracy compared to both traditional methods and human experts, with an impressive R^2 value of 0.423 and an RMSE of 0.678. These results underscore the effectiveness of deeper-net models in drug discovery, as the 20-layer ResNet50 model outperformed shallow models and even expert predictions, achieving an ASP accuracy of 90.6%. This enhanced performance highlights the value of leveraging deep learning techniques to improve drug solubility prediction, ultimately benefiting the drug discovery process by addressing the challenges of bioavailability and pharmacokinetics.

Author contributions

O.D. led the conceptualization, study design, and supervision of the research. V.V.P. contributed to data collection, analysis, and manuscript drafting. Both authors reviewed and approved the final version of the manuscript.

Acknowledgment

Author was grateful to their department.

Competing financial interests

The authors have no conflict of interest.

References

Bennett-Lenane, H., Griffin, B. T., & O'Shea, J. P. (2022). Machine learning methods for prediction of food effects on bioavailability: A comparison of support vector machines and artificial neural networks. European Journal of Pharmaceutical Sciences, 168, 106018.

Bobir, A.O., Askariy, M., Otabek, Y.Y., Nodir, R.K., Rakhima, A., Zukhra, Z.Y., & Sherzod, A.A. (2024). Utilizing deep learning and the internet of things to monitor the health of aquatic ecosystems to conserve biodiversity. Natural and Engineering Sciences, 9(1), 72-83.

Cai, H., Chen, T., Niu, R., & Plaza, A. (2021). Landslide detection using densely connected convolutional networks and environmental conditions. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 5235-5247.

Deore, A. B., Dhumane, J. R., Wagh, R., & Sonawane, R. (2019). The stages of drug discovery and development process. Asian Journal of Pharmaceutical Research and Development, 7(6), 62-67.

Fan, Y., & Yang, W. (2022). A backpropagation learning algorithm with graph regularization for feedforward neural networks. Information Sciences, 607, 263-277.

Francoeur, P. G., & Koes, D. R. (2021). SolTranNet–A machine learning tool for fast aqueous solubility prediction. Journal of Chemical Information and Modeling, 61(6), 2530-2536.

Francoeur, P. G., & Koes, D. R. (2021). SolTranNet–A machine learning tool for fast aqueous solubility prediction. Journal of Chemical Information and Modeling, 61(6), 2530–2536.

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: Machine intelligence approach for drug discovery. Molecular Diversity, 25, 1315-1360.

Huuskonen, J. (2000). Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. Journal of Chemical Information and Computer Sciences, 40(3), 773-777.

Huuskonen, J. (2000). Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. Journal of Chemical Information and Computer Sciences, 40(3), 773–777.

Lovrić, M., Pavlović, K., Žuvela, P., Spataru, A., Lučić, B., Kern, R., & Wong, M. W. (2021). Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? Journal of Chemometrics, 35(7-8), e3349.

P. Vijayakumar, Sivasubramaniyan, G., & Saraswati Rao, M. (2019). Bibliometric analysis of Indian Journal of Nuclear Medicine (2014–2018). Indian Journal of Information Sources and Services, 9(1), 122-127.

Rika, R., Bob, S. R., & Suparni, S. (2023). Comparative analysis of support vector machine and convolutional neural network for malaria parasite classification and feature extraction. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 14(3), 194-217.

Rutba-Aman, Rahnuma Tasmin et al. (2023). Unveiling the Veiled: Leveraging Deep Learning and Network Analysis for De-Anonymization in Social Networks, Journal of Primeasia, 4(1), 1-6, 40042

Salman, R., & Banu, A. A. (2023). DeepQ residue analysis of computer vision dataset using support vector machine. Journal of Internet Services and Information Security, 13(1), 78-84.

Sovannarith, H., Phet, A., & Chakchai, S. (2023). A novel video-on-demand caching scheme using hybrid fuzzy logic least frequency and recently used with support vector machine. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 14(1), 15-36.

Surendar, A., Veerappan, S., Sadulla, S., & Arvinth, N. (2024). Lung cancer segmentation and detection using KMP algorithm. Onkologia i Radioterapia, 18(4).

Tanvir Anjum Labir, Poly Rani Ghosh et al. (2023). Enhancing Emotion Recognition through Deep Learning and Brain-Computer Interface Technology, Journal of Primeasia, 4(1), 1-6, 40046

Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N., & Villa, A. E. (2001). Estimation of aqueous solubility of chemical compounds using E-state indices. Journal of Chemical Information and Computer Sciences, 41(6), 1488-1493.

Wang, G., & Qiao, J. (2021). An efficient self-organizing deep fuzzy neural network for nonlinear system modeling. IEEE Transactions on Fuzzy Systems, 30(7), 2170-2182.

Wu, K., Zhao, Z., Wang, R., & Wei, G. W. (2018). TopP–S: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. Journal of Computational Chemistry, 39(20), 1444-1454.

Zheng, H., Wu, Y., Deng, L., Hu, Y., & Li, G. (2021, May). Going deeper with directly-trained larger spiking neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 12, pp. 11062-11070).