big data and cognitive computing

MDPI

Article

Screening of Potential Indonesia Herbal Compounds Based on Multi-Label Classification for 2019 Coronavirus Disease

Aulia Fadli [1], Wisnu Ananta Kusuma [1,2,*], Annisa [1], Irmanida Batubara [2,3] and Rudi Heryanto [2,3]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor 16680, Indonesia; fadliaulia@apps.ipb.ac.id (A.F.); annisa@apps.ipb.ac.id (A.)
[2] Tropical Biopharmaca Research Center, IPB University, Bogor 16680, Indonesia; ime@apps.ipb.ac.id (I.B.); rudi_heryanto@apps.ipb.ac.id (R.H.)
[3] Department of Chemistry, Faculty of Mathematics and Natural Sciences, IPB University, Bogor 16680, Indonesia
* Correspondence: ananta@apps.ipb.ac.id

Abstract: Coronavirus disease 2019 pandemic spreads rapidly and requires an acceleration in the process of drug discovery. Drug repurposing can help accelerate the drug discovery process by identifying new efficacy for approved drugs, and it is considered an efficient and economical approach. Research in drug repurposing can be done by observing the interactions of drug compounds with protein related to a disease (DTI), then predicting the new drug-target interactions. This study conducted multilabel DTI prediction using the stack autoencoder-deep neural network (SAE-DNN) algorithm. Compound features were extracted using PubChem fingerprint, daylight fingerprint, MACCS fingerprint, and circular fingerprint. The results showed that the SAE-DNN model was able to predict DTI in COVID-19 cases with good performance. The SAE-DNN model with a circular fingerprint dataset produced the best average metrics with an accuracy of 0.831, recall of 0.918, precision of 0.888, and F-measure of 0.89. Herbal compounds prediction results using the SAE-DNN model with the circular, daylight, and PubChem fingerprint dataset resulted in 92, 65, and 79 herbal compounds contained in herbal plants in Indonesia respectively.

Keywords: coronavirus disease 2019; drug repurposing; drug-target interaction; health; multilabel classification; stack autoencoder-deep neural network

Citation: Fadli, A.; Kusuma, W.A.; Annisa; Batubara, I.; Heryanto, R. Screening of Potential Indonesia Herbal Compounds Based on Multi-Label Classification for 2019 Coronavirus Disease. *Big Data Cogn. Comput.* **2021**, *5*, 75. https://doi.org/10.3390/bdcc5040075

Academic Editors: Andrea Prati, Carlos A. Iglesias, Luis Javier García Villalba, Vincent A. Cicirello and Min Chen

Received: 7 November 2021
Accepted: 6 December 2021
Published: 9 December 2021

1. Introduction

Coronavirus infectious disease 2019 (COVID-19) is an infectious disease that causes its victim's fever, cough, respiratory problems, pneumonia, and even death [1]. COVID-19 is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2]. On 11 March 2020, WHO declared COVID-19 a pandemic that has affected the world's socio-economic conditions globally. In addition, the highly contagious nature of the SARS-CoV-2 virus also worsened the impact of COVID-19 [3]. As of 9 October 2021, COVID-19 cases reached 237,995,075 cases globally, with the death toll reaching 4,857,370 people (https://www.worldometers.info/coronavirus accessed on 9 October 2021).

The COVID-19 pandemic has spread rapidly and requires an acceleration in the process of drug discovery to fight this disease. Drug repurposing is one of the processes that can help to accelerate the drug discovery process to fight COVID-19 [4]. In drug repurposing, the drug discovery process is conducted by identifying new efficacy for approved drugs, and it is considered an efficient and economical approach [5]. Research in drug repurposing can be done by observing the interactions of drug compounds with protein related to a disease (drug–target interaction or DTI), then predicting the new drug–target interactions [3].

The search for potential drugs through drug repurposing should consider the ease of access to medicinal ingredients to be more accessible by the public, especially for the

people of Indonesia. In this case, exploring herbal compounds can be a good choice in searching for potential drug candidates [6]. Indonesia is one of the countries with the highest number of herbal plant species globally [7]. Herbal compounds provide better efficacy, fewer side effects, and lower prices than conventional compounds [8]. Exploration of herbal compounds candidates can be done by predicting DTI between herbal compounds and proteins of diseases as well as in the drug repurposing process [3].

One of the newest approaches in predicting DTI in drug repurposing is the feature-based chemogenomics approach. The feature-based chemogenomics approach utilizes feature information of drug compounds and diseases proteins represented in a set of descriptors to predict compound–protein interactions [9]. Research [10] used graph approach to create DTI features by combining several information such as drug–drug similarity, drug–disease association, protein–disease association, and protein–protein interaction to construct a heterogenous graph and captures the topological properties of each graph node. However, this method cannot predict the interaction of new drugs or targets. Another method that can be used to create DTI features is using protein descriptors for protein features and molecular fingerprint for compound features. Protein descriptors created protein features by analyzing its amino acid sequences [11] while molecular fingerprint simplifies chemical information in compounds by analyzing the molecular structure into a graph and representing it through binary vectors [12]. Research [13] proposed a method called DeepConv-DTI that applies convolutional neural network (CNN) in predicting binary classification DTI using amino acid composition (AAC) as protein features and circular fingerprints as compound features by analyzing compound's molecule as a graph. Choosing the right type of fingerprint to represent the features of the compound is important in the process of searching for potential drugs [14].

Research related to DTI and the prediction of herbal compounds in COVID-19 cases has previously been carried out by [3,6]. Research [6] conducted binary classification DTI in COVID-19 case using several machine learning models such as SVM, MLP, and Random Forest. The feature extraction process is done in the dataset using PubChem fingerprint for compound features and dipeptide composition (DC) for protein features. This research's results indicated that using SVM, RF, and MLP models gives good results in the prediction of binary classification DTI. The four models were used to predict herbal compounds and resulted in 63 herbal compounds candidate that target 3CLPro, PLPro, and RdRp protein in COVID-19 case. Research [3] conducted binary classification DTI in COVID-19 case using Stacked AutoEncoder Deep Neural Network (SAE-DNN) model with chemogenomics feature-based approach. The feature extraction process is done using PubChem fingerprint for compound data and dipeptide composition (DC), autocorrelation descriptors (ACD), and position-specific scoring matrix (PSSM) for protein data. This research identified that fingerprint for compound features, and DC for protein features gave the best results for predicting DTI with an accuracy of 94%, AUROC of 0.97, and F-measure of 0.82. Herbal compounds prediction results showed that there were 929 interactions between herbal compounds and COVID-19 target proteins.

There are several drawbacks to using binary classification in predicting DTI. First, this method simplifies the DTI problem by modeling high-dimensional compound-protein and their complex associations into a binary classification model without considering the relationship between compounds or proteins [15]. Second, binary classification models tend to have class imbalance problems where positive interaction data is much less than negative interaction data [16]. Therefore, a balance is needed between positive and negative interaction data to produce a good performance. The process of balancing this data is done by taking a random sample of negative interaction data. This process can result in false-negative rates and bias in the model results [15]. In addition, there is a paradigm in DTI where a compound can interact with more than one target protein, and one target protein can interact with more than one compound [17]. Binary classification models tend not to pay attention to possible correlations between labels which may have vital information to increase the accuracy of DTI predictions [18].

Multilabel classification to predict DTI can be used to overcome binary classification problems. In multilabel classification, the training process is conducted to produce a model that maps input vectors to one or more classes. In the multilabel classification for DTI, m compounds are samples, and *n* proteins are target classes. The sample is characterized as an input vector which is then used to predict the target (protein) of the compound with a multilabel learning algorithm [17]. The use of protein as a class label can reduce the dimensions of the input because it no longer requires feature extraction of the protein. Prediction of the target is only determined based on the pattern of the existing compound structure. In addition, from a machine learning perspective, apart from being able to predict several interactions at once, the multilabel classification model can also identify possible correlations between class labels (proteins) to increase the performance of DTI predictions [18]. Research [17] shows that multilabel classification for DTI problems can outperform binary classification with better computational speed, especially for large datasets.

Several studies related to the multilabel classification of DTI have been conducted before. Research [19] conducted a multilabel DTI search using a deep belief network (DBN) model with a binary relevance data transformation approach on protease and kinase data taken from the DUD-E site. Feature extraction on compounds was carried out using the PubChem fingerprint and Klekota-Roth fingerprint descriptors. As a result, the DBN method can be used as a model to predict multi-target DTI with an accuracy range of 97–99% and an AUC range of 83–99%. Research [20] predicted multi-target DTI using the ensemble tree model on the golden standard dataset from research [21]. First, the data are reconstructed using the Neighborhood Regularized Logistic Matrix Factorization (NRLMF) method to overcome the imbalanced data problem. The ensemble tree model with data reconstruction using the proposed NRLMF produces good predictive ability in multi-target DTI.

Research [3,6] that use a binary classification approach for DTI prediction in COVID-19 case simplified DTI problems and could bias the model. Therefore, it is necessary to try using a multilabel classification approach to predict DTI in COVID-19. This study conducted a multilabel classification approach to predict DTI using the SAE-DNN algorithm. SAE is used as a pre-training model for DNN by unsupervised learning. DNN uses an algorithm adaptation approach in solving multilabel classification problems and has good performance in multilabel classification [22]. The feature extraction process was conducted on the compound data using four compound fingerprints: PubChem fingerprint, daylight fingerprint, MACCS fingerprint, and circular fingerprint. The trained SAE-DNN model is then used to predict herbal compounds. Then a search for herbal plants containing predicted herbal compounds was made on the KNapSAcK site (http://www.knapsackfamily.com/KNApSAcK/ accessed on 2 November 2021) [23].

2. Materials and Methods

2.1. Dataset

This study used three datasets: protein dataset, drug-target interaction dataset, and herbal compound dataset. Protein dataset obtained from GeneCards [24], which yielded 1567 genes related to COVID-19. The data are filtered by selecting genes categorized as protein-coding and producing a total of 1498 genes (proteins). Protein data can be seen in Supplementary Spreadsheet 1. The drug–target interaction dataset obtained from the SuperTarget [25] and DrugBank [26] site yielded 58,446 interactions. All of the drug–target interaction data can be seen in Supplementary Spreadsheet 2. Herbal compound data were obtained from HerbalDB [27], which consist of 403 herbal compounds. Data acquisition is made on 12–13 July 2021. Feature extraction on the compound was carried out with four fingerprints with two different types: substructural fingerprint (PubChem fingerprint [28] and MACCS fingerprint (or referred as MDL keys [29]) and topological fingerprint (daylight fingerprint [30] and circular fingerprint (ECFPs) [31]). In chemoinformatics, a fingerprint is one way to represent the chemical structure [12].

2.2. Workflow

DTI multilabel prediction consists of three steps:

1. Data preprocessing step, which includes feature extraction on compounds and class data transformation.
2. Multilabel modeling step using SAE-DNN model. Hyperparameter tuning is also conducted to find the optimal parameter of SAE-DNN for all feature extraction datasets.
3. Post-processing step including model evaluation and herbal compounds prediction.

2.3. Data Preprocessing

The first stage of data preprocessing is the feature extraction on the compound. Feature extraction aims to form a representation of compounds' chemical structure. One of the most commonly used feature extraction processes to represent compounds' chemical structure is molecular fingerprints. Molecular fingerprinting simplifies chemical information in compounds by analyzing the molecular structure into a graph and representing it through binary vectors [12].

The feature extraction process used four fingerprints with two different types, substructure-based fingerprint (PubChem fingerprint [28] and MACCS (MDL) fingerprint [29]) and topological fingerprint (daylight fingerprint [30] and circular fingerprint (ECFPs) [31]). In a substructure-based fingerprint, an array is formed to represent the chemical substructure of a compound, with each substructure assigned to a specific location in the array. For each substructure that occurs in the compound, the position of the corresponding substructure in the fingerprint vector is 1; otherwise, the position of the substructure is 0 [32]. Topology-based fingerprints are formed by analyzing the number of molecular fragments that emerge from a specified path or radius of a molecule, then each path or radius is encrypted with a hash. The bit value in topology-based fingerprint array is 1 if there is a molecular fragment at a certain path length or radius; otherwise, it is 0 [33].

The illustration of the feature extraction process on compounds can be seen in Figure 1.

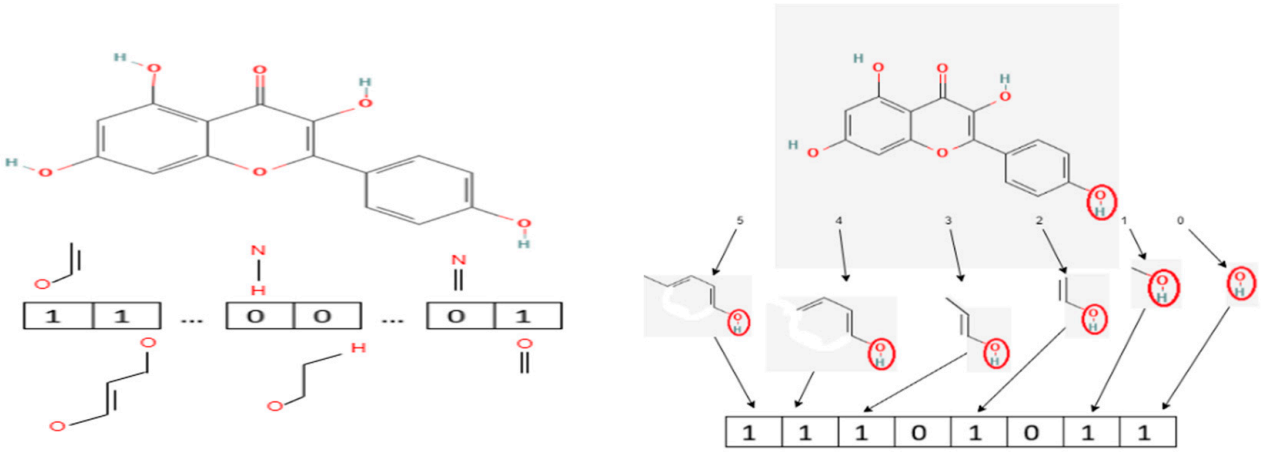

Figure 1. (**a**) Illustration of the substructure-based fingerprint to represent the chemical structure of a compound. Bit 1 indicates that the substructures they represent are present in the molecule. In contrast, bit 0 indicates the substructure is not present in the molecule (**b**) Illustration of the topological-based fingerprint to represent the chemical structure of a compound. In this case, a linear path-based (daylight) fingerprint with a path length of 5. Every fragment found from the starting point (circled) to a certain path length is hashed to the corresponding bit in the fingerprint. Circular fingerprints follow a similar approach, but instead of using path length, it used the radius of the starting point to find the molecule fragments.

The feature extraction process is done using the following steps:

1. Identify each unique compound in the interaction data. There are 49,862 unique compounds in the interaction data
2. Identify PubChem ID of each compound
3. Identify the SMILES (Simplified Molecular-Input Line-Entry System), which represents the chemical structure of each compound. SMILES data can be seen in Supplementary Spreadsheet 3.
4. Form fingerprint of each compound according to the SMILES of each compound.
5. The fingerprint feature retrieval process produces a feature vector C (C = [c_1, c_2, c_3, . . . , c_n] with n = the number of substructures on the fingerprint), which will be used as input to the DNN.

The compound feature extraction process produces 881 attributes for the PubChem fingerprint, 1024 for the daylight fingerprint, 166 for the MACCS fingerprint, and 1024 for the circular fingerprint.

The next step is transforming the class attributes from a single label to a multilabel problem. The transformation is done by creating an array P (P = [p_1, p_2, p_3, . . . , p_m] with m = many proteins). In each data row, the value of p in the P array is one of the compounds in that row that interacts with the p protein and is 0 if the compound in the row does not or is not known to have an interaction with the p protein.

It was transforming class data into a multilabel problem by first identifying the unique protein in the interaction data. There are 467 unique proteins in the interaction data. The formation of multilabel data resulted in 49,862 rows of data representing unique compounds and 27 unique classes representing the set of proteins that interacted with these compounds. The data are dominated by compounds that have only one protein that interacts with the compound. Details of the number of compounds that interact with at least one protein can be seen in Figure 2.

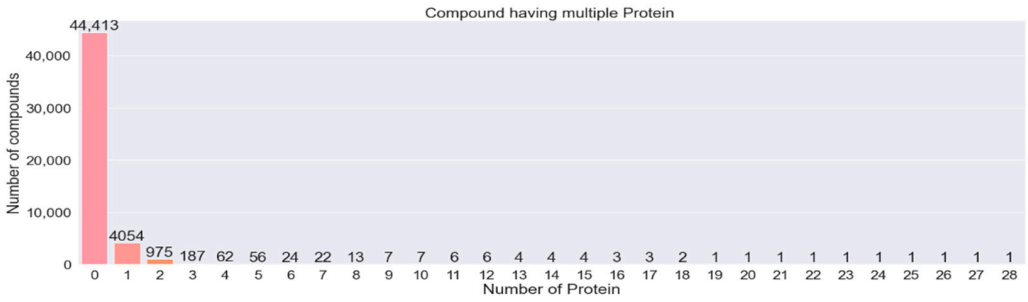

Figure 2. Details of the number of compounds that interact with at least one protein.

The data are then separated into two forms: data table X, compound data containing feature extraction results as predictor variables, and data table Y, a data class in the form of multilabel. An example of all feature extraction data and multilabel class data can be seen in Figure 3.

2.4. SAE-DNN Model

In the SAE-DNN model, SAE is used to pre-trained DNN model to initialize initial weight for DNN. DNN used an algorithm adaptation approach to predict multilabel DTI. Initial weight initialization using SAE training results on DNN modeling is carried out to produce an optimal model compared to a model with random weights [34]. SAE-DNN architecture can be seen in Figure 4. The Pre-training DNN process using SAE can be seen in Algorithm 1 [34].

PubChem Fingerprint [C_1,C_2,...,C_{881}] (49863×881)

Compound ID	C_1	C_2	...	C_{10}	...	C_{881}
145994598	0	0	...	1	...	0
92337	0	0	...	1	...	0
3698	0	0	...	1	...	0
54454	0	0	...	1	...	0
448281	0	0	...	1	...	0
...	...	...	...	...	...	...

MACCS Fingerprint [C_1,C_2,...,C_{166}] (49863×166)

Compound ID	C_1	C_2	...	C_{10}	...	C_{166}
145994598	0	0	...	1	...	0
92337	0	0	...	0	...	0
3698	0	0	...	0	...	0
54454	0	0	...	0	...	0
448281	0	0	...	0	...	0
...	...	...	...	...	...	...

Multilable Class [P_1,P_2,...,P_{467}] (49863×467)

Compound ID	IKBKG	TNF	ACE2	PTGS1	IL6	...
145994598	1	0	0	0	0	...
92337	1	0	0	1	0	...
3698	0	1	0	0	0	...
54454	0	1	0	0	1	...
448281	0	0	1	0	0	...
...	...	...	...	...	...	...

Daylight Fingerprint [C1,C2,...,C1024] (49863×1024)

Compound ID	C_1	C_2	C_3	...	C_{1024}
145994598	0	1	1	...	0
92337	0	1	0	...	0
3698	0	0	1	...	0
54454	0	0	0	...	0
448281	0	1	1	...	0
...	...	...	...	...	...

Circular Fingerprint [C_1,C_2,...,C_{1024}] (49863×1024)

Compound ID	C_1	C_2	C_3	...	C_{1024}
145994598	0	0	0	...	0
92337	0	1	0	...	0
3698	0	0	1	...	0
54454	0	1	0	...	0
448281	0	0	0	...	0
...	...	...	...	...	...

Figure 3. Example of all feature extraction data and multilabel class data.

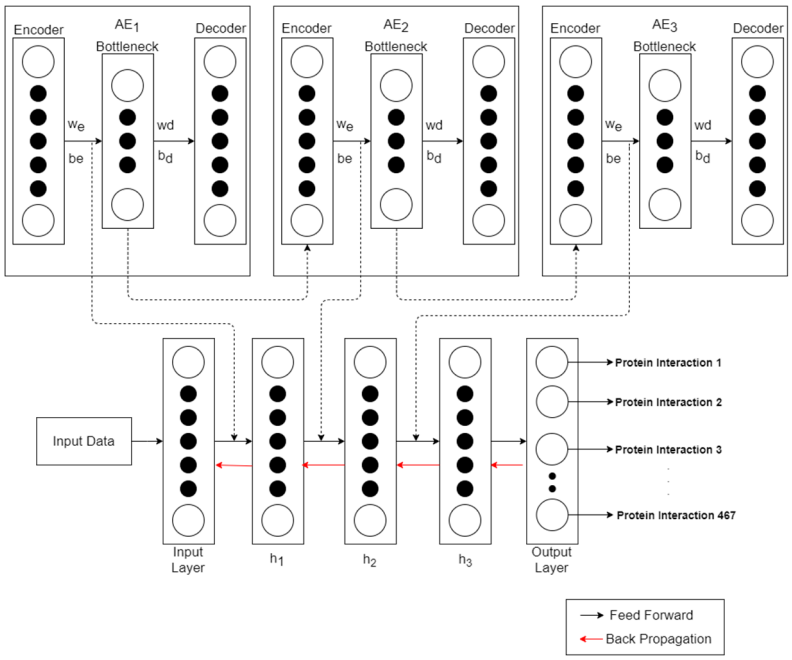

Figure 4. SAE-DNN architecture.

In SAE architecture, the input data (x) that enters the encoder layer will be converted to a new data representation form h which is formulated with Equation (1).

$$h = \varnothing(w_e x + b_e) \tag{1}$$

where $\varnothing$ is activation function, w_e is weight, and b_e is bias at encoder layer. Then the decoder layer will return the new data representation from the bottleneck layer to the initial data form (x') with Equation (2).

$$x' = \varnothing(w_d h + b_d) \tag{2}$$

where $\varnothing$ is activation function, w_d is weight, and b_d is bias at decoder layer.

The AE model is trained to reduce errors in data reconstruction between x' and h. The error metric used is the mean squared error (MSE). In the SAE training process, the new data representation at the current AE bottleneck is used as input for the next AE.

In DNN architecture, the learning process starts from the feed forward process where the input data moves through the existing layers to the output layer. The hidden layer maps the input from the layer to the output to be sent to the next layer according to Equation (3).

$$y_j = f(x_j);\ x_j = \sum_{i=1}^{N} y_i w_{i,j} + b_j \tag{3}$$

where i is the current layer, j is the next layer, y is output from the layer, $f(x_j)$ is activation function, x is input for layer, N is the number of nodes in the layer, $w_{i,j}$ is a weight that connects i-th layer to j-th layer, and b is bias from the layer.

The backpropagation process is then carried out to update the weight and bias values in the DNN architecture to reduce the error value in training. The weight update is carried out using stochastic gradient descent, which aims to optimize the objective function and the learning process of DNN. Equation (4) shows the formulation of the error value, and Equation (5) shows the formulation of the stochastic gradient. Finally, the model training process is repeated until the maximum iteration or results converge to a value.

$$e_j = t_j(n) - y_j(n) \tag{4}$$

$$\Delta w = \alpha \Delta w(t-1) - e\, f'(C) \tag{5}$$

where α is the learning rate, Δw is the change in the weight value, t is the batch size of the data, e is the error, and $f'(C)$ is the cost derivative function used during the backpropagation process to calculate the error gradient.

The pre-training DNN process using SAE can be seen in Algorithm 1 [34].

Algorithm 1 pre-training DNN using Stacked AutoEncoder

INPUT: Feature vector C
OUTPUT: weight (w_e) and bias (b_e) for DNN hidden layers
1. Initialize Max Iteration, N as number of AutoEncoder (AE)
repeat
  2. Train initial AE using C as input
  for i = 1: N **do:**
    3. Save weight (w_e) and bias (b_e) in AE encoder layer
    4. Delete AE decoder layer
    5. Retrieve data representation at AE bottleneck layer
    6. Train next AE using retrieved data representation
  end
until Max Iteration
7. Save weight (w_e) and bias (b_e) on all AE encoder layer

The number of autoencoder layers in SAE is adjusted to the number of hidden layers in DNN. The weights and biases in the DNN hidden layer use the weights and biases from the SAE training, while the weights and biases in the output layer are initialized randomly. After getting the initial weights and biases for the DNN, the next step is to build the DNN model. The DNN adaptation process in solving multilabel classification problems can be done by adjusting the number of nodes in the output layer according to the number of classes (proteins) in the data. Each node in the output layer uses a binary cross-entropy loss function formulated by Equation (6).

$$H_p(P(y)) = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(p(y_i) + (1 - y_i) \log(1 - p(y_i)) \tag{6}$$

DNN architecture applies batch normalization and dropout processes to improve model performance [3]. Batch normalization normalizes the input from each layer by making all inputs have a mean close to 0 and a standard deviation close to 1 to speed up the DNN training process [35]. The dropout process can reduce the complexity of the DNN architecture and prevent overfitting by temporarily eliminating several nodes in the layer randomly during the model training process [36]. After the batch normalization process, the dropout process is carried out to produce a more stable training process, faster convergence, and better generalization [37].

Each node in the output layer will produce a class probability value from the input. If the probability value of the class is above 0.5, then the input class is 1, otherwise, it is 0. The output of the output layer is an array $\hat{P}$ ($\hat{P} = [\hat{p_1}, \hat{p_2}, \hat{p_3}, \ldots, \hat{p_m}]$) which is the prediction result of each node in the output layer. SAE-DNN modeling uses the "TensorFlow-GPU" library version 2.3 and uses GPU to speed up the training process.

The hyperparameter tuning process is carried out using Bayesian Optimization (BO). BO builds a probabilistic model that selects the best hyperparameter from several possible parameters and includes the best hyperparameter to search the other best hyperparameters in the next iteration to speed up the search process for all the best hyperparameters [38]. The implementation of BO is carried out using the "Keras-tuner" library [39]. The hyperparameter search space can be seen in Table 1.

Table 1. The hyperparameter search space.

Hyperparameter	Values
HL_0 Node	100–2000
HL_i Node	$0.5 \times (HL_0)$–$0.75 \times (HL_0)$
Hidden layer	1–6
Optimizer	Adam, adagrad
Learning rate	0.01–0.1
Dropout rate	0.2–0.7

2.5. Postprocess Step

Model evaluation is carried out using iterative stratification, a modification of k-fold cross validation that aims to balance the number of combinations of labels from multilabel data in each fold [40,41] with a total fold of k = 5. The evaluation metrics used are accuracy, recall, precision, and F-measure metrics. The accuracy value measures how well the test data predict. Precision measures the percentage of positive predictions against a positive class. Recall measures the accuracy of the positive prediction of the model. F-measure measures the performance of the minority class [3].

Prediction of herbal compounds is made by predicting the set of proteins (classes) that interact with the data of herbal compounds using the SAE-DNN model with optimal hyperparameters. A set of proteins is considered to interact with herbal compounds if the probability value of the prediction results is above 0.5. Prediction of herbal compounds was carried out using two models, the SAE-DNN model trained with the PubChem fingerprint feature and the best SAE-DNN model from the results of a comparison of four feature extractions. Then a search for herbal plants containing predicted herbal compounds was made on the KNapSAcK site (http://www.knapsackfamily.com/KNApSAcK/ accessed on 2 November 2021).

3. Results

First, we present the comparison between SAE-DNN and DNN model without pre-training (DNN only) for all feature extraction dataset using default parameter such as: HL_0 Node = 1024, HL_i Node = 0.5, hidden layer = 3, Optimizer = Adam, activation function = ReLU, learning rate = 0.01, dropout rate = 0.5. Second, we present the performance comparison for all the feature extraction datasets using optimal hyperparameters. Third, we show the herbal prediction result using the SAE-DNN model.

3.1. Performance Comparison between SAE-DNN and DNN Only

Model performance is presented with the mean and standard deviation of each metric. Figure 5 shows the performance results of SAE-DNN and DNN without a pre-training process for all feature extraction datasets.

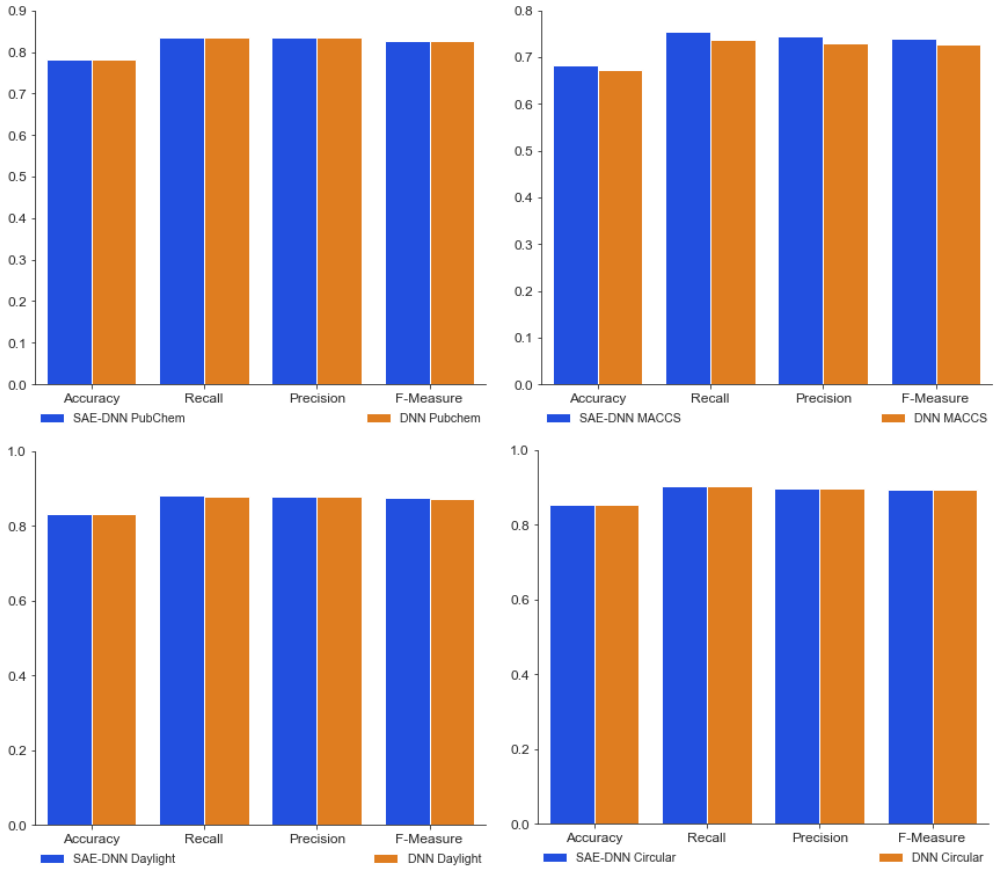

Figure 5. Performance results of SAE-DNN and DNN only for all feature extraction dataset.

SAE-DNN model produces slightly better performance than DNN without pre-training with higher average values of accuracy, recall, precision, and F-measure than the performance of DNN. This indicates that the SAE-DNN model is better able to predict multilabel classes (high accuracy), better at predicting the positive class of each label (high recall), better at predicting positive each label (high precision), and able to recognize minority classes well (high F-measure).

Even though SAE-DNN produces slightly better performance than DNN without pre-training, the use of SAE for DNN pre-training has several advantages, including preventing layer activation outputs from exploding or vanishing during the training of the DL technique [42] and helping DNN achieve better convergence and better generalization power [34]. One way to analyze the generalization performance of learning algorithms is the stability of its prediction performance [43]. Figure 6 shows the standard deviation of SAE-DNN and DNN without pre-training process metrics.

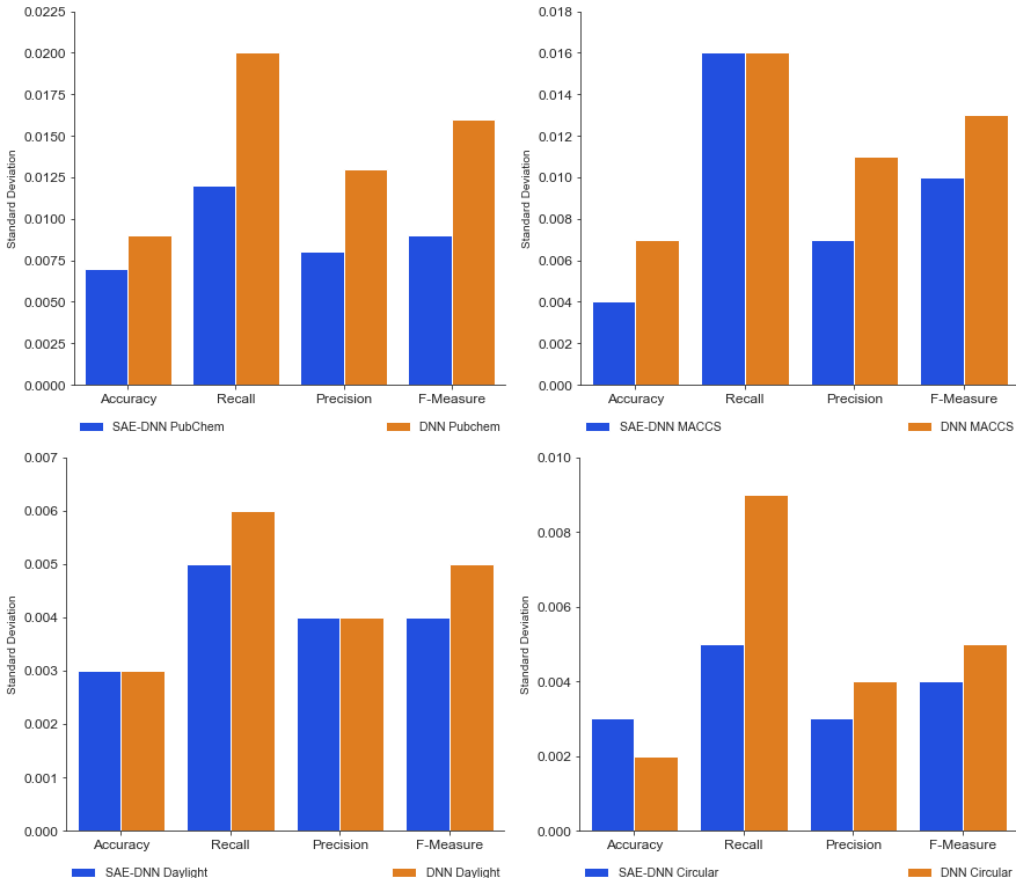

Figure 6. Standard deviation for all metrics in SAE-DNN and DNN without pre-training results.

In general, SAE-DNN produced a lower standard deviation for all metrics compared to DNN without pre-training. A lower standard deviation value indicates that the model's performance is more stable for each fold in the cross-validation process. These results imply that the SAE-DNN model has better generalization power than the DNN without pre-training.

3.2. SAE-DNN Performance Comparison for All the Feature Extraction Datasets

SAE-DNN model is trained using optimal hyperparameters from the hyperparameter tuning process. Optimal hyperparameters can be seen in Table 2.

Table 2. Optimal hyperparameter for SAE-DNN model.

Hyperparameter	Model			
	PubChem	**Daylight**	**MACCS**	**Circular**
HL_0 Node	1500	2000	1024	2000
HL_i Node	0.5	0.5	0.5	0.75
Hidden layer	2	2	3	2
Optimizer	Adam	Adam	Adam	Adam
Learning rate	0.01	0.01	0.01	0.01
Dropout rate	0.5	0.5	0.5	0.5

SAE-DNN comparison results for all the feature extraction datasets can be seen in Table 3.

Table 3. SAE-DNN comparison results for all the feature extraction datasets.

Metrics	Model			
	PubChem	**Daylight**	**MACCS**	**Circular**
Accuracy	0.78747 ± 0.005	0.82814 ± 0.004	0.68272 ± 0.004	**0.83160 ± 0.007**
Recall	0.86178 ± 0.004	0.89306 ± 0.012	0.75462 ± 0.016	**0.91836 ± 0.005**
Precision	0.84641 ± 0.003	0.87854 ± 0.004	0.74407 ± 0.007	**0.88848 ± 0.005**
F-measure	0.84572 ± 0.003	0.87808 ± 0.006	0.74089 ± 0.010	**0.89368 ± 0.005**

Bold indicated the best results.

Topological fingerprints (daylight fingerprint and circular fingerprint) have better performance than substructure-based fingerprints (PubChem fingerprint and MACCS fingerprint). SAE-DNN model with the circular fingerprint feature produces the best average metric value compared to other feature extraction processes, with an accuracy value of 0.83160, recall 0.91836, precision 0.88848, and F-measure 0.89368. The standard deviation of each metric in the circular, daylight, and PubChem models is also relatively low, indicating that the model's performance for each fold tends to be stable. The low performance of the MACCS fingerprint can be assumed due to the lack of explanatory features used, considering that the features extracted from the MACCS feature are only 166. Based on these results, it can be concluded that using the circular fingerprint feature in the SAE-DNN model has the best performance compared to other models. Prediction of herbal compounds was carried out using three SAE-DNN models, namely SAE-DNN with PubChem fingerprint, daylight fingerprint, and circular fingerprint dataset. SAE-DNN model with the MACCS fingerprint dataset was not used to predict herbal compounds due to the low performance of the model.

3.3. Comparison with Other Approaches from the Literature

From a methodological point of view, some recent studies regarding DTI prediction commonly used a binary classification approach. In our proposed method, DTI prediction is done using the multilabel classification approach and takes several advantages over using the binary classification approach. First, the proposed method does not require a process of balancing data between positive data and negative data to achieve fair results, whereas the existing binary classification approach needs to randomly sample the negative DTI in order to balance the data, such as in research [3], which can result in false-negative rates and bias in the model results [15]. Second, the proposed method does not require to include a feature extraction process on protein data which can decrease data dimensions and speed up the training process.

From a machine learning performance point of view, we compare the DTI prediction performance between SAE-DNN and other deep learning models implemented in research [44,45]. Although these studies used a binary approach to predict DTI, comparisons can be made by looking at the model's performance in predicting positive classes. In terms of DTI, only the positive class is considered validated information, while the negative class cannot be validated due to the lack of experimental data on drug–target pairs [46]. Therefore, the comparison is done using recall and f-measure metrics. SAE-DNN outperforms other deep learning such as standard artificial neural network (ANN) and deep belief network (DBN) method from Research [44] with the best f-measure of 0.89368 compared to standard ANN with f-measure of 0.88 and DBN with an f-measure of 0.885. SAE-DNN also outperforms the proposed ComboNet method [45] with the best recall of 0.918 compared to the ComboNet recall of 0.8.

3.4. Herbal Compounds Prediction

The first prediction of herbal compounds was carried out using the SAE-DNN model trained with circular fingerprint datasets. Since the herbal compound data are taken from the HerbalDB website only contains the PubChem fingerprint feature, we must first search for the PubChem ID of each compound. PubChem ID is used to determine SMILES and

look for each herbal compound's daylight fingerprint and circular fingerprint. PubChem ID search for herbal compounds yields 305 herbal compounds that have PubChem ID. Prediction results with the SAE-DNN model with circular fingerprint dataset produced 169 compounds interacting with COVID-19 proteins. Of the 169 compounds, 79 compounds interacted with more than one protein, while the rest interacted with only one protein. Details of the number of herbal compounds that interact with the protein set predicted by SAE-DNN with the circular fingerprint feature can be seen in Figure 7.

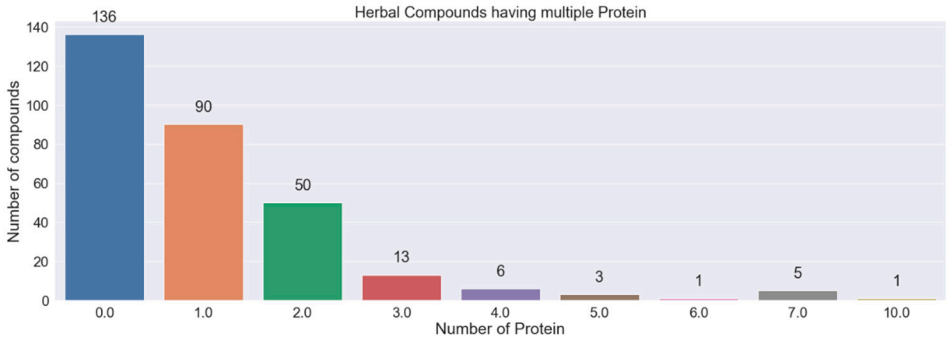

Figure 7. Details of the number of herbal compounds that interact with the protein set predicted by SAE-DNN with circular fingerprint datasets.

This study did not conduct molecular docking on the predicted results of compound-protein interactions from the SAE-DNN model. Potential compounds for COVID-19 can be determined from the relevance score of predicted proteins that interact with these compounds which shows the value of the relevance of the protein to COVID-19 disease on the GeneCards site. The relevance score is calculated based on several factors, including how often the protein appears in a publication and disease pathways. Herbal plants search results show that of the 169 compounds predicted to interact with the COVID-19 protein according to the results of the SAE-DNN prediction with a circular fingerprint dataset; there are 92 compounds contained in herbal plants in Indonesia with a total of 378 herbal plants. Ten compound–protein interactions with the highest relevance score predicted by SAE-DNN with the circular fingerprint dataset can be seen in Table 4. All compound–protein interactions of SAE-DNN prediction results with circular fingerprint dataset can be seen in Supplementary Spreadsheet 4.

Table 4. Ten compound-protein interactions with the highest protein relevance score of SAE-DNN prediction results with circular fingerprint dataset.

Compounds	Proteins	Probability	Protein Relevance Score	Total Herbal Plants	Commonly Used Herbal Plants
Rutin	DPP4, PTGS1	0.534, 0.628	12.445, 0.895	1	*Carmellia sinensis* (tea leaves)
Damnacanthal	TNF	0.819	12.213	2	*Morinda citrifolia* L. (noni)
Ascorbic acid	F3, TLR4, FURIN, PPT1, PVR, PPARA, FADS2	0.998, 0.999, 0.909, 0.994, 0.863, 0.937, 0.969	11.928, 10.370, 4.519, 0.965, 0.965, 0.895, 0.653	5	*Mangifera indica* (mango) *Carica papaya* (papaya)
Palmitoleic acid	F3, TLR4, FURIN, PPT1, PVR, PPARA, FADS2	0.997, 0.999, 0.961, 0.994, 0.915, 0.932, 0.982	11.928, 10.370, 4.519, 0.965, 0.965, 0.895, 0.653	4	*Mangifera indica* (mango) *Punica granatum* (pomegranate)
Petunidin	F3, TLR4, FURIN, PPT1, PVR, PPARA, FADS2	0.997, 0.999, 0.961, 0.994, 0.915, 0.932, 0.982	11.928, 10.370, 4.519, 0.965, 0.965, 0.895, 0.653	1	*Lagerstroemia indica* (crepe-myrtle)
Naringin	F3, TLR4, FURIN, PPT1, PVR, PPARA, FADS2	0.997, 0.999, 0.961, 0.994, 0.915, 0.932, 0.982	11.928, 10.370, 4.519, 0.965, 0.965, 0.895, 0.653	4	*Punica granatum* (pomegranate) *Citrus aurantium* (bitter orange)
Malvidin	F3, TLR4, PPT1, PPARA	0.996, 0.931, 0.883, 0.832	11.928, 10.370, 0.965, 0.895	3	*Impatiens balsamina Melastoma malabathricum*
Sterculic acid	F3, TLR4, PPT1, PPARA	0.996, 0.931, 0.883, 0.832	11.928, 10.370, 0.965, 0.895	2	*Cassia fistula Sterculia foetida*
Ricinoleic acid	F3, TLR4, FADS2, PPARA	0.975, 0.999, 0.999, 0.715	11.928, 10.370, 0.965, 0.895	2	*Ganoderma lucidum Ricinus communis* (ricinus)
P-coumaric acid	F3, FADS2, PPARA, PPT1, ELOVL5	0.983, 0.99, 0.983, 1.0, 0.999	10.370, 0.965, 0.895, 0.653, 0.653	22	*Mangifera indica* (mango) *Punica granatum* (pomegranate)

The scientific names of species are italicized. The genus name is always capitalized and is written first; the specific epithet follows the genus name and is not capitalized.

Based on the results in Table 4. Ten herbal compounds were predicted to have interactions with four proteins with high relevance score, which is DPP4 (12.445), TNF (12.213), TLR4 (11.928), and F3 (10.37) protein. According to research [47], DPP4 protein can be a receptor for SARS-CoV-2 and help the hyper inflammation process in the body. Rutin compound is predicted to interact with DPP4 protein with a probability of 0.534, together with PTGS1 protein with a small relevance value (0.895) and a probability of 0.628. In Indonesia, Rutin compounds are only found in the Carmellia sinensis plant (tea leaves). TNF protein is also a protein that plays a role in the process of forming a cytokine storm in acute COVID-19 patients [48]. It is predicted to interact with the Damnacanthal compound with a probability of 0.819. Damnacanthal compounds are found in the *Morinda citrifolia* L. (noni) plant. Research [49] stated that TLR4 protein could be a promising drug target in COVID-19 cases because it contributes significantly to the pathogenesis of SARS-CoV-2, and its over-activation can cause an exaggerated innate immune response. This protein is predicted to interact with ascorbic acid, palmitoleic acid, petunidin, naringin, malvidin, sterculic acid, and ricinoleic acid compounds with a high probability (0.975–0.998). Among the seven compounds, ascorbic acid compounds are the compounds contained by most herbal plants (5).

Next, predictions of herbal compounds were made using the SAE-DNN model with a daylight fingerprint dataset. The prediction results of the SAE-DNN model with daylight fingerprint dataset resulted in 119 compounds interacting with COVID-19 proteins. Of the 119 compounds, 63 compounds interacted with one protein while the rest interacted with more than one protein. Details of the number of herbal compounds that interact with the protein set predicted by SAE-DNN with the daylight fingerprint feature can be seen in Figure 8.

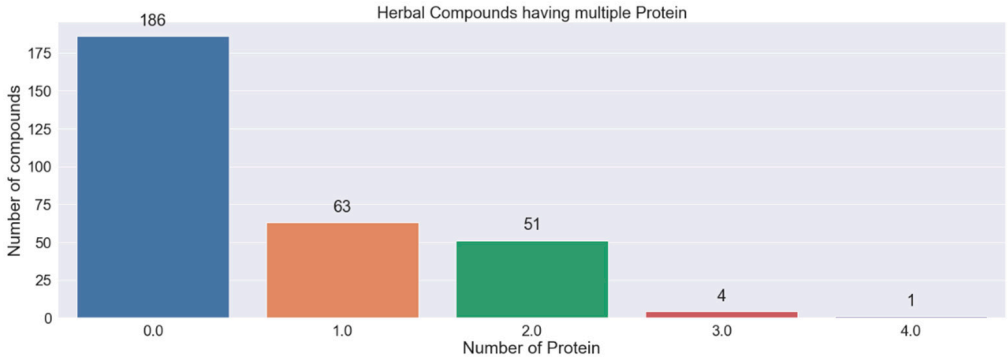

Figure 8. Details of the number of herbal compounds that interact with the protein set predicted by SAE-DNN with daylight fingerprint datasets.

Herbal plants search results at the KNApSAcK site show that of the 119 compounds predicted to interact with the COVID-19 protein according to the results of the SAE-DNN prediction with a daylight fingerprint dataset; there are 65 compounds contained in herbal plants in Indonesia with a total of 272 herbal plants. Ten compound–protein interactions with the highest relevance score predicted by SAE-DNN with daylight fingerprint dataset can be seen in Table 5. All compound–protein interactions of SAE-DNN prediction results with daylight fingerprint dataset can be seen in Supplementary Spreadsheet 5.

Table 5. Ten compound–protein interactions with the highest protein relevance score of SAE-DNN prediction results with daylight fingerprint dataset.

Compounds	Proteins	Probability	Protein Relevance Score	Total Herbal Plants	Commonly Used Herbal Plants
Hyperoside	EGFR	0.509	9.887	7	*Mangifera indica* (mango)
Safrole	TNFRSF1A	0.586	5.459	1	*Cananga odorata*
Estradiol	TNFRSF1A, CSNK2B, EIF3F	0.528, 0.923, 0.626	5.459, 0.965, 0.653	1	*Punica granatum* (pomegranate)
Tetrahydroxyflavone	TNFRSF1A, ALOX5	0.788, 0.537	5.459, 0.895	5	*Cucumis sativus* (cucumber)
Myristic acid	TNFRSF1A, ALOX5, EIF3F	0.912, 0.509, 0.608	5.459, 0.895, 0.653	16	*Mangifera indica* (mango)
Rhamnetin	TNFRSF1A, ALOX5, EIF3F	0.916, 0.526, 0.603	5.459, 0.895, 0.653	6	*Averrhoa carambola* (starfruit)

Table 5. *Cont.*

Compounds	Proteins	Probability	Protein Relevance Score	Total Herbal Plants	Commonly Used Herbal Plants
A-terpinene	TTR	0.564	2.638	20	*Cuminum cyminum* L. (white cumin)
Epicatechin	AR	0.820	2.428	19	*Punica granatum* (pomegranat3)
Proanthocyanidin a2	AR	0.627	2.428	5	*Garcinia mangostana* (mangosteen)
Momordicilin	AR	0.813	2.428	1	*Momordica charantia*

The scientific names of species are italicized.

Based on the results in Table 5, ten herbal compounds with the highest relevance score predicted by the SAE-DNN model with the daylight fingerprint feature interacted with seven proteins with four proteins having the highest relevance score, namely EGFR protein (9.887), TNFRSF1A (5.459), TTR (2.638), and AR (2.428). EGFR protein became the protein with the highest relevance from the prediction results of the SAE-DNN model with daylight fingerprint feature. Research [50] stated that EGFR protein is involved in the infection process in lung cells, triggers a proinflammatory response, and is a potential drug target in the treatment of COVID-19. EGFR protein is predicted to interact with Hyperoside compounds with a probability of 0.509. Hyperoside compounds are found in seven herbal plants in Indonesia, one of which is Mangifera indica (mango). According to the GeneCards website, the TNFRSF1A protein is a variant of the TNF protein and plays a role in inflammatory processes in the body. The TNFRSF1A protein is predicted to interact with safrole, estradiol, tetrahydroxyflavone, myristic acid, and rhamnetin compounds, with the highest probability of rhamnetin compounds being 0.916. Rhamnetin compounds are found in Syzygium aromaticum (cloves) and Averrhoa carambola (starfruit) plants. The direct impact of the TTR protein on COVID-19 is not yet known. According to the GeneCards website, the TTR protein is associated with respiratory failure, which may be one of the effects of COVID-19. According to research [51], AR protein affects the severity of COVID-19 in patients. AR protein regulates the transcription of the TMPRSS2 protein, which is the host for the SARS-CoV-2 spike protein. AR protein is predicted to interact with Epicatechin, Proanthocyanidin a2, and Momordicilin compounds. Epicatechin had the highest probability of interacting with AR protein of 0.82.

Next, predictions of herbal compounds were made using the SAE-DNN model with PubChem fingerprint dataset. The prediction results of the SAE-DNN model with the PubChem fingerprint dataset resulted in 187 compounds predicted to interact with the COVID-19 protein. Of the 187 interactions, 15 compounds interact with two proteins, while 172 interact with only one protein. Details of the number of herbal compounds that interact with the protein set predicted by SAE-DNN with the PubChem fingerprint feature can be seen in Figure 9.

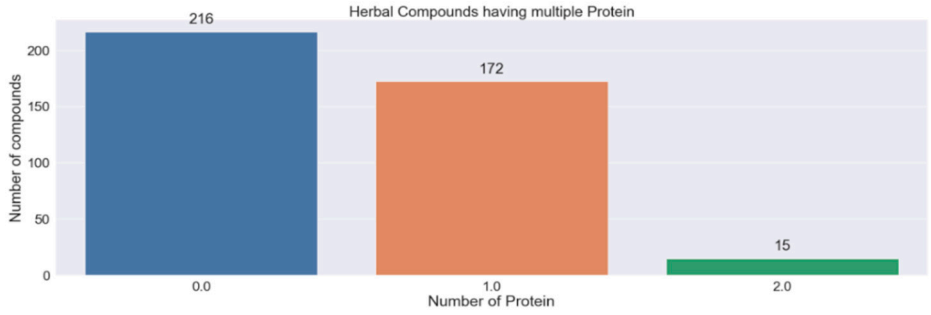

Figure 9. Details of the number of herbal compounds that interact with the protein set predicted by SAE-DNN with PubChem fingerprint datasets.

Herbal plants search results at the KNApSAcK site show that of the 187 compounds predicted to interact with the COVID-19 protein according to the results of the SAE-DNN prediction with a PubChem fingerprint dataset; there are 79 compounds contained in herbal plants in Indonesia with a total of 454 herbal plants. Ten compound-protein interactions with the highest relevance score predicted by SAE-DNN with PubChem fingerprint dataset

can be seen in Table 6. All compound–protein interactions of SAE-DNN prediction results with PubChem fingerprint dataset can be seen in Supplementary Spreadsheet 6.

Table 6. Ten compound–protein interactions with the highest protein relevance score of SAE-DNN prediction results with PubChem fingerprint dataset.

Compounds	Proteins	Probability	Protein Relevance Score	Total Herbal Plants	Commonly Used Herbal Plants
Glucobrassicin	EGFR	0.986	9.887	8	*Raphanus sativus* (radish) *Brassica oleracea* (wild cabbage)
Cuminaldehyde	EGFR	0.923	9.887	4	*Cuminum cyminum* L. (cumin) *Eucalyptus globulus*
P-cymene	EGFR	0.904	9.887	23	*Mangifera indica* (mango) *Nigella sativa* (black cumin)
Methyl cinnamate	EGFR	0.895	9.887	1	*Ocimum* (basil)
Garcimangosone d	EGFR	0.856	9.887	1	*Garcinia mangostana* (mangosteen)
Ethyl cinnamate	EGFR	0.85	9.887	1	*Durio zibethinus* (durian)
Kaempferol	EGFR	0.844	9.887	1	*Carthamus tinctorius* (safflower)
Cinnamic acid	EGFR	0.817	9.887	3	*Glycine max* (soybean) *Ocimum basilicum* (basil)
P-coumaric acid	EGFR	0.808	9.887	22	*Mangifera indica* (mango) *Punica granatum* (pomegranate)
Cinnamaldehyde	EGFR	0.801	9.887	4	*Carica papaya* (papaya) *Pogostemon cablin* (patchouli)

The scientific names of species are italicized.

From the results in Table 6, ten compounds with the highest protein relevance score interacted with the EGFR protein. Glucobrassicin compound has the highest probability to interact with EGFR protein according to the prediction results of SAE-DNN with PubChem fingerprint. This compound is found in eight herbal plants in Indonesia, with one of the commonly used plants being Brassica oleracea (cabbage). Of the ten protein–compound interactions with the highest relevance score predicted by SAE-DNN with PubChem fingerprint, P-cymene compounds are the compounds contained by the most herbal plants, which are 23 herbal plants.

There are similarities in the protein predicted by the three SAE-DNN models, namely AR, EGFR, and PRKCA proteins, with relevance values of 9.887 (EGFR), 2.428 (AR), and 1.003 (PRKCA). The SAE-DNN model with circular fingerprint and PubChem fingerprint both predicts EGFR protein interacting with P-cymene compound, while the SAE-DNN model with daylight fingerprint predicts EGFR protein interacting with Hyperoside compound. For AR proteins, the three models gave different predictions regarding the compounds interacting with these proteins. As for PRKCA protein, the circular and daylight feature SAE-DNN model predicts that this protein interacts with Catalpol and Sinigrin compounds, while the PubChem feature SAE-DNN model predicts this protein interacts with 31 different compounds.

There are also similar compounds that emerged from the prediction results of the three SAE-DNN models, namely Hyperoside, Aloin, Garcimangosone d, Rhamnetin, Anisaldehyde, Laurotetanine, Momordin I, Isoquercetin, and Cycloeucalenone compounds. Details of these compounds can be seen in Table 7.

Table 7. The details of the compounds that appear in the prediction results of the three SAE-DNN models.

Compounds	Protein Predicted by SAE-DNN Model			Species	Activity	References
	Circular	Daylight	PubChem			
Hyperoside	AHR, AKT1	EGFR	PRKCA	*Mangifera indica* (mango)	Served as an anti-inflammatory	[52]
Aloin	MBL2	LGALS3	PRKCA	*Aloe vera*	Indicate to induce anti-inflammatory	[53,54]
Garcimangosone d	RELA, NFKB1	RELA, NFKB1	EGFR	*Garcinia mangostana* (mangosteen)	*Garcinia mangostana* can be used to cure inflammation	[55]
Rhamnetin	CSNK2B, ALOX5, EIF3F	TNFRSF1A, ALOX5, EIF3F	PRKCA	*Averrhoa carambola* (starfruit)	Have good anti-inflammatory activity	[56]
Anisaldehyde	TXNRD1, PLOD1, P4HA1, TFRC, PLOD3, PLOD2	PLOD1, P4HA1, PLOD3, PLOD2	EGFR	*Pimpinella anisum*	-	-
Laurotetanine	ACHE	ADRB2	PRKCA	*Litsea cubeba*	Possess anti-inflammatory properties	[57]
Momordin i	PTGS1, TOP1	TOP1	PRKCA	*Basella rubra* L.	-	-
Isoquercetin	RAC1	RAC1	PRKCA	*Mangifera indica* (mango)	Have anti-inflammatory properties	[58]
Cycloeucalenone	TOP1	TOP1	CSNK2A1	*Musa sapientum*	-	-

The scientific names of species are italicized.

There are several similarities from the prediction results of the SAE-DNN model using daylight fingerprint and circular fingerprint. These results can occur because both fingerprints belong to topology-based fingerprints. The difference between these two fingerprints lies in creating array fingerprints where daylight builds an array of fingerprints based on a specified path of a molecule while circular builds an array based on the specified radius of a molecule [33].

Regarding similar compounds that emerged from the prediction results of the three SAE-DNN models, these compounds were predicted to have interactions with COVID-19 proteins. This is supported by several literature studies showing that Hyperoside, Aloin, Rhamnetin, Laurotetanine, and Isoquercetin compounds have anti-inflammatory properties [52–54,56–58], which can help fight the hyper inflammation process in COVID-19 patients. For Garcimangosone d, its efficacy on COVID-19 or the inflammatory process in the body is not yet known. This compound is found in the *Garcinia mangostana* plant which is a plant commonly used to treat various disease such as inflammation and fever in several Asian countries [55]. Usually, certain databases such as DrugBank and Stitch [59] (http://stitch.embl.de/ accessed on 27 November 2021) can be used to verify compound–protein interaction based on the databases collection of knowledge compounds and their known interactions with proteins. However, the herbal compounds from the prediction results of the three SAE-DNN models in this study have not been found to have interactions with predicted proteins in these databases, thus its compound–protein interactions still cannot be verified. This is due to the lack of experiments related to herbal compounds. Further research is needed to verify compound–protein interactions and determine the potential value of herbal compounds from SAE-DNN prediction results in this study.

4. Conclusions

The multilabel classification approach to search for potential compounds that interact with COVID-19 proteins using the SAE-DNN model has been successfully carried out. The results showed that the SAE-DNN model was able to predict the interaction between drug-target in cases of COVID-19 with a pretty good performance and outperformed DNN without pre-training. The results also show that using the circular fingerprint feature as the predictor variable of the model produces the best average metric value with an accuracy of 0.831, recall 0.918, precision 0.888, and F-measure 0.893.

The prediction results of herbal compounds using the SAE-DNN model with the circular fingerprint dataset resulted in 92 herbal compounds contained in herbal plants in Indonesia, while the prediction results for herbal compounds using the SAE-DNN model with daylight fingerprint dataset resulted in 65 herbal compounds contained in herbal plants in Indonesia and using SAE-DNN model with PubChem fingerprint dataset resulted in 79 compounds contained in herbal plants in Indonesia. Hyperoside, Aloin, Rhamnetin, Laurotetanine, and Isoquercetin are predicted to interact with COVID-19 proteins according to prediction results of the SAE-DNN with circular, daylight, and PubChem dataset and are known to have anti-inflammatory properties regarding several literature studies.

Supplementary Materials: The following are available online at https://ipb.link/supplementary-files, Spreadsheet 1: protein data, Spreadsheet 2: DTI data, Spreadsheet 3: SMILES data, Spreadsheet 4: SAE-DNN with circular fingerprint dataset herbal prediction results, Spreadsheet 5: SAE-DNN with daylight fingerprint dataset herbal prediction results, Spreadsheet 6: SAE-DNN with PubChem fingerprint dataset herbal prediction results.

Author Contributions: Conceptualization, W.A.K.; methodology, A.F., W.A.K. and A.; software, A.F.; validation, W.A.K., I.B. and R.H.; formal analysis, A.F., W.A.K. and A.; investigation, A.F.; data curation, A.F.; writing—original draft preparation, A.F. and W.A.K.; writing—review and editing, A.F., W.A.K. and I.B.; visualization, A.F.; supervision, W.A.K. and A.; project administration, W.A.K.; funding acquisition, W.A.K. and I.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Ministry of Research, Technology and Higher Education, Indonesia, under Competitive Research Grant from Directorate of Higher Education, Indonesia, 2021, grant from Directorate of Higher Education, Indonesia, 2021, contract No. 1/E1/KP. PTNBH/2021 and the APC was funded by IPB University "Agromaritime Institution Research Assignment Grant FY 2021 No 7826/IT3.L1/PT.01.03/P/B/2021".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. Protein data can be found here: https://www.genecards.org/Search/Keyword?queryString=covid-19 accessed on 5 December 2021. DTI data can be found here: http://insilico.charite.de/supertarget/ accessed on 5 December 2021 and https://go.drugbank.com accessed on 5 December 2021. Herbal compound data can be found here: http://herbaldb.farmasi.ui.ac.id/ accessed on 5 December 2021.

Acknowledgments: We thank Ministry of Research, Technology and Higher Education, Indonesia for funding this research and IPB University for funding the APC.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Chen, N.; Zhou, M.; Dong, X.; Qu, J.; Gong, F.; Han, Y.; Qiu, Y.; Wang, J.; Liu, Y.; Wei, Y.; et al. Epidemiological and Clinical Characteristics of 99 Cases of 2019 Novel Coronavirus Pneumonia in Wuhan, China: A Descriptive Study. *Lancet* **2020**, *395*, 507–513. [CrossRef]
2. Gorbalenya, A.E.; Baker, S.C.; Baric, R.S.; de Groot, R.J.; Drosten, C.; Gulyaeva, A.A.; Haagmans, B.L.; Lauber, C.; Leontovich, A.M.; Neuman, B.W.; et al. The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-NCoV and Naming It SARS-CoV-2. *Nat. Microbiol.* **2020**, *5*, 536–544. [CrossRef]
3. Sulistiawan, F.; Kusuma, W.A.; Ramadhanti, N.S.; Tedjo, A. Drug-Target Interaction Prediction in Coronavirus Disease 2019 Case Using Deep Semi-Supervised Learning Model. In Proceedings of the 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 17–18 October 2020; pp. 83–88. [CrossRef]
4. Yadav, M.; Dhagat, S.; Eswari, J.S. Emerging Strategies on in Silico Drug Development against COVID-19: Challenges and Opportunities. *Eur. J. Pharm. Sci.* **2020**, *155*, 105522. [CrossRef]
5. Huang, F.; Zhang, C.; Liu, Q.; Zhao, Y.; Zhang, Y.; Qin, Y.; Li, X.; Li, C.; Zhou, C.-Z.; Jin, N.; et al. Identification of Amitriptyline HCl, Flavin Adenine Dinucleotide, Azacitidine and Calcitriol as Repurposing Drugs for Influenza A H5N1 Virus-Induced Lung Injury. *PLoS Pathog.* **2020**, *16*, 1–16. [CrossRef]
6. Erlina, L.; Paramita, R.I.; Kusuma, W.A.; Fadilah, F.; Tedjo, A.; Pratomo, I.P.; Ramadhanti, N.S.; Nasution, A.K.; Surado, F.K.; Fitriawan, A.; et al. Virtual Screening on Indonesian Herbal Compounds as COVID-19 SupportiveTherapy: Machine Learning and Pharmacophore Modeling Approaches. Available online: https://www.researchsquare.com/article/rs-29119/v1 (accessed on 5 November 2021).
7. Salim, Z.; Munadi, E. *Info Komoditi Tanaman Obat*; Badan Pengkajian dan Pengembangan Perdagangan Kementerian Perdagangan Republik Indonesia: Jakarta, Indonesia, 2017.
8. Ekor, M. The Growing Use of Herbal Medicines: Issues Relating to Adverse Reactions and Challenges in Monitoring Safety. *Front. Pharmacol.* **2013**, *4*. [CrossRef]
9. Larasati, L.; Kusuma, W.A.; Annisa, A. Model Prediksi Interaksi Senyawa Dan Protein Untuk Drug Repositioning Menggunakan Deep Semi-Supervised Learning. *J. Teknol. Inf. Dan Ilmu Komput.* **2020**, *7*, 727. [CrossRef]
10. Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A Network Integration Approach for Drug-Target Interaction Prediction and Computational Drug Repositioning from Heterogeneous Information. *Nat. Commun.* **2017**, *8*. [CrossRef] [PubMed]
11. Ong, S.A.K.; Lin, H.H.; Chen, Y.Z.; Li, Z.R.; Cao, Z. Efficacy of Different Protein Descriptors in Predicting Protein Functional Families. *BMC Bioinform.* **2007**, *8*, 300. [CrossRef]
12. Fernández-De Gortari, E.; García-Jacas, C.R.; Martinez-Mayorga, K.; Medina-Franco, J.L. Database Fingerprint (DFP): An Approach to Represent Molecular Databases. *J. Cheminform.* **2017**, *9*, 9. [CrossRef]
13. Lee, I.I.; Keum, J.; Nam, H.I. DeepConv-DTI: Prediction of Drug-Target Interactions via Deep Learning with Convolution on Protein Sequences. *PLoS Comput. Biol.* **2019**, *15*, e1007129. [CrossRef]
14. Zagidullin, B.; Wang, Z.; Guan, Y.; Pitkänen, E.; Tang, J. Comparative Analysis of Molecular Fingerprints in Prediction of Drug Combination Effects. *Brief. Bioinform.* **2021**, *22*, bbab291. [CrossRef]
15. Mei, S.; Zhang, K. A Multi-Label Learning Framework for Drug Repurposing. *Pharmaceutics* **2019**, *11*, 466. [CrossRef]

16. Mahmud, S.M.H.; Chen, W.; Meng, H.; Jahan, H.; Liu, Y.; Hasan, S.M.M. Prediction of Drug-Target Interaction Based on Protein Features Using Undersampling and Feature Selection Techniques with Boosting. *Anal. Biochem.* **2020**, *589*, 113507. [CrossRef] [PubMed]
17. Chu, Y.; Shan, X.; Chen, T.; Jiang, M.; Wang, Y.; Wang, Q.; Salahub, D.R.; Xiong, Y.; Wei, D.Q. DTI-MLCD: Predicting Drug-Target Interactions Using Multi-Label Learning with Community Detection Method. *Brief. Bioinform.* **2021**, *22*, bbaa205. [CrossRef]
18. Pliakos, K.; Vens, C.; Tsoumakas, G. Predicting Drug-Target Interactions With Multi-Label Classification and Label Partitioning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *18*, 1596–1607. [CrossRef]
19. Fitriawan, A.; Wasito, I.; Syafiandini, A.F.; Amien, M.; Yanuar, A. Multi-Label Classification Using Deep Belief Networks for Virtual Screening of Multi-Target Drug. In Proceedings of the 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA): Recent Progress in Computer, Control, and Informatics for Data Science, Tangerang, Indonesia, 3–5 October 2016; pp. 102–107. [CrossRef]
20. Pliakos, K.; Vens, C. Drug-Target Interaction Prediction with Tree-Ensemble Learning and Output Space Reconstruction. *BMC Bioinform.* **2020**, *21*, 1V. [CrossRef] [PubMed]
21. Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of Drug–Target Interaction Networks from the Integration of Chemical and Genomic Spaces. *Bioinformatics* **2008**, *24*, i232–i240. [CrossRef] [PubMed]
22. Maxwell, A.; Li, R.; Yang, B.; Weng, H.; Ou, A.; Hong, H.; Zhou, Z.; Gong, P.; Zhang, C. Deep Learning Architectures for Multi-Label Classification of Intelligent Health Risk Prediction. *BMC Bioinform.* **2017**, *18*, 121–131. [CrossRef]
23. Afendi, F.M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L.K.; et al. KNApSAcK Family Databases: Integrated Metabolite-Plant Species Databases for Multifaceted Plant Research. *Plant Cell Physiol.* **2011**, *53*, e1. [CrossRef]
24. Stelzer, G.; Rosen, N.; Plaschkes, I.; Zimmerman, S.; Twik, M.; Fishilevich, S.; Stein, T.I.; Nudel, R.; Lieder, I.; Mazor, Y.; et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinform.* **2016**, *2016*, 1.30.1–1.30.33. [CrossRef]
25. Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J.; et al. SuperTarget and Matador: Resources for Exploring Drug-Target Relationships. *Nucleic Acids Res.* **2008**, *36*, D919–D922. [CrossRef]
26. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672. [CrossRef]
27. Yanuar, A.; Mun'im, A.; Bertha, A.; Lagho, A.; Syahdi, R.R.; Rahmat, M.; Suhartanto, H. Medicinal Plants Database and Three Dimensional Structure of the Chemical Compounds from Medicinal Plants in Indonesia. *arXiv* **2011**, arXiv:1111.7183.
28. PubChem Substructure Fingerprint V1.3. Available online: https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt (accessed on 14 July 2021).
29. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280. [CrossRef] [PubMed]
30. Daylight Fingerprints-Screening and Similarity. Available online: https://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed on 15 July 2021).
31. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Modeling* **2010**, *50*, 742–754. [CrossRef]
32. Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **2002**, *29*, 97–101. [CrossRef]
33. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63. [CrossRef]
34. Bahi, M.; Batouche, M. Drug-Target Interaction Prediction in Drug Repositioning Based on Deep Semi-Supervised Learning. *IFIP Adv. Inf. Commun. Technol.* **2018**, *522*, 302–313. [CrossRef]
35. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Available online: http://proceedings.mlr.press/v37/ioffe15.pdf (accessed on 27 August 2021).
36. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958. [CrossRef]
37. Chen, G.; Chen, P.; Shi, Y.; Hsieh, C.-Y.; Liao, B.; Zhang, S. Rethinking the Usage of Batch Normalization and Dropout in the Training of Deep Neural Networks. *arXiv* **2019**, arXiv:1905.05928.
38. Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R.P.; de Freitas, N. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148–175. [CrossRef]
39. O'Malley, T.; Bursztein, E.; Long, J.; Chollet, F.; Jin, H.; Invernizzi, L. Others Keras Tuner 2019. Available online: https://github.com/keras-team/keras-tuner (accessed on 27 August 2021).
40. Szymański, P.; Kajdanowicz, T. A Network Perspective on Stratification of Multi-Label Data. In Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications; 2017. Available online: http://proceedings.mlr.press/v74/szyma%C5%84ski17a.html (accessed on 27 August 2021).
41. Sechidis, K.; Tsoumakas, G.; Vlahavas, I. On the Stratification of Multi-Label Data. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) LNAI*; 2011; Volume 6913, pp. 145–158. Available online: http://lpis.csd.auth.gr/publications/sechidis-ecmlpkdd-2011.pdf (accessed on 27 August 2021). [CrossRef]

42. Boulila, W.; Driss, M.; Al-Sarem, M.; Saeed, F.; Krichen, M. Weight Initialization Techniques for Deep Learning Algorithms in Remote Sensing: Recent Trends and Future Perspectives. *arXiv* **2021**, arXiv:2102.07004.
43. Charles, Z.; Papailiopoulos, D. Stability and Generalization of Learning Algorithms That Converge to Global Optima. Available online: http://proceedings.mlr.press/v80/charles18a/charles18a.pdf (accessed on 19 November 2021).
44. El-Behery, H.; Attia, A.F.; El-Feshawy, N.; Torkey, H. Efficient Machine Learning Model for Predicting Drug-Target Interactions with Case Study for Covid-19. *Comput. Biol. Chem.* **2021**, *93*, 107536. [CrossRef] [PubMed]
45. Jin, W.; Stokes, J.M.; Eastman, R.T.; Itkin, Z.; Zakharov, A.V.; Collins, J.J.; Jaakkola, T.S.; Barzilay, R. Deep Learning Identifies Synergistic Drug Combinations for Treating COVID-19. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2105070118. [CrossRef]
46. Hao, M.; Bryant, S.H.; Wang, Y. A New Chemoinformatics Approach with Improved Strategies for Effective Predictions of Potential Drugs. *J. Cheminform.* **2018**, *10*, 1–9. [CrossRef] [PubMed]
47. Du, H.; Wang, D.W.; Chen, C. The Potential Effects of DPP-4 Inhibitors on Cardiovascular System in COVID-19 Patients. *J. Cell. Mol. Med.* **2020**, *24*, 10274. [CrossRef]
48. Zhang, F.; Mears, J.R.; Shakib, L.; Beynor, J.I.; Shanaj, S.; Korsunsky, I.; Nathan, A.; Accelerating Medicines Partnership Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP RA/SLE) Consortium; Donlin, L.T.; Raychaudhuri, S. IFN-γ and TNF-α Drive a CXCL10+ CCL2+ Macrophage Phenotype Expanded in Severe COVID-19 Lungs and Inflammatory Diseases with Tissue Inflammation. *Genome Med.* **2021**, *13*, 64. [CrossRef]
49. Aboudounya, M.M.; Heads, R.J. COVID-19 and Toll-Like Receptor 4 (TLR4): SARS-CoV-2 May Bind and Activate TLR4 to Increase ACE2 Expression, Facilitating Entry and Causing Hyperinflammation. *Mediat. Inflamm.* **2021**, *2021*, 8874339. [CrossRef]
50. Vagapova, E.R.; Lebedev, T.D.; Prassolov, V.S. Viral Fibrotic Scoring and Drug Screen Based on MAPK Activity Uncovers EGFR as a Key Regulator of COVID-19 Fibrosis. *Sci. Rep.* **2021**, *11*, 11234. [CrossRef] [PubMed]
51. Wambier, C.G.; Goren, A.; Vaño-Galván, S.; Ramos, P.M.; Ossimetha, A.; Nau, G.; Herrera, S.; McCoy, J. Androgen Sensitivity Gateway to COVID-19 Disease Severity. *Drug Dev. Res.* **2020**, *81*, 771–776. [CrossRef]
52. Kim, S.-J.; Um, J.-Y.; Hong, S.-H.; Lee, J.-Y. Anti-Inflammatory Activity of Hyperoside through the Suppression of Nuclear Factor-KB Activation in Mouse Peritoneal Macrophages. *Am. J. Chin. Med.* **2011**, *39*, 171–181. [CrossRef]
53. Ma, Y.; Tang, T.; Sheng, L.; Wang, Z.; Tao, H.; Zhang, Q.; Zhang, Y.; Qi, Z. Aloin Suppresses Lipopolysaccharide-Induced Inflammation by Inhibiting JAK1-STAT1/3 Activation and ROS Production in RAW264.7 Cells. *Int. J. Mol. Med.* **2018**, *42*, 1925–1934. [CrossRef]
54. Park, M.-Y.; Kwon, H.-J.; Sung, M.-K. Evaluation of Aloin and Aloe-Emodin as Anti-Inflammatory Agents in Aloe by Using Murine Macrophages. *Biosci. Biotechnol. Biochem.* **2009**, *73*, 828–832. [CrossRef]
55. Santo, B.L.S.D.E.; Santana, L.F.; Junior, W.H.K.; Araújo, F.D.O.D.; Bogo, D.; Freitas, K.D.C.; Guimarães, R.D.C.A.; Hiane, P.A.; Pott, A.; Filiú, W.F.D.O.; et al. Medicinal Potential of Garcinia Species and Their Compounds. *Molecules* **2020**, *25*, 4513. [CrossRef]
56. Jnawali, H.N.; Lee, E.; Jeong, K.-W.; Shin, A.; Heo, Y.-S.; Kim, Y. Anti-Inflammatory Activity of Rhamnetin and a Model of Its Binding to c-Jun NH 2-Terminal Kinase 1 and P38 MAPK. *J. Nat. Prod.* **2014**, *77*, 258–263. [CrossRef]
57. Xing, X.; Wang, H. Anti-Asthmatic Effect of Laurotetanine Extracted from *Litsea Cubeba* (Lour.) Pers. Root on Ovalbumin-Induced Allergic Asthma Rats, and Elucidation of Its Mechanism of Action. *Trop. J. Pharm. Res.* **2021**, *18*, 1277–1283. [CrossRef]
58. Talasaz, A.H.; Sadeghipour, P.; Aghakouchakzadeh, M.; Kakavand, H.; Ariannejad, H.; Connors, J.M.; Hunt, B.J.; Berger, J.S.; van Tassell, B.W.; Middeldorp, S.; et al. Use of Novel Antithrombotic Agents for COVID-19: Systematic Summary of Ongoing Randomized Controlled Trials. *J. Thromb. Haemost.* **2021**, *19*, 3080–3089. [CrossRef] [PubMed]
59. Kuhn, M.; von Mering, C.; Campillos, M.; Jensen, L.J.; Bork, P. STITCH: Interaction Networks of Chemicals and Proteins. *Nucleic Acids Res.* **2008**, *36*, D684–D688. [CrossRef] [PubMed]