Journal of Advanced Computer Science & Technology, 12 (2) (2024) 53-59

Journal of Advanced Computer Science & Technology

Website: www.sciencepubco.com/index.php/JACST

Research paper

Deep Learning-Based Prediction Model for Drug-Target Interactions

Puspanjali Mohapatra[1*], Shantilata Palei[1], Aryan Sinha[1], P. Subham[2] and Adarsh Kumar[1]

[1]*International Institute of Information Technology, Bhubaneswar*
[2]*National Institute of Technology, Rourkela*
[*]*Corresponding author E-mail:puspanjali@iiit-bh.ac.in*

Abstract

This paper comprehensively studies deep learning approaches for drug-target interaction (DTI) prediction in drug discovery. We evaluate the performance of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformers on DTI prediction tasks. Our results demonstrate that the CNN model consistently outperforms both RNN and transformer models in accuracy. Additionally, we investigate the impact of transfer learning on DTI model performance, showing that pre-trained fine-tuning significantly enhances the results. These insights contribute to selecting and optimising deep learning models for DTI prediction, thereby advancing drug discovery efforts. Notably, our findings highlight the potential of combining CNNs with the BindingDB dataset and utilizing transformers as pre-trained models for real-world DTI cases.

Keywords: *Drug-target interaction, transfer learning, drug discovery, deep learning, SMILES, CNN, RNN, Transformer.*

1. Introduction

In this paper, we explore transfer learning techniques to enhance the prediction of Drug-Target Interactions (DTI) [13]. The motivation behind this study stems from the desire to reduce the time and cost associated with traditional drug discovery processes. By leveraging transfer learning, we aim to apply the knowledge gained from related datasets to a target dataset, enabling the development of a more generalized model for DTI prediction [4]. Our objective is to assess the effectiveness and accuracy of transfer learning in this context, utilizing the publicly available dataset, BindingDB.

By developing a computational tool that utilizes transfer learning, we aim to leverage existing knowledge from related datasets and apply it to predicting drug-target interactions [14, 15]. This approach can significantly lower the expenses and time required for experimental validation [8]. By exploring the effectiveness of transfer learning, we can harness its benefits in drug discovery and contribute to developing more efficient and cost-effective methodologies [7].

The main objective of our research is to evaluate the effectiveness and accuracy of transfer learning in the context of DTI prediction [1]. We aim to establish a framework for analyzing unidentified molecules and building a general model for real-world testing. To achieve this, we employ transfer learning techniques, specifically leveraging knowledge from related datasets to enhance the performance of DTI models [11]. Our investigations focus on the publicly accessible dataset BindingDB, providing a practical and relevant context for our research.

The paper's contribution encompasses various techniques, including encoding methods for input data representation, DTI prediction algorithms such as KronRLS and SimBoost, and the utilization of specialized frameworks like DeepPurpose for DTI model creation. Additionally, we explore the application of transfer learning in drug development and propose its incorporation to improve the performance of DTI models. Evaluating these models and techniques involves using five-fold cross-validation, receiver operating characteristic (ROC) curve analysis, and metrics such as precision, recall, and F1 score. Overall, this research aims to advance DTI prediction by harnessing the power of transfer learning and providing insights into its impact on drug discovery efforts [3].

To evaluate the performance of the proposed models and techniques, we employ rigorous evaluation methodologies [16, 9]. Five-fold cross-validation is utilized, which involves randomly splitting the dataset into five subsets for training and testing, providing robust and reliable performance metrics [6]. We analyze the receiver operating characteristic (ROC) curves to determine optimal classification thresholds and utilize the area under the ROC curve (AUC) as an evaluation metric. Furthermore, precision, recall, and F1 score are employed to assess the performance, particularly when dealing with imbalanced datasets and prioritizing correctly identifying positive samples.

In summary, our research aims to leverage transfer learning techniques to improve the prediction of drug-target interactions. By exploring encoding methods, DTI prediction algorithms, and the incorporation of transfer learning, we strive to enhance the efficiency and cost-

effectiveness of drug discovery processes. Through comprehensive evaluation and analysis, our findings contribute to advancing the field of DTI prediction and offer valuable insights for future drug discovery efforts.
The organization of this paper is as follows: Section 2 reviews previous research on DTI prediction, providing a foundation and context for our study. Section 3 discusses preparing and pre-processing datasets, detailing the methodologies employed to ensure data quality and relevance. In Section 4, we describe the dataset training models, including applying various prediction algorithms and integrating transfer learning techniques. Section 5 outlines the performance parameters used to evaluate the models. Section 6 presents the results analysis, interpreting the outcomes of our experiments and comparing them with existing approaches. Finally, Section 7 concludes the paper, summarizing the key findings, implications for drug discovery, and potential directions for future research.

2. Previous Research on DTI Prediction

Studies regarding machine learning techniques to predict DTIs have been used for quite some time. Table 1 mentions the prior research papers on optimization techniques used in feature selection.

Table 1: Optimization Techniques used in Feature Selection

Year	Techniques Used	Application
2022	Matrix Factorization [8]	DTI prediction
2018	Computational Methods in Drug Discovery [5]	DTI prediction
2019	CNN [10]	Drug discovery
2021	Deep unsupervised learning, autoencoders [12]	Drug discovery
2010	Diffusion tensor imaging, SVM [12]	Diagnosis of schizophrenia and autism spectrum disorder

ML techniques are effective for DTI prediction. They have the potent model training and evaluation format for drug discovery. More research is needed to improve the accuracy of ML-based DTI prediction methods. However, the results of previous studies suggest that ML techniques have the potential to revolutionize drug discovery.

3. Dataset Preparation and Pre-processing

Dataset preparation involves gathering, curating, and organizing data representing drug-target interactions. It includes processes such as data collection from reliable sources, data cleaning to remove inconsistencies and errors, and data integration from multiple datasets. Pre-processing techniques are applied to transform raw data into a suitable model training and evaluation format.
Neglecting proper dataset preparation can lead to biased or incomplete results, as models may be trained on data that does not accurately represent real-world scenarios. Inaccurate or inconsistent data introduces noise and hinders model performance, making it challenging to draw reliable conclusions in drug discovery.
Dataset preparation is crucial for addressing specific challenges in DTI prediction, such as data sparsity, imbalanced classes, and the integration of heterogeneous data sources. Appropriate pre-processing techniques mitigate these challenges and ensure models are trained on balanced and representative datasets.
The BindingDB, a widely used benchmark dataset, is chosen to evaluate the proposed method. BindingDB provides experimentally measured binding affinities between small molecules and target proteins, making it suitable for training and evaluating DTI prediction models. The dataset is obtained from a dedicated website, ensuring its integrity and reliability. Data cleaning and pre-processing steps are applied to remove duplicates, normalize target names, and standardize SMILES strings representing ligand structures.
After pre-processing, the dataset is binarized using a threshold on the binding affinity values. The DeepPurpose framework is utilized for this task, converting the affinity values into binary labels of active or inactive interactions. Selecting an appropriate threshold is crucial for model performance, and multiple threshold values are evaluated to determine the optimal one.
The pre-processed and binarized datasets are divided into training, validation, and testing sets. Random splitting ensures an even distribution of data across the sets. Although stratified splitting can address class imbalance, it is not necessary in this study due to the absence of severe imbalance.
By emphasizing dataset preparation and pre-processing, the chapter highlights the significance of well-curated datasets and the efforts made to ensure data quality and suitability. In the subsequent sections, these prepared datasets serve as a foundation for feature engineering, model development, and evaluation.

4. Dataset Training Models

Understanding and identifying potential targets and medications depend on the prediction of drug targets. We can better comprehend all of these compounds thanks to machine learning, enabling us to identify possible therapeutics. Transformer models, Recurrent Neural Networks, and Convolutional Neural Networks have lately contributed to increasing the accuracy of drug target predictions.

4.1. Convolutional Neural Networks (CNNs)

CNNs are practical and efficient for detecting spatial and structural patterns in data. CNN models may be used to comprehend and interpret chemical structure in medication target prediction. CNNs may comprehend hierarchical features, which will comprehend suitable local and global features using convolutional filters and pooling processes. This aids the model's ability to identify the crucial molecular traits connected to the medication and the target.

4.2. Recurrent Neural Networks (RNNs)

RNNs are adept at managing sequential and temporal data. RNN models can capture dependencies and connections among many molecular components in drug target prediction. RNNs may model molecular structures or sequences of chemical compounds by processing the molecular sequence step by step to understand long-term relationships and contextual information.

4.3. Transformer Models

Transformer models, made famous by the attention mechanism, have transformed several natural language processing jobs and have shown promise in predicting pharmacological targets [2]. Transformers are excellent at identifying interdependencies and connections among various input elements. They can learn contextual representations of chemical structures or descriptors using self-attention techniques. Transformer models are highly suited for applications requiring the prediction of drug targets because they can efficiently capture distant dependencies and interactions.

5. Performance Parameters

5.1. Area Under the Precision-Recall Curve (AUPRC)

The AUPRC is a metric that assesses the performance of binary classification models, such as those used in DTI prediction. To calculate the AUPRC, we first construct a precision-recall curve by plotting the precision (or positive predictive value) against the recall (or sensitivity) at various classification thresholds. This curve provides a visual representation of the trade-off between precision and recall.

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall})\, d\text{Recall} \tag{1}$$

5.2. Area Under the Receiver Operating Characteristic Curve (AUROC)

The AUROC is another widely used metric for evaluating binary classification models, including DTI prediction models. It quantifies the model's ability to discriminate between positive and negative instances by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various classification thresholds.

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR} \tag{2}$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{3}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}} \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{5}$$

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)) \tag{6}$$

6. Results Analysis

The table 2 shows the Epoch values of a directly trained model. Step by step, you can observe that the loss values decrease and accuracy increases. Figure 1 shows CNN model training and validation results after training.

Table 2: Epoch Results for CNN using direct training

Traning Loss	Training Accuracy	Validation Loss	Validation Accuracy
0.5326	0.7326	0.4719	0.7667
0.4265	0.7972	0.4319	0.7916
0.3778	0.8264	0.4165	0.8007
0.3250	0.8569	0.3962	0.8228
0.2854	0.8849	0.2696	0.8470

Table 3 represents the Epoch values of the Pre-trained and Fine-tuned model. The performance is a bit better than that of the Direct Training, as more Transfer learning is used here. CNN model training and validation results after transfer learning are displayed in Figure 2.

Figure 1: CNN Model Training and Validation Results after Training

Table 3: Epoch Results for CNN using pre-trained and fine-tuning

Traning Loss	Training Accuracy	Validation Loss	Validation Accuracy
0.5226	0.7414	0.4618	0.7701
0.4165	0.8001	0.4216	0.8005
0.3678	0.8422	0.4060	0.8013
0.3021	0.8691	0.3612	0.8322
0.2642	0.8946	0.2266	0.8570

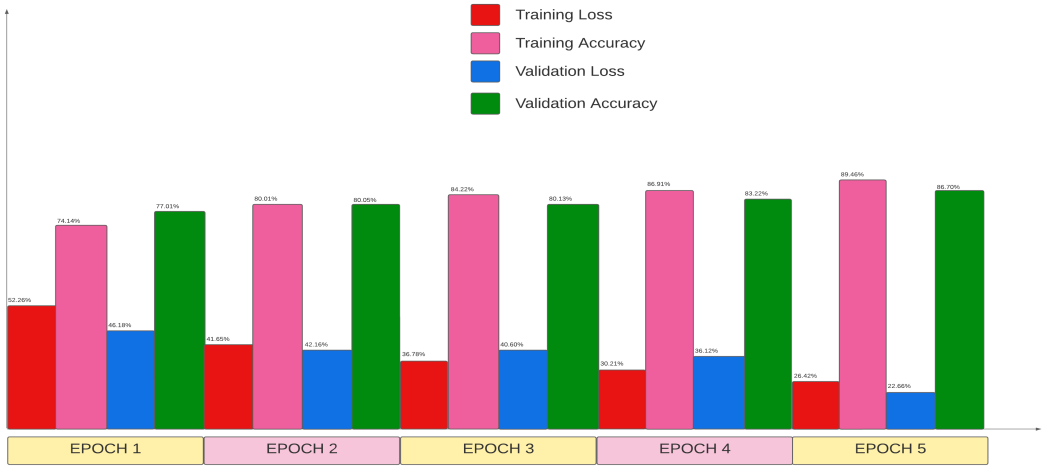

Figure 2: CNN Model Training and Validation Results after Transfer Learning

Table 4: Epoch Results for RNN using direct training

Traning Loss	Training Accuracy	Validation Loss	Validation Accuracy
0.6415	0.6909	0.5362	0.7333
0.5118	0.7515	0.4775	0.7657
0.4722	0.7734	0.4574	0.7762
0.4503	0.7861	0.4462	0.7828
0.4354	0.7949	0.4396	0.7870

Table 5: Epoch Results for RNN using pre-trained and fine-tuning

Traning Loss	Training Accuracy	Validation Loss	Validation Accuracy
0.6225	0.6969	0.5120	0.7533
0.5010	0.7631	0.4788	0.7661
0.4613	0.7832	0.4644	0.7793
0.4422	0.7896	0.4306	0.7814
0.4212	0.8090	0.4231	0.7870

Figure 3: RNN Model Training and Validation Results after Direct Training

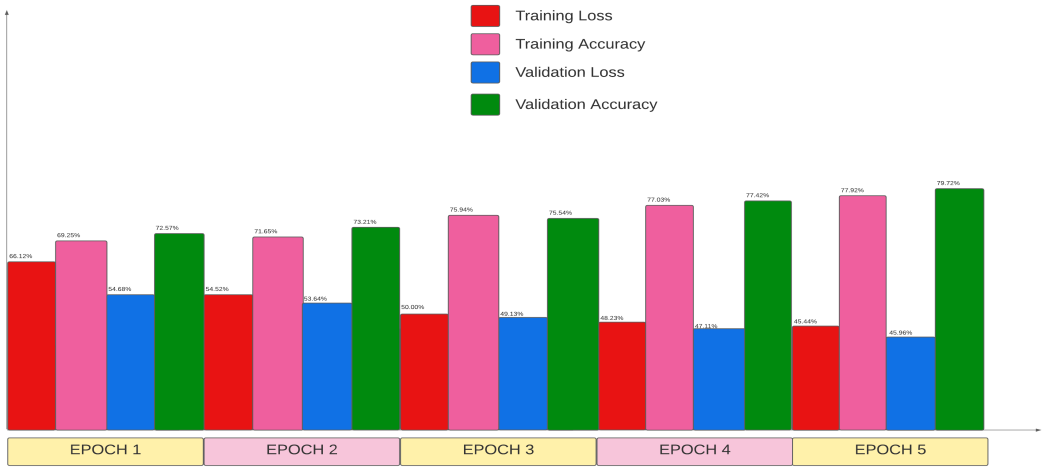

Figure 4: RNN Model Training and Validation Results after Transfer Learning

Table 4 shows us the Epoch values of a directly trained model and step by step; you can observe that the loss values decrease and accuracy increases. RNN model training and validation results after direct training are visible in Figure 3.
The table 5 represents the Epoch values of the Pre-trained and Fine-tuned models. The performance is better than that of Direct Training, as more Transfer learning is used here. Figure 4 shows RNN model training and validation results after transfer learning.
Table 6 shows the epoch values of a directly trained model. Step by step, you can observe that the loss values decrease and accuracy increases. The increase in Accuracy implies that the model is improving in performance. Figure 5 displays transformer model training and validation results after direct training.
The table 7 represents the Epoch values of the Pre-trained and Fine-tuned models. The performance is better than that of Direct Training, as more Transfer learning is used here. Transformer model training and validation results after transfer learning are shown in Figure 6.
The results are represented in table 8, a percentage value that indicates the value under the precision-recall and receiver-operating curves. The results gained using the direct training method are shown in the table 9. Surprisingly, the best-performing model for BindingDB is the CNN-CNN with AUROC of 82.46% and AUPRC of 78.45%; meanwhile, the models with Transformers have lower performance scores.

Table 6: Epoch Results for Transformer using direct training

Traning Loss	Training Accuracy	Validation Loss	Validation Accuracy
0.6776	0.6826	0.5509	0.7189
0.5495	0.7189	0.5495	0.7312
0.4999	0.7422	0.4913	0.7587
0.4715	0.7607	0.4711	0.7701
0.4554	0.7779	0.4596	0.7810

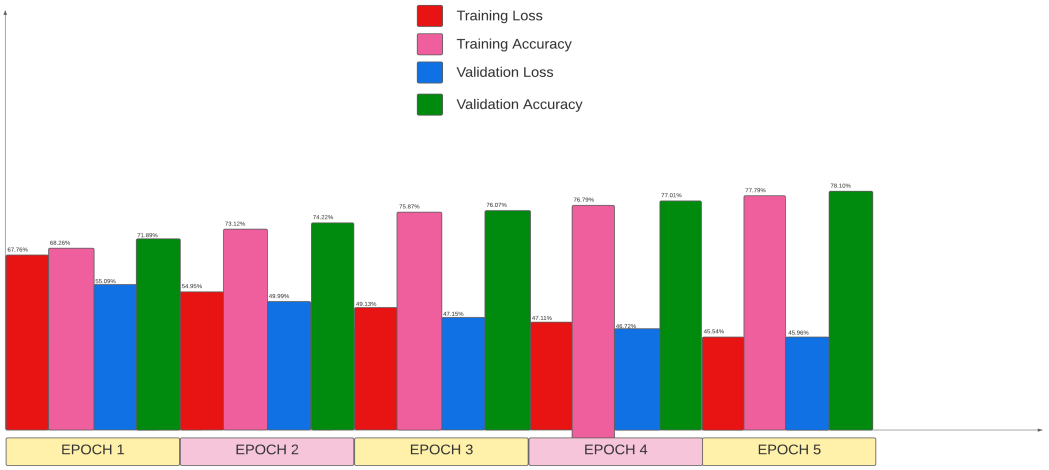

Figure 5: Transformer Model Training and Validation Results after Direct Training

Table 7: Epoch Results for Transformer using Pre-Trained and Fine-Tuning

Traning Loss	Training Accuracy	Validation Loss	Validation Accuracy
0.6612	0.6925	0.5468	0.7257
0.5482	0.7165	0.5364	0.7321
0.5000	0.7514	0.4913	0.7554
0.4823	0.7703	0.4711	0.7742
0.4544	0.7792	0.4596	0.7972

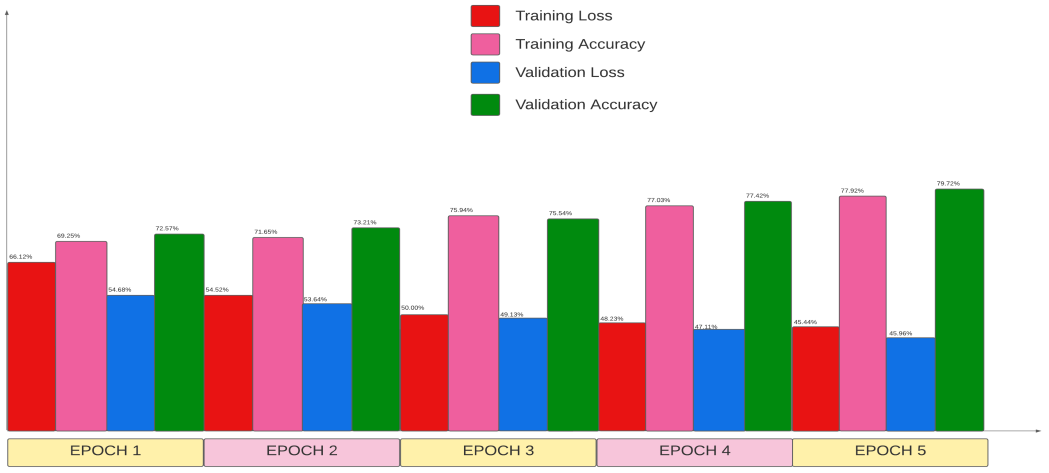

Figure 6: Transformer Model Training and Validation Results after Transfer Learning

Table 8: Direct training Model Comparision

Drug Encoder	Target Encoder	Average AUPRC	Average AUROC
CNN	CNN	78.45%	82.46%
RNN	RNN	72.48%	76.79%
Transformer	Transformer	56.57%	74.75%

Table 9: Pre-trained and Fine-Tuned Model Comparision

Drug Encoder	Target Encoder	Average AUPRC	Average AUROC
CNN	CNN	78.73%	82.68%
RNN	RNN	72.36%	77.35%
Transformer	Transformer	58.43%	75.32%

7. Conclusion

Our extensive experiments and analysis indicate that the CNN model excels when trained directly, while the Transformer model performs better when fine-tuned using pre-trained weights. In the direct training approach, the CNN model outperformed the RNN model in accuracy, precision, recall, and F1-score. Its ability to effectively capture local patterns and spatial dependencies in the input data makes it particularly well-suited for sentiment classification tasks. The CNN model achieved high accuracy and demonstrated robust performance across various evaluation metrics. Conversely, when fine-tuned with pre-trained weights, the Transformer model surpassed both CNN and RNN models. Leveraging the power of self-attention mechanisms and its ability to model long-range dependencies, the Transformer model effectively captured the contextual information required for sentiment analysis tasks. By transferring knowledge from pre-trained models, the Transformer model achieved higher convergence rates and demonstrated improved performance compared to training from scratch. It is important to note that model performance may vary depending on the specific dataset and task. However, based on our experiments, when trained directly, the CNN model is a strong choice for sentiment classification. In contrast, the Transformer model with pre-trained fine-tuning offers significant advantages in terms of performance and convergence. Further research and experimentation are recommended to explore the potential of other deep learning architectures and transfer learning techniques for sentiment analysis tasks. The choice of the model should be based on the task's specific requirements and the available resources.

References

[1] Dongmin Bang, Bonil Koo, and Sun Kim. Transfer learning of condition-specific perturbation in gene interactions improves drug response prediction. *Bioinformatics*, 40(Supplement_1):i130–i139, 2024.

[2] Chenjing Cai, Shiwei Wang, Youjun Xu, Weilin Zhang, Ke Tang, Qi Ouyang, Luhua Lai, and Jianfeng Pei. Transfer learning for drug discovery. *Journal of Medicinal Chemistry*, 63(16):8683–8694, 2020.

[3] Ruolan Chen, Xiangrong Liu, Shuting Jin, Jiawei Lin, and Juan Liu. Machine learning for drug-target interaction prediction. *Molecules*, 23(9):2208, 2018.

[4] Yanyi Chu, Xiaoqi Shan, Tianhang Chen, Mingming Jiang, Yanjing Wang, Qiankun Wang, Dennis Russell Salahub, Yi Xiong, and Dong-Qing Wei. Dti-mlcd: predicting drug-target interactions using multi-label learning with community detection method. *Briefings in bioinformatics*, 22(3):bbaa205, 2021.

[5] Ali Ezzat. *Challenges and solutions in drug-target interaction prediction*. PhD thesis, 2018.

[6] Wenbo Guo, Yawen Dong, and Ge-Fei Hao. Transfer learning empowers accurate pharmacokinetics prediction of small samples. *Drug Discovery Today*, page 103946, 2024.

[7] Yi-Sue Jung, Yoonbee Kim, and Young-Rae Cho. Comparative analysis of network-based approaches and machine learning algorithms for predicting drug-target interactions. *Methods*, 198:19–31, 2022.

[8] Mohammad Reza Keyvanpour, Faraneh Haddadi, and Soheila Mehrmolaei. Dtip-tc2a: An analytical framework for drug-target interactions prediction methods. *Computational Biology and Chemistry*, 99:107707, 2022.

[9] Ashish Khanna, Deepak Gupta, Siddhartha Bhattacharyya, Aboul Ella Hassanien, Sameer Anand, and Ajay Jaiswal. International conference on innovative computing and communications. *Proceedings of ICICC*, 2, 2021.

[10] Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6):e1007129, 2019.

[11] Bin Liu. *Addressing class imbalance in multi-label data*. PhD thesis, Aristotle University Of Thessaloniki, Greece, 2021.

[12] Seyedeh Zahra Sajadi, Mohammad Ali Zare Chahooki, Sajjad Gharaghani, and Karim Abbasi. Autodti++: deep unsupervised learning for dti prediction by autoencoders. *BMC bioinformatics*, 22:1–19, 2021.

[13] Wen Shi, Hong Yang, Linhai Xie, Xiao-Xia Yin, and Yanchun Zhang. A review of machine learning-based methods for predicting drug–target interactions. *Health Information Science and Systems*, 12(1):1–16, 2024.

[14] Derwin Suhartono, Muhammad Rizki Nur Majiid, Alif Tri Handoyo, Pandu Wicaksono, and Henry Lucky. Towards a more general drug target interaction prediction model using transfer learning. *Procedia Computer Science*, 216:370–376, 2023.

[15] AA Toropov, AP Toropova, and E Benfenati. Qsar-modeling of toxicity of organometallic compounds by means of the balance of correlations for inchi-based optimal descriptors. *Molecular diversity*, 14:183–192, 2010.

[16] Orhan Gazi Yalçın and T Istanbul. *Applied neural networks with TensorFlow 2: API oriented deep learning with python*. Springer, 2021.