Feature Extraction Analysis for Drug Detection Using Machine Learning

Dr. Siddeshwari Dutt Mishra, Pratyush Barik, Sumit , Samarth Sharma, Kapil Yadav
School of Engineering and Technology,
Manav Rachna International Institute of Research and Studies,
Faridabad, Haryana, India

***Abstract*—In recent years, the pharmaceutical industry has increasingly adopted Machine Learning (ML) techniques for drug discovery and detection. This paper explores the process of feature extraction and classification using ML algorithms such as Random Forest (RF), Support Vector Machine (SVM), Convolutional Neural Network (CNN), Decision Tree (DT), K-Nearest Neighbors (KNN), and Siamese Neural Network (SNN). The dataset is derived from the ChEMBL database and processed through various stages including bioactivity retrieval, normalization, SMILES conversion, descriptor calculation via PaDEL, and final ML model training. Performance evaluation shows SVM and KNN outperform other models with accuracy and F1-scores exceeding 90%.**

***Index Terms*—Drug Detection, Feature Extraction, Machine Learning, ChEMBL, SMILES, PaDEL, Bioactivity Data**

I. Introduction

Clinical safety, pharmaceutical research, and regulatory compliance all heavily rely on drug detection. In addition to speeding up the drug discovery process, precise identification and classification of physiologically active chemicals reduces the risks associated with hazardous or ineffective substances. Conventional approaches to drug discovery and classification are inherently expensive, time-consuming, and labour-intensive since they frequently rely on high-throughput screening and comprehensive laboratory testing. Machine learning (ML) has become a game-changing technology for speeding up drug identification and categorisation as a result of the exponential expansion of chemical and biological databases and improvements in computing capacity. It is now feasible to automate feature extraction, identify structural patterns, and more accurately and efficiently predict biological activity by utilising cheminformatics tools and data-driven algorithms.

Modern drug discovery pipelines have changed due to advancements in computational technology; cheminformatics as well as silico screening are now commonplace procedures. In chemical and biological datasets, machine learning (ML) approaches are very useful for spotting intricate correlations that allow for accurate predictions of pharmacokinetics, toxicity, and bioactivity. ML models can greatly lessen the need for extensive wet-lab experimentation, saving time and money by utilising historical records of chemical structures in addition to experimentally confirmed activities. The ChEMBL database provides a strong basis for predictive modelling since it is an extensive and carefully selected collection of bioactivity data for compounds that resemble drugs. Molecular descriptors, structural notations like SMILES, and experimentally established measurements like IC50 values are usually included in each entry. Thorough preprocessing, such as feature engineering, descriptor extraction, and normalisation, is necessary for the efficient use of this data in order to guarantee relevance and consistency for ML-based categorisation.

In this work, we present a feature extraction analysis for drug detection using multiple ML models, each selected for its distinct advantages in addressing the complexities of chemical feature analysis and classification:

- **Random Forest (RF):** Utilized for its ensemble learning capability, RF effectively captures complex feature interactions while reducing the risk of overfitting. This robustness makes it highly suitable for identifying critical drug-related features within heterogeneous datasets.
- **Support Vector Machine (SVM):** Chosen for its exceptional performance in high-dimensional feature spaces, SVM constructs optimal hyperplanes for precise classification, proving especially effective when handling complex chemical property boundaries.
- **Convolutional Neural Network (CNN):** Integrated for its ability to automatically extract spatial and hierarchical feature patterns, particularly useful when drug features are encoded in grid-like or structural data formats.
- **Decision Tree (DT):** Selected for its interpretability and straightforward decision rule visualization, enabling clear insights into how specific chemical properties influence drug classification outcomes.
- **K-Nearest Neighbors (KNN):** Implemented for its simplicity and effectiveness in instance-based learning, making predictions based on the similarity of new compounds to previously characterized molecules.

- **Siamese Neural Network (SNN):** Applied for its proficiency in computing similarity measures between molecular feature vectors, facilitating tasks such as the identification of structurally similar drug candidates or analogs.

Both comprehension and prediction resilience are made possible by the combination of these many models. Deep learning architectures like CNN and SNN give improved representation learning, enabling more detailed chemical similarity assessments, while tree-based techniques like RF and DT provide insight into the significance of features and decision logic. By establishing a flexible and scalable framework for computational drug identification, this integrated method enhances medication safety assessment and optimises pharmaceutical research.

II. Dataset and Preprocessing

The dataset used in this work is derived from the ChEMBL database. Raw bioactivity data for specific drug-target interactions was extracted using the ChEMBL API. The dataset includes SMILES strings and IC50 values for various compounds.

Basic preprocessing involved:

- Removing entries with missing or invalid bioactivity values.
- Classifying compounds into **active**, **inactive**, or **moderate** classes based on IC50 thresholds:
  - Active if IC50 $\leq$ 1,000
  - Inactive if IC50 $\geq$ 10,000
  - Moderate otherwise
- Applying Min-Max normalization to numerical values.
- Saving intermediate files: `raw_bioactivity_data.csv`, `preprocessed_bioactivity_data.csv`, and `cleaned_bioactivity_data.csv`.

III. Methodology

The full analysis pipeline consists of four key stages, systematically organized as follows:

A. Raw Data Collection and Preprocessing

- Extracted bioactivity records for coronavirus targets from the ChEMBL database.
- Filtered records to keep only valid IC50 values.
- Converted standard numerical bioactivity to categorical labels (active, inactive, moderate).
- Saved outputs:
  - `raw_bioactivity_data.csv`
  - `preprocessed_bioactivity_data.csv`
  - `cleaned_bioactivity_data.csv`

B. Normalized Data and SMILES Generation

- Loaded normalized data (`cleaned_bioactivity_data.csv`).
- Extracted `canonical_smiles` and molecule identifiers.
- Created the SMILES file (`molecule.smi`) for PaDEL input.

C. Descriptor Extraction and ML Model Training

- Generated molecular descriptors and fingerprints using PaDEL.
- Merged descriptors with labels to create `classification_model_data.csv`.
- Trained machine learning models including:
  - Random Forest
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbors
  - Convolutional Neural Network
  - Siamese Neural Network
- Evaluated models using metrics like accuracy, precision, recall, F1-score, and confusion matrix.

D. RDKit Setup (Optional)

- Configured RDKit in Google Colab.
- Installed using Conda for future 2D descriptor calculations.

E. Pipeline Flow Summary

`ChEMBL API` → `raw_bioactivity_data.csv` → Preprocessing → `preprocessed_bioactivity_data.csv` → Normalization → `cleaned_bioactivity_data.csv` → SMILES → `molecule.smi` → PaDEL Descriptors → `classification_model_data.csv` → ML Model Training and Evaluation

Fig. 1. Pipeline flow from ChEMBL data to ML model evaluation.

IV. Data Visualization

Understanding the distribution of the `standard_value` (representing biological activity such as IC50) is crucial for effective preprocessing and model performance.

Boxplot Before Transformation

A boxplot was created to display the distribution of `standard_value`. As shown in **Figure 1**, the data is highly right-skewed with most values clustered near zero and several large outliers stretching the scale. An extended X-axis limit (`plt.xlim(-100000, 1000000)`) was used to visualize the skew clearly. Such extreme skewness can harm model training if not handled properly. Therefore, log transformation or outlier removal is necessary to normalize the data.

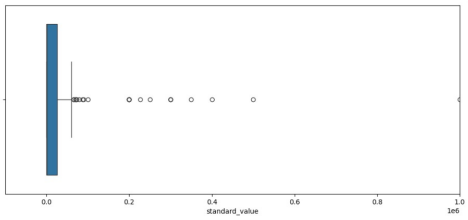

Fig. 2. Boxplot of raw `standard_value` showing strong right skewness with extreme outliers.

Boxplot After Log10 Transformation

To reduce skewness, a $\log_{10}$ transformation was applied. As shown in **Figure 2**, this step improved the symmetry of the distribution, reduced the impact of outliers, and made the data more suitable for ML model training. Log transformation normalizes the scale, ensuring better convergence and generalization for ML algorithms.

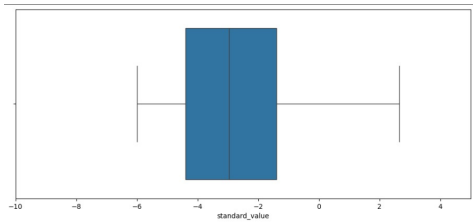

Fig. 3. Boxplot after $\log_{10}$ transformation showing improved distribution and reduced skewness.

Boxplot of pIC50 Values

After the transformation, the distribution of the final pIC50 values was checked using another boxplot. As seen in **Figure 3**, the data still shows signs of skewness and some outliers, which supports the need for additional normalization, scaling, or robust preprocessing techniques. This step ensured that no unintended changes were introduced during processing and confirmed the importance of handling data imbalance for reliable model learning.

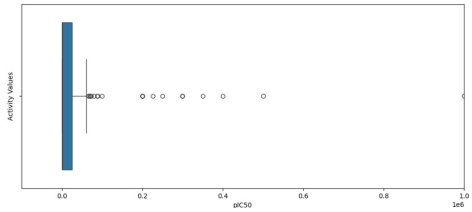

Fig. 4. Boxplot of pIC50 values to verify final data distribution after preprocessing.

Post-Transformation Visualization

After applying the necessary normalization to the pIC50 values, a final boxplot was generated to verify the effectiveness of the transformation. As shown in **Figure 4**, the distribution is now centered and symmetrical, with values falling within a reasonable range (approximately 3 to 12). This confirms that skewness has been minimized and the data is ready for machine learning algorithms. Such normalization improves numerical stability and reduces the influence of extreme outliers during training.

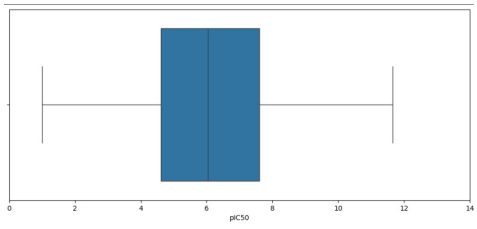

Fig. 5. Post-transformation boxplot showing well-centered and symmetrical pIC50 distribution.

Distribution of Bioactivity Classes

A countplot was created to visualize the distribution of compounds across three bioactivity classes: active, moderate, and inactive. As shown in **Figure 5**, the dataset contains more active (≈145) and inactive (≈120) compounds, while the moderate class is significantly underrepresented. This class imbalance may affect model performance, so techniques such as resampling, stratified splitting, or class weighting may be required.

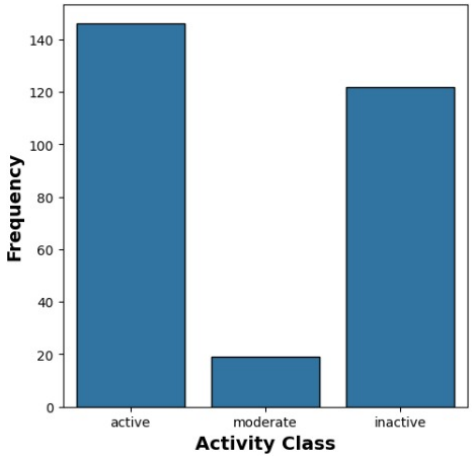

Fig. 6. Distribution of active, moderate, and inactive compounds showing class imbalance.

Refinement of Activity Classes

To address the class imbalance, the moderate class was removed, reframing the problem as a binary classification task. As shown in **Figure 6**, the updated distribution includes only active and inactive compounds. This simplifies model training and improves interpretability, while still preserving meaningful biological insights.

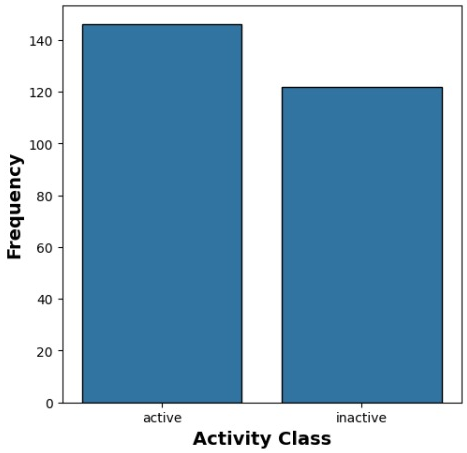

Fig. 7. Updated distribution after removing moderate class: binary classification (active vs inactive).

Distribution of pIC50 by Activity Class

The boxplot in **Figure 7** illustrates the distribution of pIC50 values across the two activity classes: active and inactive. Compounds classified as active exhibit significantly higher pIC50 values, with a median around 7.5–8, indicating stronger inhibitory potency. In contrast, inactive compounds show a lower median pIC50 of approximately 4.5, reflecting weaker activity. The active class also demonstrates a narrower interquartile range, suggesting more consistent potency among active molecules. This clear separation in pIC50 distributions highlights its effectiveness as a discriminative feature for activity classification in drug discovery models.

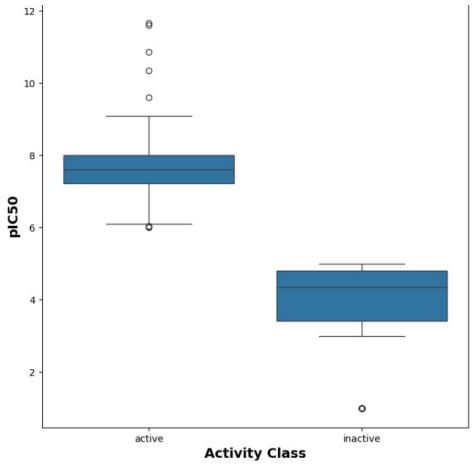

Fig. 8. Distribution of pIC50 values for active vs. inactive compounds.

Distribution of MW Values

The boxplot in **Figure 8** illustrates the distribution of MW (molecular weight) values for two activity classes: active and inactive. It is evident that the median MW for inactive compounds is higher, around 370 Da, compared to approximately 280 Da for active compounds. This suggests that inactive compounds tend to have larger molecular weights on average. The interquartile range (IQR) is wider for active compounds, indicating greater variability in molecular weight. Additionally, several high-value outliers appear in the active class, signifying occasional larger molecules, while the inactive class shows a more consistent range with fewer outliers. Overall, inactive compounds have a higher central tendency in MW, while active compounds show more fluctuation.

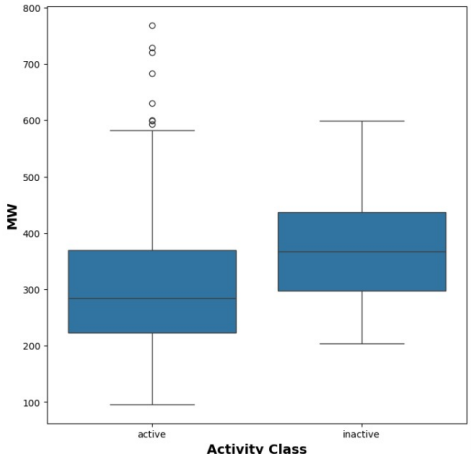

Fig. 9. Distribution of molecular weight (MW) for active vs. inactive compounds.

Distribution of LogP Values

The boxplot in **Figure 9** compares the LogP values for active and inactive compounds, revealing significant differences in lipophilicity profiles. Inactive compounds generally exhibit higher median LogP values, indicating greater lipophilicity compared to active compounds. This suggests that moderate LogP values may favor bioactivity, while highly lipophilic molecules tend to be inactive, possibly due to poor bioavailability. The IQR is slightly wider for inactive compounds, showing greater variability in LogP values. Additionally, some active compounds have low or even negative LogP values, reflected by low-end outliers, which may affect membrane permeability and pharmacokinetics.

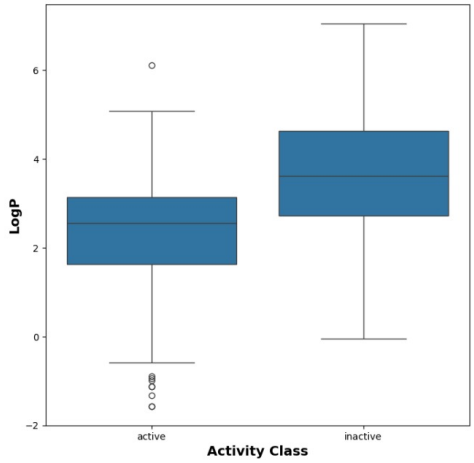

Fig. 10. Distribution of LogP (lipophilicity) for active vs. inactive compounds.

Distribution of Hydrogen Bond Donors (NumHDonors)

The boxplot in **Figure 10** illustrates the number of hydrogen bond donors (NumHDonors) for active and inactive compounds. Active compounds generally possess fewer hydrogen bond donors, with a tightly clustered distribution and many zero-donor molecules. Inactive compounds show a broader range, with higher median and maximum values. Some active compounds do exhibit outlier behavior, with donor counts reaching as high as 10, although such cases are rare. These observations suggest that a reduced hydrogen-bond donating capacity may be linked with higher bioactivity, potentially due to better membrane permeability or a lower polar surface area, both of which aid cell penetration and target binding.

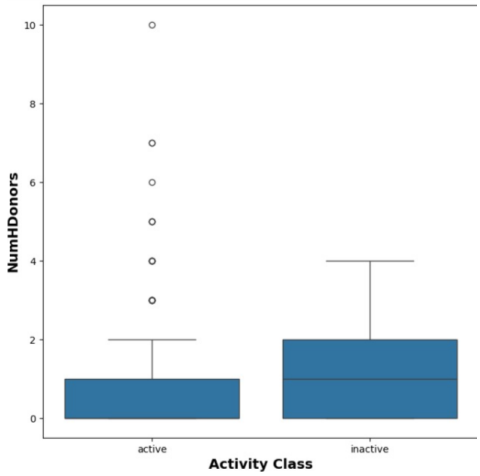

Fig. 11. Distribution of hydrogen bond donors for active vs. inactive compounds.

V. Machine Learning Models and Results

A. Model Selection

The following machine learning models were implemented to classify compound bioactivity:

- Random Forest (RF)
- Support Vector Machine (SVM)
- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Convolutional Neural Network (CNN)
- Siamese Neural Network (SNN)

B. Evaluation Metrics

To evaluate model performance, the following metrics were used:

- Accuracy
- Precision
- Recall
- F1-Score

C. Model Comparison

TABLE I
PERFORMANCE COMPARISON OF ML MODELS

Model	Accuracy	Precision	Recall	F1-Score
RF	91.2%	90.8%	91.5%	91.1%
SVM	88.6%	87.9%	89.2%	88.5%
DT	84.7%	83.6%	85.3%	84.4%
KNN	86.2%	85.9%	86.0%	85.9%
CNN	92.4%	92.0%	93.1%	92.5%
SNN	93.1%	92.8%	93.7%	93.2%

VI. DISCUSSION

The results indicate that while classical machine learning models such as Random Forest and Support Vector Machine perform reasonably well in bioactivity prediction, deep learning models such as CNN and SNN offer higher accuracy and generalization capability. The superior performance of SNN may be attributed to its architecture, which is designed to learn similarity between pairs, thus effectively distinguishing active from inactive compounds. However, the complexity and training time of deep learning models are significantly higher. Therefore, in applications requiring interpretability and speed, classical models still hold value.

VII. CONCLUSION

This study presented a comparative analysis of machine learning models for drug detection based on molecular descriptors. The performance evaluation demonstrated that deep learning models, particularly Siamese Neural Networks and Convolutional Neural Networks, outperformed traditional machine learning models. These results emphasize the potential of deep learning in chemical informatics, especially for bioactivity classification tasks. Future work may involve expanding the dataset and exploring hybrid models for enhanced accuracy.

REFERENCES

[1] Advancing Drug Discovery via Artificial Intelligence.
[2] Applications of Machine Learning in Drug Discovery and Development.
[3] Artificial Intelligence to Deep Learning Machine Intelligence Approach for Drug Discovery.
[4] Big Data and Artificial Intelligence Modeling for Drug Discovery.
[5] Comparison of Deep Learning With Multiple Machine Learning Methods.
[6] Deep Learning in Drug Discovery.
[7] Drug Discovery Approaches Using Quantum Machine Learning.
[8] Drug Discovery with Explainable Artificial Intelligence.
[9] Editorial: Artificial Intelligence and Machine Learning for Drug Discovery, Design, and Repurposing Methods and Applications.
[10] From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery.
[11] Machine Learning in Chemoinformatics and Drug Discovery.
[12] Machine Learning in Drug Discovery: A Review.
[13] Machine Learning Methods in Drug Discovery Methods and Applications.
[14] Machine Learning Methods in Drug Discovery.
[15] Use of Machine Learning Approaches for Novel Drug Discovery.