



# Binding Activity Classification of Anti-SARS-CoV-2 Molecules using Deep Learning across Multiple Assays

Bilge Eren Yamasan<sup>1</sup>, Selçuk Korkmaz<sup>2</sup>

<sup>1</sup>Department of Biophysics, Trakya University Faculty of Medicine, Edirne, Türkiye

<sup>2</sup>Department of Biostatistics and Medical Informatics, Trakya University Faculty of Medicine, Edirne, Türkiye

**Background:** The coronavirus disease-2019 (COVID-19) pandemic, caused by severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2), has urgently necessitated effective therapeutic solutions, with a focus on rapidly identifying and classifying potential small-molecule drugs. Given traditional methods' labor-intensive and time-consuming nature, deep learning has emerged as an essential tool for efficiently processing and extracting insights from complex biological data.

**Aims:** This study aims to utilize deep learning techniques, particularly deep neural networks (DNN) enhanced with the synthetic minority oversampling technique (SMOTE), to enhance the classification of binding activities in anti-SARS-CoV-2 molecules across various bioassays.

**Methods:** We used 11 bioassay datasets covering various SARS-CoV-2 interactions and inhibitory mechanisms. These assays ranged from spike-ACE2 protein-protein interaction to ACE2 enzymatic activity and 3CL enzymatic activity. To address the prevalent class imbalance in these datasets, the SMOTE technique was employed to generate new samples for the minority class. In our model-building approach, we divided the dataset into 80% training and 20% test sets, reserving 10% of the training set for validation. Our approach involved employing a DNN that integrates ReLU and sigmoid activation functions, incorporates batch normalization, and uses Adam optimization. The hyperparameters and architecture of the DNN were optimized through various tests on layers, minibatch sizes, epoch sizes, and learning rates. A 40% dropout rate was incorporated to mitigate overfitting. For model evaluation, we computed

performance metrics, such as balanced accuracy (BACC), precision, recall, F1 score, Matthews' correlation coefficient (MCC), and area under the curve (AUC).

**Results:** The performance of the DNN across 11 bioassay test sets revealed varying outcomes, significantly influenced by the ratios of active-to-inactive compounds. Assays, such as AlphaLISA and CoV-PPE, demonstrated robust performance across various metrics, including BACC, precision, recall, and AUC, when configured with more balanced ratios (1:3 and 1:1, respectively). This suggests the effective identification of active compounds in both cases. In contrast, assays with higher imbalance ratios, such as 3CL (1:38) and cytopathic effect (1:15), demonstrated higher recall but lower precision, highlighting challenges in accurately identifying active compounds among numerous inactive compounds. However, even in these challenging settings, the model achieved favorable BACC and recall scores. Overall, the DNN model generally performed well, as indicated by the BACC, MCC, and AUC values, especially when considering the degree of dataset imbalance in each assay.

**Conclusion:** This study demonstrates the significant impact of deep learning, particularly DNN models enhanced with SMOTE, in improving the identification of active compounds in bioassay datasets for COVID-19 drug discovery, outperforming traditional machine learning models. Furthermore, this study highlights the efficacy of advanced computational techniques in addressing high-throughput screening data imbalances.

## INTRODUCTION

The outbreak of the novel coronavirus, severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2), causing coronavirus disease-2019 (COVID-19), has posed unprecedented challenges to global health, triggering an urgent need for effective therapeutic interventions.<sup>1</sup> This scenario underscores the importance of rapidly identifying and developing potential therapeutic agents, particularly

small molecules with promising binding activity against various viral targets. These molecules are important in the fight against the pandemic. Accurate classification of these molecules is pivotal for accelerating drug discovery and development processes. In this critical quest, the role of computational tools, especially machine learning (ML) models, has become increasingly valuable. Offering the potential to significantly accelerate the identification

**Corresponding author:** Selçuk Korkmaz, Department of Biostatistics and Medical Informatics, Trakya University Faculty of Medicine, Edirne, Türkiye

**e-mail:** selcukkorkmaz@trakya.edu.tr

**Received:** January 29, 2024 **Accepted:** February 15, 2024 **Available Online Date:** xxxxxx • **DOI:** 10.4274/balkanmedj.galenos.2024.2024-1-73

Available at [www.balkanmedicaljournal.org](http://www.balkanmedicaljournal.org)

**ORCID iDs of the authors:** B.E.Y. 0000-0002-6525-2503; S.K. 0000-0003-4632-6850.

**Cite this article as:** Yamasan BE, Korkmaz S. Binding Activity Classification of Anti-SARS-CoV-2 Molecules using Deep Learning across Multiple Assays. *Balkan Med J*;

Copyright@Author(s) - Available online at <http://balkanmedicaljournal.org/>

and classification of promising drug candidates, these tools have emerged as key contributors in the battle against COVID-19, enabling researchers to efficiently analyze complex biological data and optimize experimental efforts through the predictive power of advanced algorithms.<sup>2,3</sup>

In antiviral research, traditional methods for classifying the binding activities of small molecules, while effective, often involve labor-intensive and time-consuming experimental procedures.<sup>4</sup> The advent of quantitative high-throughput screening (qHTS) techniques has revolutionized this landscape by generating extensive datasets. However, this abundance of data presents a dual challenge: its complexity and volume often exceed the processing capabilities of conventional analytical methods.<sup>5</sup> ML methods, and more specifically deep learning, have emerged as a potent solution to these challenges.<sup>6,7</sup> Deep learning offers a powerful approach to extracting meaningful insights from complex data by providing robust models that can efficiently handle large datasets and navigate vast chemical spaces and intricate biological interactions.<sup>8</sup> This advancement in the computational analysis is both a response to the immediate challenges posed by the pandemic and a significant leap forward in the field of drug discovery.

Deep learning, also known as deep neural networks (DNN), has demonstrated significant success in various fields, including drug discovery, by offering potent tools to model complex biological interactions.<sup>6</sup> Its application in classifying the binding activities of molecules, especially in the context of antiviral research against SARS-CoV-2, represents a promising avenue for enhancing the screening and analysis of potential antiviral agents.<sup>9</sup> However, the effectiveness of deep learning models is contingent upon their ability to generalize across a range of diverse experimental conditions and assays—a challenge frequently encountered in biological datasets. Our study tackles this challenge by concentrating on the classification of binding activity for anti-SARS-CoV-2 molecules, utilizing deep learning techniques across multiple assays. This methodology is influenced by recent strides in computational drug discovery, as showcased in projects like REDIAL-2020. In this initiative, multiple ML models were applied to predict various activities, encompassing live viral infectivity, viral entry, viral replication, and host cell toxicity.<sup>9</sup>

We used 11 SARS-CoV-2 bioassays from the National Center for Advancing Translational Sciences (NCATS) Open Data Portal for COVID-19 (<https://opendata.ncats.nih.gov/covid19>). These bioassays are derived from qHTS and frequently exhibit an imbalance, typically characterized by a notably larger number of inactive compounds than active ones. This imbalance poses a significant challenge for ML and deep learning algorithms because the limited data for the less represented (minority) class makes learning more difficult.<sup>10</sup> Balancing the datasets through the application of oversampling techniques is one approach to address this challenge.<sup>6</sup> In this study, we utilized the synthetic minority oversampling technique (SMOTE)<sup>11</sup>, a widely used method for oversampling, to establish a balance between active and inactive compounds in each bioassay. Furthermore, we generated molecular descriptors for every compound in each bioassay. Subsequently, we applied the data

balancing method to each dataset and proceeded to train the DNN models on these adjusted datasets. The predictive performance of these models was assessed using a range of metrics.

## METHODS

### *Datasets*

In our study, we used 11 experimental bioassay datasets from the NCATS OpenData Portal, including AlphaLISA and TruHit counter screens for spike-ACE2 protein-protein interaction<sup>12</sup>, ACE2 enzymatic activity<sup>13</sup>, and 3CL enzymatic activity<sup>14</sup> assays targeting SARS-CoV-2 entry and replication. We also employed SARS-CoV-2 cytopathic effect (CPE) and host tox counter screen (cytotox) assays in Vero E6 cells<sup>15</sup>, along with SARS-CoV pseudotyped particle entry (CoV-PPE), its counter screen (CoV-PPE-cs), and similar assays for middle east respiratory syndrome (MERS-PPE and MERS-PPE-cs).<sup>16</sup> Additionally, a human fibroblast toxicity (hCYTOX) assay was included to evaluate compound cytotoxicity.

KC et al.<sup>9</sup> previously employed these bioassays. We used PaDEL software to calculate 1878 molecular descriptors for each compound within each bioassay. Refer to Yap<sup>17</sup> for more comprehensive details on these descriptors. We applied a preprocessing step. First, empty columns, missing values, and variables with zero variance were eliminated. Subsequently, the datasets underwent centering and scaling through z-score transformation. Table 1 summarizes the number of compounds (both before and after preprocessing, categorized as active and inactive compounds), the active-to-inactive ratio, and the number of variables retained post-preprocessing for each bioassay dataset.

### *Data balancing using SMOTE*

SMOTE, a popular oversampling method, generates new synthetic samples for the minority class instead of replicating existing ones.<sup>11</sup> It begins by identifying the k-nearest neighbors (k-NN) of each minority class sample and then synthesizes new samples in the direction of these neighbors. This procedure includes several steps: first, determining the difference between a sample from the minority class and its nearest neighbor; next, multiplying this difference by a randomly chosen number between 0 and 1; then, adding this product to the original sample, thereby generating a new synthetic sample along the line connecting the two samples; and finally, repeating this process for every sample in the minority class. Such a technique expands the decision-making region for the minority class, prompting classifiers to establish broader and more generalized decision regions. This stands in contrast to narrow and overly specific decision regions, ultimately enhancing the overall generalization of the classifier.

### *Deep neural networks*

DNNs, a specialized area of ML, are distinguished by their architecture, which includes several non-linear hidden layers between the input and output layers.<sup>18</sup> The ability of DNNs to decode complex relationships between inputs and outputs is enhanced by their multiple layers and adjustable weights.<sup>6</sup> The training of

**TABLE 1.** Characteristics of the Bioassays Used in the Study.

Assay	Before preprocessing	After preprocessing	Number of active compounds	Number of inactive compounds	Active-to-inactive ratio	Number of remaining variables
3CL	11,400	10,061	260	9,801	1:38	1,330
ACE2	3,250	2,780	174	2,606	1:15	1,046
AlphaLISA	3,250	2,780	729	2,051	1:3	1,046
CoV-PPE	5,099	4,190	2,051	2,139	1:1	1,317
CoV-PPE-cs	5,099	4,190	963	3,227	1:3	1,317
CPE	9,179	8,204	507	7,697	1:15	1,328
cytotox	9,179	8,204	1,410	6,794	1:5	1,328
hCYTOX	4,409	3,849	373	3,476	1:9	1,046
MERS-PPE	2,671	2,245	508	1,737	1:3	1,044
MERS-PPE-cs	2,671	2,245	183	2,062	1:11	1,044
TruHit	3,250	2,780	810	1,970	1:2	1,046

ACE2, angiotensin-converting enzyme 2; PPE-cs, pseudotyped particle entry-counter screen; CPE, cytopathic effect; hCYTOX, human fibroblast toxicity; MERS, middle east respiratory syndrome.

a DNN begins with randomly assigned weights. Subsequently, a loss function is used to evaluate the model's predictions against actual classes, resulting in a loss score. This score is crucial for the backpropagation algorithm, which calculates derivatives of the loss function concerning the weights in each layer using the chain rule.<sup>19</sup> Subsequently, an optimizer, such as a gradient descent algorithm, carefully adjusts the weights based on these derivatives to lower the loss. This iterative process of assessment, calculation, and refinement aims to minimize the loss function, leading to a DNN that is finely tuned to make predictions that closely align with true classifications.

### Model building

For our model, we divided each dataset into an 80% training set and a 20% test set. Additionally, we randomly selected 10% of each training set to serve as validation sets. We used the training sets to train the DNN models and the test sets to assess their predictive performances. The validation sets served two main purposes: first, to fine-tune the hyperparameters of the network, and second, to identify the optimal cut-off points for the predicted probabilities generated by the DNN models. To select optimal cut-off points, we generated a plot illustrating balanced accuracy (BACC) as a function of the predicted probabilities derived from the validation sets. Then, we identified the point at which the DNN attained the highest BACC, deeming it as the optimal cut-off point. This strategy was offered by Korkmaz<sup>6</sup>. We applied the SMOTE method for data balancing in each training set. The ReLU function activated the input and hidden layers, and a sigmoid function was used for the output layer. Batch normalization enhances network performance and stability in each layer. We utilized a binary loss function and the Adam optimization method to tune hyperparameters through five-fold cross-validation. The model architecture was selected based on minimizing cross-validation loss, leading to a three-hidden-layered network. After evaluating various minibatch sizes, we determined 64 to be optimal,

setting the epoch size at 150. A 40% dropout rate was applied to each layer to prevent overfitting, and the learning rate was set at 0.001 as it most effectively minimized cross-validation loss. We used Python v3.7.1 and R v4.3.2 for all analyses. We utilized the Keras (v3.0) and TensorFlow (v2.13) libraries for DNN model building and the imbalanced-learn library to apply the SMOTE method in Python. Receiver operating characteristic (ROC) curves were created using the pROC (v1.18.5) package in R.

### Performance metrics

We computed various performance metrics, including BACC, precision, recall, F1-score, MCC, and AUC, to evaluate the validation and test set performances of the DNN models.

$$\begin{aligned}
 BACC &= \left[ \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right] / 2 \\
 Precision &= \frac{TP}{TP + FP} \\
 Recall &= \frac{TP}{TP + FN} \\
 F1\ Score &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \\
 MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \\
 AUC &= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left( I[Y_{1i} > Y_{0j}] + \frac{1}{2} I[Y_{1i} = Y_{0j}] \right)
 \end{aligned}$$

Here, TP represents true positives, TN denotes true negatives, FP signifies false positives, and FN indicates false negatives. The variable  $i$  iterates over all  $n_0$  data points that are accurately labeled as 1, while  $j$  covers all  $n_1$  data points that are correctly identified as 0.  $Y_{1i}$  refers to the  $i$ -th data point with a true label of 1,  $Y_{0j}$  to the  $j$ -th data point with a true label of 0, and  $I$  represents the indicator function.

These measures are defined in Korkmaz<sup>6</sup>.

## RESULTS

Performance metrics were acquired for each bioassay. The results for the validation set are summarized in Table 2, while the results for the test set are presented in Table 3. Additionally, ROC curves are shown in Figures 1, 2 for both validation and test sets to visualize the classification ability of the DNN model.

The performance of the DNN model across 11 bioassay test sets provides valuable insights, especially when considering the imbalance ratio of active-to-inactive compounds. The subsequent section provides a comprehensive analysis of each assay:

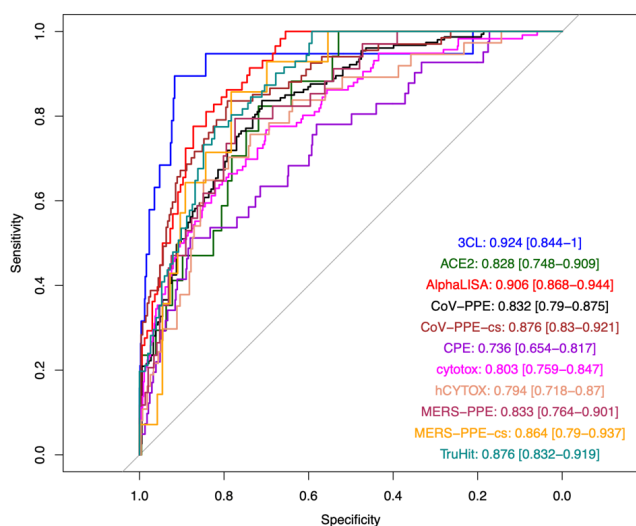
**3CL assay:** With an imbalance ratio of 1:38, this assay presented a substantial challenge. The model achieved a BACC of 0.731, a noteworthy accomplishment given the high level of imbalance. However, the precision was low at 0.075, indicating challenges in accurately identifying active compounds, a likely consequence of the high number of inactive compounds. The recall of 0.700 suggests that the model was sensitive to detecting active compounds. However, this sensitivity came at the cost of precision. The MCC of 0.172 and AUC of 0.783 further reflect these trends.

**ACE2 assay:** With a more manageable imbalance ratio of 1:15, the model showed improved performance with a BACC of 0.772.

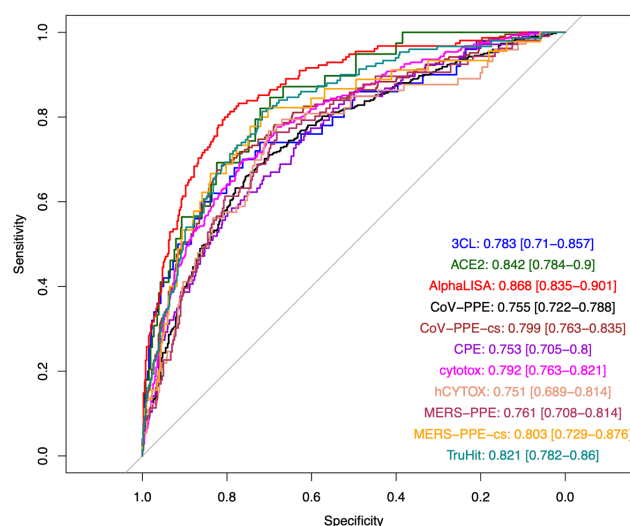
**TABLE 2.** Prediction Metrics of the Deep Neural Network Model for Validation Sets.

Assay	BACC	Precision	Recall	F1-score	MCC	AUC
3CL	0.906	0.218	0.895	0.351	0.418	0.924
ACE2	0.771	0.194	0.824	0.315	0.308	0.828
AlphaLISA	0.828	0.595	0.862	0.704	0.594	0.906
CoV-PPE	0.773	0.707	0.837	0.766	0.546	0.832
CoV-PPE-cs	0.814	0.500	0.836	0.626	0.532	0.876
CPE	0.696	0.221	0.512	0.309	0.270	0.736
cytotox	0.735	0.352	0.776	0.484	0.367	0.803
hCYTOX	0.749	0.369	0.649	0.471	0.396	0.794
MERS-PPE	0.784	0.450	0.794	0.574	0.472	0.833
MERS-PPE-cs	0.820	0.250	0.857	0.387	0.388	0.864
TruHit	0.799	0.671	0.775	0.719	0.577	0.876

BACC, balanced accuracy; MCC, Matthews correlation coefficient; AUC, area under the curve; Precision, positive predictive value; Recall, sensitivity; ACE2, angiotensin-converting enzyme 2; PPE-cs, pseudotyped particle entry-counter screen; CPE, cytopathic effect; hCYTOX; human fibroblast toxicity; MERS, middle east respiratory syndrome.



**FIG. 1.** ROC curves for validation sets of 11 anti-SARS-CoV-2 bioassays (values represent the AUC and corresponding 95% confidence intervals). ROC, receiver operating characteristic; SARS-CoV-2, severe acute respiratory syndrome-coronavirus-2; AUC, area under the curve



**FIG. 2.** ROC curves for the test sets of 11 anti-SARS-CoV-2 bioassays (values represent the AUC and corresponding 95% confidence intervals). ROC, receiver operating characteristic; SARS-CoV-2, severe acute respiratory syndrome-coronavirus-2; AUC, area under the curve

**TABLE 3.** Prediction Metrics of the Deep Neural Network Model for Test Sets.

Assay	BACC	Precision	Recall	F1-score	MCC	AUC
3CL	0.731	0.075	0.700	0.135	0.172	0.783
ACE2	0.772	0.175	0.846	0.289	0.294	0.842
AlphaLISA	0.803	0.603	0.813	0.692	0.561	0.868
CoV-PPE	0.708	0.730	0.659	0.693	0.417	0.755
CoV-PPE-cs	0.750	0.556	0.675	0.610	0.470	0.799
CPE	0.690	0.121	0.764	0.208	0.190	0.753
cytotox	0.730	0.379	0.699	0.491	0.374	0.792
hCYTOX	0.734	0.212	0.767	0.332	0.289	0.751
MERS-PPE	0.723	0.426	0.764	0.547	0.384	0.761
MERS-PPE-cs	0.763	0.236	0.822	0.366	0.331	0.803
TruHit	0.765	0.515	0.813	0.630	0.476	0.821

BACC, balanced accuracy; MCC, Matthews correlation coefficient; AUC, area under the curve; Precision, positive predictive value; Recall, sensitivity; ACE2, angiotensin-converting enzyme 2; PPE-cs, pseudotyped particle entry-counter screen; CPE, cytopathic effect; hCYTOX, human fibroblast toxicity; MERS, middle east respiratory syndrome.

The higher precision of 0.175, compared to 3CL, indicates better identification of true positives, whereas the high recall of 0.846 suggests that the model effectively identified the most active compounds. The F1-score of 0.289 and MCC of 0.294 align with these findings, while an AUC of 0.842 indicates good overall model performance.

**AlphaLISA assay:** This assay had a relatively balanced ratio of 1:3, reflected in strong performance across metrics, including a high BACC of 0.803, precision of 0.603, and recall of 0.813. The F1-score of 0.692 and MCC of 0.561 are particularly noteworthy, suggesting a well-balanced model. The high AUC of 0.868 indicates excellent overall model performance.

**CoV-PPE assay:** Exhibiting a balanced ratio of 1:1, this assay demonstrated robust performance metrics. It achieved the highest precision of 0.730 among all assays, indicating an effective identification of active compounds. However, the recall is slightly lower at 0.659, suggesting that some active compounds were missed. The F1-score of 0.693 and MCC of 0.417, along with an AUC of 0.755, indicate good overall performance.

**CoV-PPE-cs assay:** Similar to AlphaLISA in terms of imbalance (1:3), the model exhibited a balanced performance, with a precision of 0.556 and a recall of 0.675. The BACC of 0.750, F1-score of 0.610, MCC of 0.470, and AUC of 0.799 indicate a model that performs well in both identifying active compounds and avoiding false positives.

**CPE assay:** Despite a challenging imbalance ratio of 1:15, the model achieved a notable recall of 0.764 but struggled with precision (0.121), indicating difficulties in accurately identifying true active compounds among numerous inactive compounds. The BACC of 0.690, F1-score of 0.208, MCC of 0.190, and AUC of 0.753 suggest a model that is more sensitive than specific.

**Cytotox assay:** With an imbalance ratio of 1:5, the model demonstrated moderate precision (0.379) and high recall (0.699), suggesting an effective balance in identifying active compounds.

The BACC of 0.730, F1-score of 0.491, MCC of 0.374, and AUC of 0.792 reflect a well-performing model in a moderately imbalanced setting.

**hCYTOX assay:** Here, the ratio of 1:9 posed a moderate difficulty, and the model managed a high recall (0.767) but a relatively low precision (0.212). The BACC of 0.734, F1-score of 0.332, MCC of 0.289, and AUC of 0.751 indicate a model that effectively identifies active compounds, although with some limitations in precision.

**MERS-PPE assay:** With an imbalance ratio of 1:3, the model performed well, achieving a recall of 0.764 and a precision of 0.426, suggesting an effective balance in identifying active compounds. The BACC of 0.723, F1-score of 0.547, MCC of 0.384, and AUC of 0.761 support the model's good performance.

**MERS-PPE-cs assay:** An imbalance ratio of 1:11 presented challenges. The model exhibited a high recall (0.822) and relatively low precision (0.236), indicating sensitivity to active compounds. The BACC of 0.763, F1-score of 0.366, MCC of 0.331, and AUC of 0.803 indicate a model that is more sensitive than specific.

**TruHit assay:** With a favorable imbalance ratio of 1:2, the model delivered a robust performance, especially notable in recall (0.813) and precision (0.515), demonstrating its effectiveness in discriminating active compounds. The BACC of 0.765, F1-score of 0.630, MCC of 0.476, and AUC of 0.821 reflect a well-balanced and effective model.

Overall, the DNN model performance varied notably across different bioassays and was heavily influenced by the imbalance ratio of each dataset. Assays with more balanced ratios typically demonstrated better precision and overall metric performance. In contrast, assays with higher imbalances tended to have higher recall but struggled with precision. The BACC, MCC, and AUC values across all assays suggest that the model generally performed well, particularly when considering the varying degrees of dataset imbalance.

## DISCUSSION

In drug discovery research, a diverse range of ML algorithms has been effectively employed for the activity classification of compounds. This aims at extracting valuable insights from bioassay datasets. Algorithms, such as Naïve Bayes (NB)<sup>20,21</sup>, Random forests (RF)<sup>22,23</sup>, k-NN<sup>24</sup>, support vector machines (SVM)<sup>7,25</sup>, and DNN<sup>6</sup>, are commonly employed for these classification tasks. A significant challenge in handling bioassay datasets, especially those derived from qHTS, is their inherent class imbalance. This is typically characterized by a substantially higher number of inactive compounds compared to active compounds. This imbalance presents a challenge for ML models due to the limited data available for the minority class. To address this specific challenge, Korkmaz<sup>6</sup> explored various data balancing methods during the training phase of DNNs. The results of this study suggest that the impact of imbalanced data on neural network performance can be alleviated by employing data balancing methods.

In our study, we used a deep learning approach to test the binding activities of 11 anti-SARS-CoV-2 bioassays. We undertook a comprehensive trial-and-error process to optimize the model architecture, experimenting with various numbers of layers, minibatch sizes, epochs, and learning rate levels to identify the most effective hyperparameters. We evaluated the model performance by selecting performance metrics that reflect the imbalanced nature of our datasets, including BACC, F1-score, and MCC.

KC et al.<sup>9</sup> analyzed these bioassays in their research by utilizing a variety of 24 ML models. These models included the following methods: NB, extreme gradient boosting, RF, logistic regression, multilayer perceptron, and SVM, alongside 22 types of molecular descriptors.<sup>9</sup> In contrast, our study focused on developing a deep learning model, employing PaDEL software to generate a comprehensive set of 1878 molecular descriptors. A notable methodological divergence lies in our approach to class imbalance. KC et al.<sup>9</sup> performed random undersampling to balance the active and inactive classes. However, we chose to implement an advanced oversampling technique, SMOTE, which creates synthetic samples for the minority class. Our DNN model exhibited superior performance over the ML classifiers employed by KC et al.,<sup>9</sup> particularly in terms of accuracy and AUC, underscoring its more robust overall capability.

It is crucial to emphasize the significance of discovering new and effective drugs to combat COVID-19, a global health challenge that has had profound impacts on society. The development of such drugs is not only a matter of immediate health concern but also of long-term preparedness for potential future outbreaks. In this context, deep learning methods have emerged as invaluable tools in the drug discovery process. Their ability to analyze complex datasets and uncover patterns that might elude traditional analytical methods offers a promising avenue for identifying novel active molecules. The capacity of deep learning to handle vast and varied data - a common scenario in drug discovery - makes it particularly suitable for rapidly evolving scenarios, such as the COVID-19 pandemic. By incorporating deep learning techniques, researchers can accelerate the identification and validation of potential therapeutic candidates, thereby shortening the time from

laboratory research to clinical trials and eventual public availability. The use of these advanced computational approaches not only enhances the efficiency of the drug discovery process but also increases the likelihood of identifying effective compounds. This, in turn, could save lives and mitigate the impacts of the pandemic.

## CONCLUSION

This study highlights the pivotal role of deep learning in revolutionizing the drug discovery process, especially within the scope of the COVID-19 outbreak. By employing sophisticated deep learning methods, notably the DNN model augmented with SMOTE to address class imbalances, our research demonstrates a marked improvement in identifying active compounds from bioassay datasets. This approach outperforms conventional ML models, as evidenced by comparisons with the work of KC et al.,<sup>9</sup> and also plays a vital role in addressing the challenges arising from the inherently imbalanced characteristics of HTS data. Our results highlight the capability of deep learning to expedite the identification of effective therapeutic agents against COVID-19. Advocating for the broader implementation of advanced computational techniques in drug discovery, this study signals progress toward a more efficient, precise, and swift identification of potential drugs to address present and upcoming health crises.

**Data Sharing Statement:** The data that support the findings of this study are openly available in Kaggle at <https://www.kaggle.com/datasets/selcukorkmaz/anti-sars-cov-2-assays>

**Authorship Contributions:** Concept- B.E.Y., S.K., Design- B.E.Y., S.K., Supervision- S.K., Materials- B.E.Y., S.K., Data Collection or Processing- B.E.Y., S.K., Analysis or Interpretation- B.E.Y., S.K., Literature Search- B.E.Y., S.K., Writing- B.E.Y., S.K., Critical Review- B.E.Y., S.K.

**Conflict of Interest:** No conflict of interest was declared by the authors.

**Funding:** The authors declared that this study received no financial support.

## REFERENCES

- Gül F, Kasapoğlu US, Sabaz MS, et al. The Impact of CoronaVac Vaccination on 28-day Mortality Rate of Critically Ill Patients with COVID-19 in Türkiye. *Balkan Med J*. 2023;40:435-444. [CrossRef]
- Pham TH, Qiu Y, Zeng J, Xie L, Zhang P. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nat Mach Intell*. 2021;3:247-257. [CrossRef]
- Schultz B, DeLong LN, Masny A, et al. A machine learning method for the identification and characterization of novel COVID-19 drug targets. *Sci Rep*. 2023;13:7159. [CrossRef]
- Tabana Y, Babu D, Fahlman R, Siraki AG, Barakat K. Target identification of small molecules: an overview of the current applications in drug discovery. *BMC Biotechnol*. 2023;23:44. [CrossRef]
- Shockley KR. Quantitative high-throughput screening data analysis: challenges and recent advances. *Drug Discovery Today*. 2015;20:296-300. [CrossRef]
- Korkmaz S. Deep Learning-Based Imbalanced Data Classification for Drug Discovery. *J Chem Inf Model*. 2020;60:4180-4190. [CrossRef]
- Korkmaz S, Zararsiz G, Goksuluk D. MLVİS: A Web Tool for Machine Learning-Based Virtual Screening in Early-Phase of Drug Discovery and Development. *PLoS One*. 2015;10:e0124600. [CrossRef]
- Askr H, Elgeldawi E, Aboul Ella H, Elshaiyer YAMM, Gomaa MM, Hassanien AE. Deep learning in drug discovery: an integrative review and future challenges. *Artif Intell Rev*. 2023;56:5975-6037. [CrossRef]

9. KC GB, Bocci G, Verma S, et al. A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nat Mach Intell.* 2021;3:527-535. [\[CrossRef\]](#)
10. Ghosh K, Bellinger C, Corizzo R, Branco P, Krawczyk B, Japkowicz N. The class imbalance problem in deep learning. *Mach Learn.* 2022. [\[CrossRef\]](#)
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321-357. [\[CrossRef\]](#)
12. Hanson QM, Wilson KM, Shen M, et al. Targeting ACE2-RBD Interaction as a Platform for COVID-19 Therapeutics: Development and Drug-Repurposing Screen of an AlphaLISA Proximity Assay. *Acs Pharmacol Transl Sci.* 2020;3:1352-1360. [\[CrossRef\]](#)
13. Hochuli JE, Jain S, Melo-Filho C, et al. Allosteric Binders of ACE2 Are Promising Anti-SARS-CoV-2 Agents. *ACS Pharmacol Transl Sci.* 2022;5:468-478. [\[CrossRef\]](#)
14. Zhu W, Xu M, Chen CZ, et al. Identification of SARS-CoV-2 3CL Protease Inhibitors by a Quantitative High-Throughput Screening. *ACS Pharmacol Transl Sci.* 2020;3:1008-1016. [\[CrossRef\]](#)
15. Chen CZ, Shinn P, Itkin Z, et al. Drug Repurposing Screen for Compounds Inhibiting the Cytopathic Effect of SARS-CoV-2. *Front Pharmacol.* 2021;11:592737. [\[CrossRef\]](#)
16. Chen CZ, Xu M, Pradhan M, et al. Identifying SARS-CoV-2 Entry Inhibitors through Drug Repurposing Screens of SARS-S and MERS-S Pseudotyped Particles. *ACS Pharmacol Transl Sci.* 2020;3:1165-1175. [\[CrossRef\]](#)
17. Yap CW. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J Comput Chem.* 2011;32:1466-1474. [\[CrossRef\]](#)
18. Larochelle H, Bengio Y, Louradour J, Lamblin P. Exploring Strategies for Training Deep Neural Networks. *J Mach Learn Res.* 2009;1:1-40. [\[CrossRef\]](#)
19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436-444. [\[CrossRef\]](#)
20. Fang J, Yang R, Gao L, et al. Predictions of BuChE Inhibitors Using Support Vector Machine and Naive Bayesian Classification Techniques in Drug Discovery. *J Chem Inf Model.* 2013;53:3009-3020. [\[CrossRef\]](#)
21. Sun H. A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J Med Chem.* 2005;48:4031-4039. [\[CrossRef\]](#)
22. Ehrman TM, Barlow DJ, Hylands PJ. Virtual screening of Chinese herbs with Random Forest. *J Chem Inf Model.* 2007;47:264-278. [\[CrossRef\]](#)
23. Plewczynski D, von Grotthuss M, Rychlewski L, Ginalski K. Virtual high throughput screening using combined random forest and flexible docking. *Comb Chem High Throughput Screen.* 2009;12:484-489. [\[CrossRef\]](#)
24. Miller DW. Results of a new classification algorithm combining nearest neighbors and recursive partitioning. *J Chem Inf Comp Sci.* 2001;41:168-175. [\[CrossRef\]](#)
25. Korkmaz S, Zararsiz G, Goksuluk D. Drug/nondrug classification using Support Vector Machines with various feature selection strategies. *Comput Methods Programs Biomed.* 2014;117:51-60. [\[CrossRef\]](#)