The Implementation of Artificial Neural Network on PIDGINv4

Summer Project Report--Yuchen Wu

1. Introduction

In silico analysis plays an important role in the drug discovery processes. After the target for the disease is identified, pharmaceutical companies will screen millions of molecules against this target in order to find a few drug leads for further investigations. Large-scale screening experiments, which are expensive, may be replaced by virtual screening [1] in the future.

Virtual screening can be divided into two categories: the ligand-based method and the structure-based method. The structure-based method focuses on the complementarity of the compound with the binding site of the target. The ligand-based method focuses on the similarity between active compounds. The improvement of the ligand-based method is the focus of this work.

Machine learning has been widely used in virtual screening [2]. Machine learning models can be trained to identify the common characteristics of active compounds, which are important for target binding. However, there are various challenges that need to be overcome before achieving promising results. The huge dimension of chemical space is problematic when only limited bioactivity data are available, and it is difficult for extrapolation. Also, the data are often very imbalanced because the hit rates in experimental screenings are low. Hence, there is ongoing search for better machine learning implementations.

Random Forests (RF) is one type of machine learning algorithms, and it is used in PIDGINv4 [3], which is a software developed in the Dr. Andreas Bender's group and performs protein target prediction by utilising bioactivity data from both active compounds and inactive compounds.

Artificial Neural Network (ANN), which is another machine learning algorithm, has been known for its flexibility in tuning the model structures and hyperparameters. ANN is a non-linear mapping structure and has the capability of identifying unknown relationship between the input variables and the output observables. Therefore, it is possible that the ANN can identify the linkage between the structural characteristics of compounds and the target activity.

In order to investigate the application of Artificial Neural Network on compound activity predictions, the Random Forests algorithm used by PIDGINv4 is replaced by the Artificial Neural Network algorithm during this summer project.

The implementation details are covered in the Methods section. In the Results and Discussion section, the performances of RF and ANN are compared. The results show that ANN is able to generalise to new types of compounds and has high confidence in its predictions. Compared with RF, ANN achieves a higher recall at the cost of a lower precision.

The code used in this project can be found from https://github.com/andywuy/PIDGIN_with_ANN .

2. Methods

2.1. Data sources

The bioactivity data is extracted from PubChem (extracted Mar 2020) and ChEMBL (version 26). For each protein, four different bioactivity cut-offs (100 μM, 10 μM, 1 μM and 0.1 μM) are used. Each model contains the bioactivity data of compounds for a specific protein target at a specific bioactivity cut-off level. There are 11,782 models in PIDGINv4, and the bioactivity dataset used by PIDGINv4 can be obtained from https://pidginv4.readthedocs.io/en/latest/install.html.

For this summer project, ANN and RF algorithms are only implemented on 1000 randomly selected models. The list of the models used could be found from https://github.com/andywuy/PIDGIN_with_ANN .

2.2. Algorithms

The training input is the array of 2048bit RDKit [4] Extended Connectivity Fingerprints of chemical compounds in the model.

The RF algorithm is implemented using the Scikit-learn [5] Random Forest module, and the number of trees used in the Random Forests is optimized for each model.

The ANN algorithm is implemented using Tensorflow2 [6]. During the training for each model, the number of epochs is 100, the batch size is 128, the loss function is binary cross-entropy, the optimizer is the Adam optimizer, and the early stopping patience is set to 3. The neural network consists of two hidden layers (ten neurons in the first layer and 100 neurons in the second layer) and an output layer (one neuron). The Relu activation is used for the hidden layers and the Sigmoid activation is used for the output layer.

For both algorithms, the training output is the predicted target activities for compounds in the model. Zero indicates that the compound is inactive while one indicates the compound is active.

2.3. Performances Evaluation

The training data in each model is split into a training set and a test set with ratio 2:1. The performances metrics are calculated from the test set results.

For cross validation, the compounds in each model are grouped by Murcko scaffolds. The GroupShuffleSplit (number of splits = 4, train size = 0.5) is used and the area under the precision-recall curve (PRAUC) is calculated for each model.

The conformal prediction measures how confident the prediction of compound activity is, and the confidence interval is set to 95 %.

3. Results and Discussion

In this section, four performances metrics (the weighted accuracy, the Matthews Correlation Coefficient (MCC), the precision and the recall) are compared between RF and ANN. The cross validation and conformal prediction results are also analysed. The metrics are summarised in Table 1.

Metric	ANN mean	ANN std	RF mean	RF std	p value	% improve
Weighted Accuracy	0.936	0.105	0.887	0.135	3.54e-19	65.6
MCC	0.744	0.229	0.867	0.245	1.22e-14	31.7
Precision	0.666	0.268	0.927	0.188	6.89e-120	3.7
Recall	0.895	0.203	0.776	0.117	1.71e-27	66.0

Table 1: The mean and the standard deviation for different metrics across the 1000 models. The two-tail test calculates the p value for the results from ANN and RF. The % improve shows the

proportion of models whose metric value is increased by the ANN compared to RF.

3.1. Weighted Accuracy

Because the hit rates in the screening experiments are low, there are fewer active compound data than inactive compound data. To adjust the imbalance between the active class and the inactive class, the sample weight is set to be inversely proportional to class frequencies in the input data. The weighted accuracy scores are calculated for both ANN and RF. According to Table 1 and Figure 1, the ANN algorithm improves the overall weighted accuracy of the predictions, and the mean value is increased by 5.5 %. This change may look small at first glance. However, given the fact that the score from RF is already high (0.887), this improvement is substantial. The p value also indicates that there is a statistical difference between the mean values from ANN and RF.

According to Figure 2, for more than 60 % of the models, their weighted accuracies are increased after using ANN. Moreover, the extent of increase is greater than the extent of decrease (indicated by the bar height). Therefore, it explains the increase in weighted accuracy.

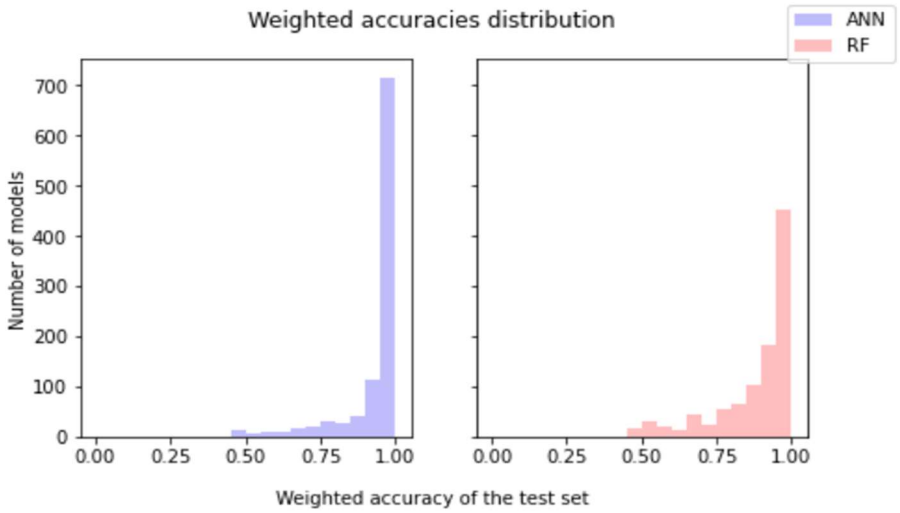

Figure 1: The weighted accuracy distribution for the 1000 models when comparing RF to ANN. The ANN achieves a higher weighted accuracy.

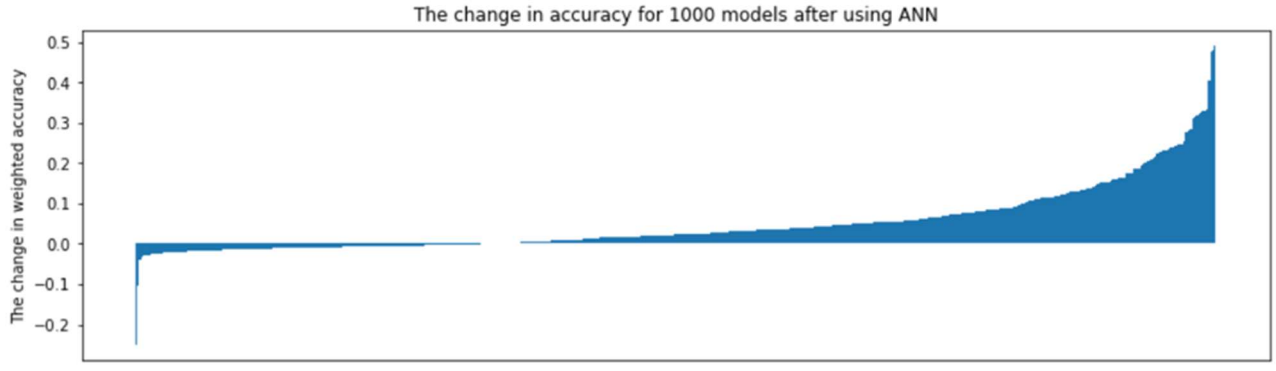

Figure 2: The bar chart for the change in weighted accuracy for each model after replacing RF with ANN. Each model is represented by a bar on the horizontal axis.

3.2. Matthews Correlation Coefficient

The Matthews Correlation Coefficient takes all four elements in the confusion matrix into account.

Although there is no perfect way to describe the confusion matrix by a single number, the MCC is generally regarded as a balanced measure and it can be used even if the classes are of very different sizes [7].

According to Table 1 and Figure 3, the ANN algorithm decreases the MCC score in general. The increase in weighted accuracy and the decrease in MCC suggests that neither the RF nor the ANN performed well in all aspects.

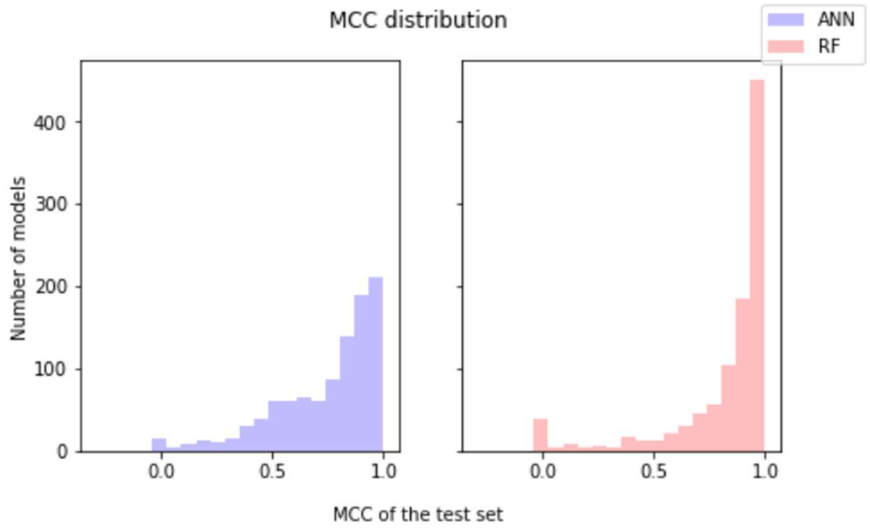

Figure 3: The MCC distribution for the 1000 models when comparing RF to ANN. The ANN achieves a lower MCC overall.

3.3. Precision and Recall

A high precision is desired when the objective is to minimise false positives, while a high recall is desired when the objective is to minimise false negatives. According to Table 1 and Figure 4, the ANN algorithm achieves a higher recall at the cost of lower precision for activity prediction. Besides, the decrease in precision is more significant than the increase in recall.

The precision-recall curve depicts the trade-off between the precision and the recall. The precision and the recall are calculated at different classification thresholds. A good model should be able to obtain a high recall and a high precision at the same time. The Receiver Operating Characteristics (ROC) curve is another tool to show the performances at different classification thresholds. It plots the true positive rate against the false positive rate. However, when the total real negatives are huge, the false positive rate (false positives / total real negatives) does not drop drastically. Therefore, it cannot be used when the dataset is highly imbalanced. Hence here, only the precision-recall curve is shown.

The precision against recall curve for each model is calculated. The averaged curve for 1000 models is shown in Figure 5. A good model should give a curve that is bending towards the top-right corner, where the recall and the precision are both one. According to the plot, it is clear that only when the recall is less than 0.87 or when the recall is greater than 0.96, ANN outperforms RF.

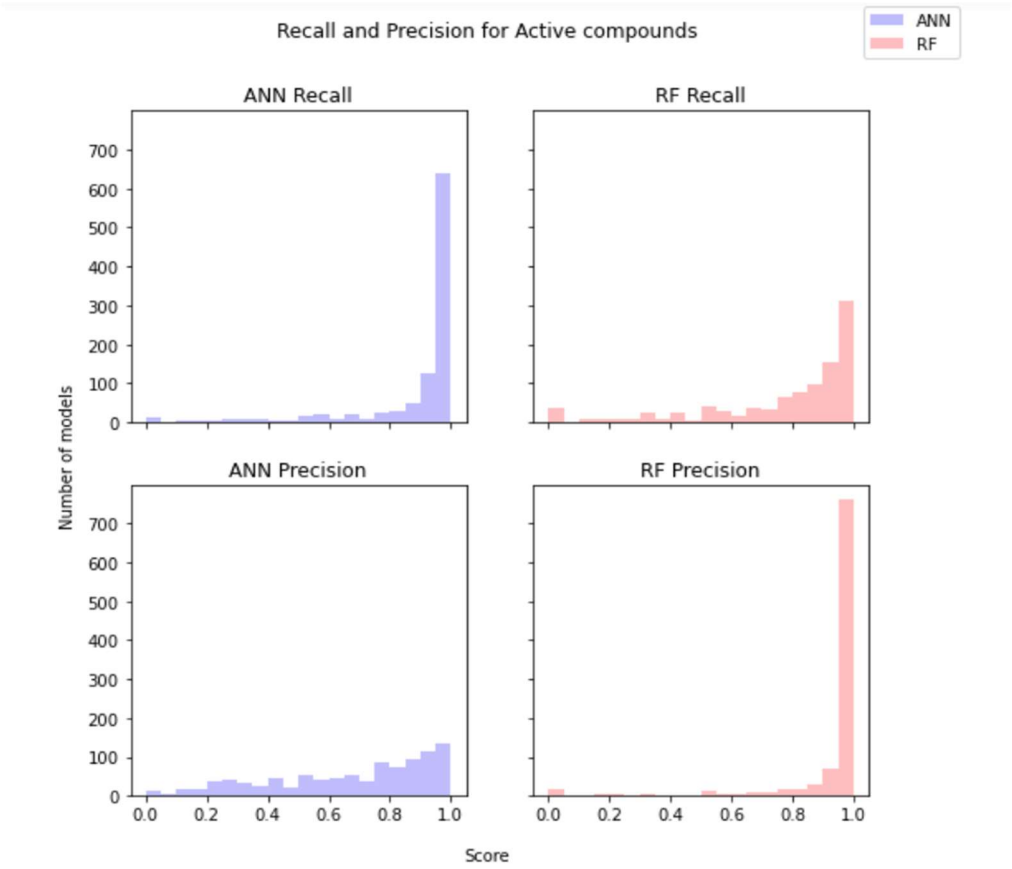

Figure 4: The precision and the recall distribution for the 1000 models when comparing RF to ANN. ANN achieves a higher recall at the cost of lower precision for activity predictions.

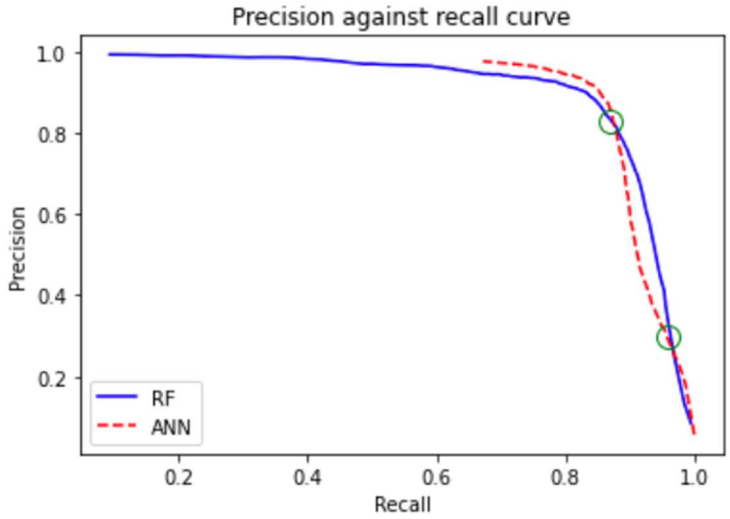

Figure 5: Averaged precision against recall curve for 1000 models. The green circles highlighted where the curves for ANN and RF cross each other.

3.4. Cross Validation

Cross validation tests the effectiveness of a model by resampling the data and ensures that every observation from the dataset has the opportunity to be included in the training set and the test set. The compounds are grouped by Murcko scaffolds, which means that similar compounds are grouped together, and the test set and the training set should include different groups of compounds.

It enables the test of the generalisation ability of the model. This ability is important in the real-world application, because the chemical compounds in the training set is just a small subset of the chemical space.

According to Figure 6, both ANN and RF perform well during the cross validation as they both achieve high averaged PRAUC scores for most models. It means that they are able to give an accurate prediction even if a new type of compound is given.

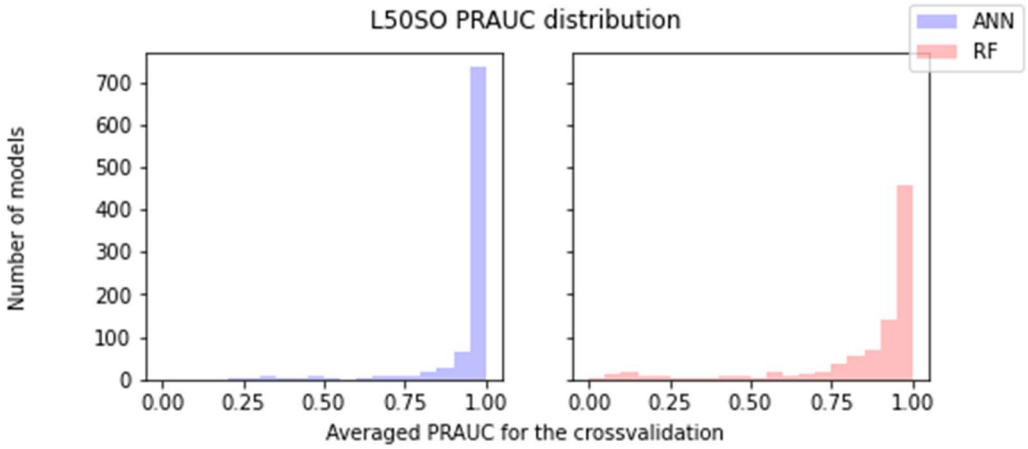

Figure 6: The distribution of PRAUC from "leaving 50% scaffolds out" cross validation.

3.5. Conformal Prediction

It is possible that the model gives a prediction with low confidence (i.e. the input could be classified to either of the two classes). Conformal prediction determines the confidence of the prediction made by the model. One-class-rate, which is defined as the proportion of predictions that are assigned to one class only within the confidence level, acts as a proxy to measure how confident the model is when making predictions.

Figure 7 showed that both ANN and RF achieve high one-class-rate (the mean value is greater than 0.95) and are very confident in predictions for most compounds. This high confidence improves the reliability of virtual screening results and may reduce the failure rate in the clinical trials.

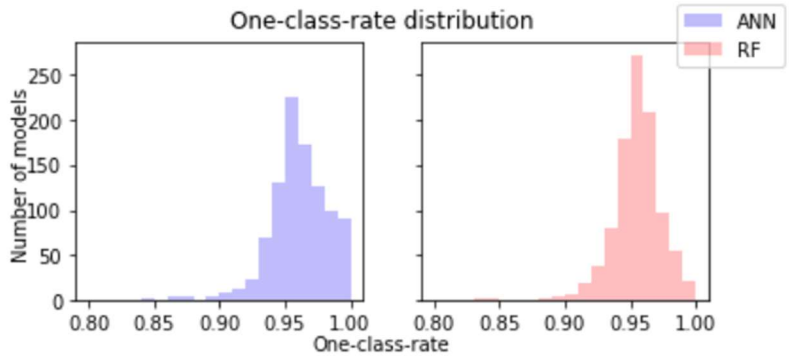

Figure 7: The distribution of the one-class-rate for ANN and RF at 95% confidence level.

4. Conclusions

The performances of Random Forests and Artificial Neural Network are compared in this summer project. The weighted accuracy and the Matthew Correlation Coefficient are used as proxies to

measure the overall performances. Although ANN achieves a higher weighted accuracy, it obtains a lower MCC. Furthermore, the precision and recall results are investigated, which shows that the ANN performs better in recalling active compounds at the cost of lower precision. Because in-silico screening aims to reduce the failure rate in the subsequent experiments, it is advantageous to have a higher precision than having a higher recall. The precision-recall curve (Figure 5) shows that at the high precision limit, the ANN outperforms RF by having a higher recall when keeping the precision the same.

The cross validation shows that both ANN and RF are able to generalise to different types of compounds. This property is important when applying the model to a wider chemical space.

The conformal prediction results show that both ANN and RF are very confident in activity predictions for most compounds, given the fact that nearly all models achieve a one-class-rate that is greater than 0.9.

In general, the ANN provides a new way to explore the bioactivity data and to carry out compound activity predictions. And ANN shows its strength in obtaining a high recall for active compounds together with a good generalisation ability and high prediction confidence.

References

[1] T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, "Structure-Based Virtual Screening for Drug Discovery: a Problem-Centric Review," *AAPS J.*, vol. 14, no. 1, pp. 133–141, 2012.

[2] A. Varnek and I. Baskin, "Machine learning methods for property prediction in chemoinformatics: Quo Vadis?," *J. Chem. Inf. Model.*, vol. 52, no. 6, pp. 1413–1437, Jun. 2012.

[3] "PIDGINv4." [Online]. Available: https://github.com/BenderGroup/PIDGINv4.

[4] "RDKit." [Online]. Available: https://github.com/rdkit/rdkit.

[5] "Scikit-Learn." [Online]. Available: https://github.com/scikit-learn/scikit-learn.

[6] "Tensorflow." [Online]. Available: https://www.tensorflow.org/.

[7] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS One*, vol. 12, no. 6, pp. e0177678–e0177678, Jun. 2017.