

RESEARCH ARTICLE

PredART: Uncertainty-quantified Machine Learning Prediction of Androgen Receptor Agonists Overcoming Imbalanced Dataset

Jidon Jang^{1,*}, Dokyun Na² and Kwang-Seok Oh^{1,3,*}

¹Data Convergence Drug Research Center, Korea Research Institute of Chemical Technology, Daejeon, Republic of Korea; ²Department of Biomedical Engineering, Chung-Ang University, Seoul, 06974, Republic of Korea; ³Department of Medicinal and Pharmaceutical Chemistry, University of Science and Technology, 176 Gajeong-ro, Yuseong, Daejeon, 34129, Republic of Korea

© 2025 The Author(s). Published by Bentham Science Publisher. This is an open access article published under CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/legalcode>

Abstract: *Aim:* This study aims to develop and validate a machine learning-based model for the accurate prediction of Androgen Receptor (AR) agonistic toxicity, addressing the challenges posed by data imbalance in existing predictive models.

Background: Anomalous agonistic activity of the androgen receptor is a known major indicator of reproductive toxicity, which can lead to prostate cancer. Machine learning-based models have been developed for the rapid prediction of such agonists. However, the existing models have exhibited biased learning outcomes and low sensitivity due to the imbalance in the available training data. In the early screening process of drug discovery, low sensitivity caused by data imbalance can hinder the detection of potentially toxic compounds.

Objective: The objective of this study is to develop a machine learning prediction model that classifies whether a drug candidate is an androgen receptor agonist or not with highly balanced performance compared to existing models.

Methods: PredART is a bootstrap aggregated k-nearest neighbor model for the balanced prediction of androgen receptor agonistic toxicity using 381 active and 8,089 inactive datasets with structural features of them.

Result: In this work, we propose an advanced model that combines the bootstrap aggregating algorithm with machine learning binary classifiers to identify androgen receptor-based reproductive toxicity while avoiding biased prediction results. The optimal model using k-nearest neighbor classifiers achieved an accuracy of 0.831, Positive Predictive Value (PPV) of 0.882, sensitivity of 0.625, specificity of 0.951, Mathews Correlation Coefficient (MCC) of 0.633 on external test data, demonstrating a significant improvement in sensitivity compared to the previous study and achieving balanced learning. Furthermore, by calculating the standard deviation among outputs of the classifiers and employing this prediction uncertainty as a screening metric to select reliable predictions, the model's performance could be further enhanced.

Conclusion: Based on the bootstrap aggregating algorithm, our prediction model effectively addressed data imbalance while evaluating the performance of various machine learning and deep learning classifiers for a benchmark. Additionally, by quantifying uncertainty, our model provided an intuitive assessment of prediction reliability during large-scale screening processes.

Keywords: Androgen receptor, reproductive toxicity, machine learning, bootstrap aggregation, uncertainty quantification, DHT.

1. INTRODUCTION

The Androgen Receptor (AR) is a nuclear receptor belonging to the steroid hormone group. It functions as a ligand-dependent transcription factor that regulates specific gene

expressions. Upon binding to its endogenous ligands, testosterone, and 5 α -dihydrotestosterone (DHT), the AR initiates male sexual differentiation. Without these androgens or a functional AR, male sexual development is impaired, leading to conditions like androgen insensitivity syndrome [1, 2].

Prostate cancer is projected to be the leading cause of cancer-related deaths among American men within the next decade [3], placing a significant focus on the AR. Given that androgenic stimulation is essential for prostate cancer devel-

*Address correspondence to these authors at the Data Convergence Drug Research Center, Korea Research Institute of Chemical Technology, Daejeon, Republic of Korea; E-mails: jdjang@krikt.re.kr; ksoh@krikt.re.kr

ARTICLE HISTORY

Received: September 24, 2024
Revised: November 14, 2024
Accepted: November 26, 2024

DOI:
[10.2174/011574893635551241220190451](https://doi.org/10.2174/011574893635551241220190451)



opment, numerous studies have highlighted the AR's critical role in this progression [4, 5]. Consequently, the AR is a key target for assessing reproductive toxicity in drug development. Predicting AR-induced toxicity for new drug candidates in advance is crucial. However, high-throughput experimental AR affinity assays for every compound are labor-intensive and costly, posing challenges in the growing field of drug discovery [6].

Recent advancements in computational power and algorithms have made it possible to predict AR binding reducing cost for experimental assay [7-9]. The Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA), led by Mansouri *et al.* (2020), aimed to screen a large chemical library to identify agonists and antagonists for AR activity. This was achieved through a consensus modeling framework, integrating predictions from *in silico* models developed by 25 research groups [10]. Among these computational methods, Artificial Intelligence (AI) has significantly enhanced the role of *in silico* predictive models in challenging tasks such as drug discovery [11, 12]. Data-driven prediction models have been widely applied in bioinformatics fields because these can assess nonlinear relationships between target properties and relevant features of existing data, enabling the prediction of new data [13-15].

AI models, particularly that adept at processing large datasets and recognizing patterns, can predict various Absorption, Distribution, Metabolism, Excretion, And Toxicity (ADMET) properties, facilitating the discovery of new drug candidates [16, 17]. The predictions from AI models are not only significantly faster than experimental assays but also outperform molecular docking simulations in speed, while efficiently handling large datasets, thereby accelerating the drug development pipeline. We present the application of various machine learning (ML) algorithms and deep neural networks to predict AR agonists and evaluate their performance.

AI-based AR agonist classification models can serve as valuable tools in rapidly predicting reproductive toxicity, as evidenced by previous studies [8, 18]. However, publicly available AR data samples are highly imbalanced [19] (see Methods for the details), presenting significant challenges for machine learning model training. When there is a substantial disparity between the number of positive and negative samples, the model tends to exhibit bias towards the more frequent class. This can lead to poor sensitivity (or Recall), compromising the model's reliability in practical applications. A low-sensitivity model is particularly problematic in drug screening, as it increases the risk of misclassifying toxic drugs as non-toxic. A recent study has demonstrated a similar bias in AR agonist prediction using ChatGPT prompts [20]. To address this challenge, Snow *et al.* [21] applied the Synthetic Minority Over-sampling Technique (SMOTE) for multiclass classification, including antagonists with experimental validation. However, the dataset used for training was limited in size, and the classification performances for AR agonists were notably lower compared to other categories. Idakwo *et al.* [19] integrated a SMOTE technique with the edited nearest neighbor algorithm for var-

ious endpoints of the Tox21 dataset [22], however, sensitivity for AR agonist prediction remained low. Piir *et al.* [23] developed a multi-class classification for AR binders using undersampled training data and a random forest algorithm, showing enhanced sensitivity. However, there is a possibility of overfitting to a single undersampled subset of the data. Furthermore, by using a balanced test dataset and considering various classification metrics including PPV (or precision) which is a trade-off with sensitivity, a more robust classification model can be developed.

To address this issue, this study combines machine learning classification models with bootstrap aggregating (Bagging) algorithms [24]. Bagging mitigates data imbalance by sampling the data to create multiple balanced training datasets while covering all available data for learning. Individual models are trained on these datasets, and their predictions are averaged (ensembled) to reduce model bias and produce more balanced prediction results. This work details the application of the bagging algorithm to overcome challenges arising from imbalanced data and demonstrates improved model performance compared to our previous work [18]. Furthermore, to enhance the reliability of prediction results, we introduce uncertainty quantification [25], providing additional screening guidance to assess the confidence in the model's predictions.

2. MATERIALS AND METHODS

2.1. Data Preparation

Data were obtained from the Tox21 Dataset which generated efficient toxicological screening tests [22, 26] which includes toxicity assay results for various nuclear receptor targets. Specifically, the Androgen Receptor (AR) data from Tox21 can be retrieved by PubChem ID 743040 with their label of activity and further supplemented with newly acquired AR-induced active compounds identified through searches in Cortellis Drug Discovery Intelligence [27]. We extracted data with EC₅₀ values below 10 μ M with data filtering applied to include only androgen receptor agonists with human cell-specific data. This conservative threshold is half of the maximum for the 'weak' potency category, as proposed in related previous works [10, 28]. Then, any duplicates are removed, and refined the dataset to exclude entries that encountered errors during structural encoding to Morgan fingerprint [29] as input of our model using RDKit python package [30]. The final dataset comprised 381 active compounds and 8,089 inactive compounds. These compounds were used for the training process. To validate model performance, 20% of the active compounds were set as holdout test data, along with an equal number of inactive compounds (76 active + 76 inactive compounds) for a balanced test set.

2.2. Model Constructions

To address the data imbalance in the training data for classification, we applied the Bagging algorithm to machine learning binary classifiers (Fig. 1). For training these classifiers, we fixed the number of active compounds and under-

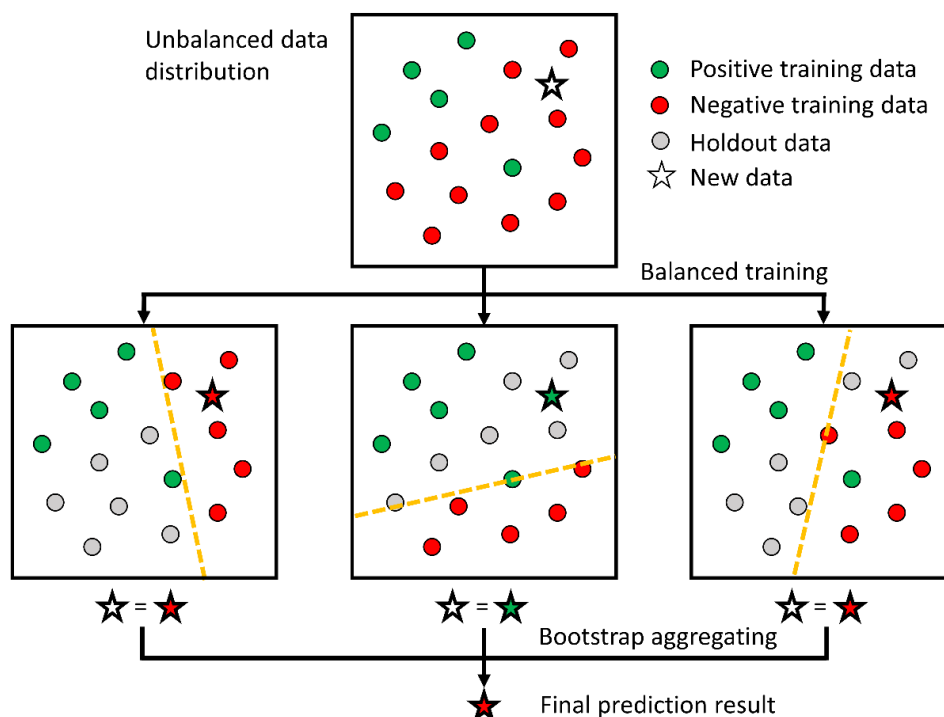


Fig. (1). Schematic diagram of bagging algorithm for unbalanced dataset. (A higher resolution / colour version of this figure is available in the electronic copy of the article).

sampling the inactive data to achieve a balanced training dataset. We employed a sliding-window sampling method to ensure that each inactive compound had an equal chance of being selected for training [31]. Randomly arranged inactive data was sampled along a shifting window at fixed intervals to sample data uniformly. In this case, the size of the window and shifting interval is 305. The number of sampling iterations for Bagging was conducted sufficiently to cover the entire inactive dataset multiple times. Consequently, we created 100 different combinations of the balanced dataset for training 100 classifiers. After 100 training iterations, Bagging was used to average the classification prediction scores consisting of 0 (inactive) and 1 (active) as the final prediction result. If the average score is below 0.5, the compound is finally classified as inactive; if it is 0.5 or above, it is classified as an active compound (toxic). The overall algorithm scheme is depicted in Fig. (2). For comparison purposes, we evaluated a total of nine classifier models, including eight conventional machine learning models and one deep neural network. The models employed were SVM [32], Random Forest [33], LightGBM [34], XGBoost [35], CatBoost [36], k-NN [37], Logistic Regression [38], Gaussian Process Classification [39], and an MPNN [40, 41]. Among these, k-NN demonstrated the best predictive performance and was chosen as the final classification model. Drug molecules were represented using Morgan fingerprints [29] with a radius of 2 and a bit vector size of 2048 from their 2D structure generated by A Simplified Molecular-Input Line-Entry System (SMILES). This fingerprint encodes the presence or absence of circular sub-structure from each atomic center with radius 2 as a binary bit across 2048 positions. Hyperpa-

rameter optimization for each model was performed using a grid search algorithm [42] with 5-fold cross-validations. The names of the models along with their optimized hyperparameter values are detailed in Table S1.

2.3. Validation with a Balanced Holdout-test Dataset

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{PPV (Precision)} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

$$\text{F1 score} = 2 \times \frac{\text{PPV} \times \text{Sensitivity}}{\text{PPV} + \text{Sensitivity}} \quad (6)$$

(TP; True positive, TN; True negative, FP; False Positive, FN; False Negative)

To evaluate model performance, 20% of both active data (76 compounds) and inactive data (76 compounds) were randomly selected to form a holdout test set and excluded during model training. The balanced test dataset, with an equal number of active and inactive data, is appropriate for a more robust evaluation of the model's performance. For instance, if the test data is excessively skewed towards inactive cases, a model predicting the majority as inactive may show a high overall accuracy but suffer from low sensitivity [43]. After training, their performance was assessed by predicting the

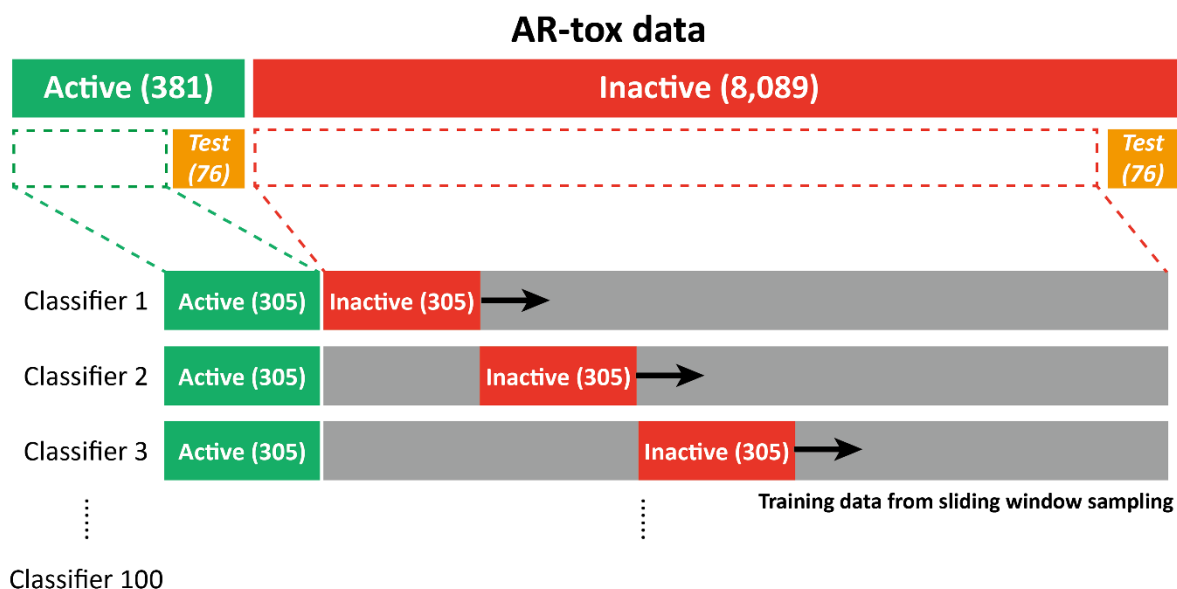


Fig. (2). Undersampling method of negative dataset and the number of each data group. (A higher resolution / colour version of this figure is available in the electronic copy of the article).

compounds in the holdout test set. The performance metrics included accuracy, positive predictive value (PPV) (or precision), sensitivity (or recall), specificity, Mathews Correlation Coefficient (MCC) [44], F1-score, and Area Under the receiver operating Characteristic curve (AUC) (see Equation 1-6 using confusion matrix). These results are summarized in Table 1.

3. RESULTS AND DISCUSSION

3.1. Model Benchmarking and Training Performance Metrics

Fig. (1) illustrates the scheme of applying the Bagging algorithm for training on imbalanced data. The training process proceeds as follows: (1) Undersample the negative data to match the number of positive data in the training set. (2) Train the model using the positive training data and the undersampled negative data. (3) After multiple iterations, average the prediction results from the trained models to determine the final prediction outcome. In order to guarantee that all negative data points have an equal chance of being selected, a sliding window method is employed during the undersampling process (Fig. 2). This method ensures that all negative examples contribute uniformly to the final prediction by providing them with a balanced selection probability.

Fig. (3) shows the results of predicting the holdout test data using models trained with various machine learning algorithms. A total of nine machine learning models were employed, including support vector machine (SVM) [32], Random Forest [33], LightGBM [34], XGBoost [35], CatBoost [36], k-Nearest Neighbor (k-NN) [37], Logistic Regression [38], Gaussian Process Classification [39], and Message Passing Neural Network (MPNN) [40, 41]. Their performance metrics are all summarized in Table S2. Of the models evaluated, the k-NN exhibited the highest MCC per-

formance metric (0.612). In contrast, the MPNN did not achieve effective performance across all metrics due to the small size of the undersampled training data compared with the complexity of deep model parameters. Notably, the sensitivity of the data has improved compared to previous results obtained from training on unbalanced data [18]. Additionally, the performance metrics of the optimized k-NN model conducted under varying numbers of undersampled inactive data, revealed that the MCC peaked at a fully balanced training (Fig. S1). As the number of inactive samples increased, sensitivity also decreased due to biased training, while undersampling fewer inactive data showed a reversed tendency of imbalanced learning with a slight decrease of MCC.

3.2. Validation for External Test Data

Fig. (4) presents the evaluation results of the model on test data obtained from Yan *et al.* [8]. These external data were derived from the previous studies and the EDKB database [45, 46]. After removing duplicates from previously used datasets, it consists of a total of 48 active compounds and 82 inactive compounds. This allowed us to assess the model's performance on completely new data and compare it with our previous study using the same dataset. Our previous work [18], trained on imbalanced data, showed poor sensitivity, with many false negative (inactive) predictions (Table 1). Despite high PPV and specificity, the low MCC, F1-score, and AUC indicated biased learning. In contrast, our current study employed models trained on balanced data using an ensemble method, resulting in improved performance metrics (Table S3). Among these models, the k-NN model shows the best performance for most metrics: accuracy of 0.831, PPV of 0.882, sensitivity of 0.625, specificity of 0.951, MCC of 0.633, F1-score of 0.732, and AUC of 0.788. This demonstrates a significant enhancement in

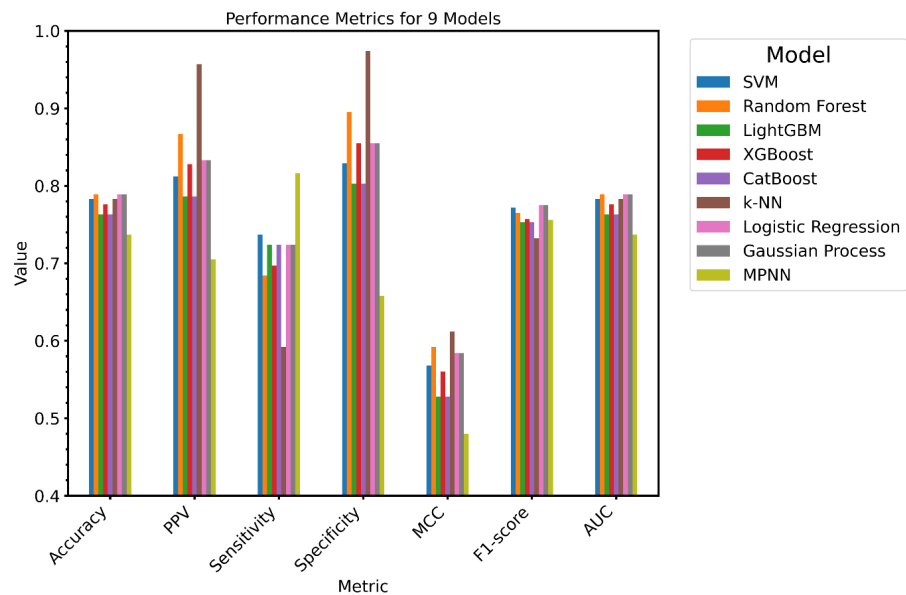


Fig. (3). Performance metrics of various machine learning algorithms for test data split. (A higher resolution / colour version of this figure is available in the electronic copy of the article).

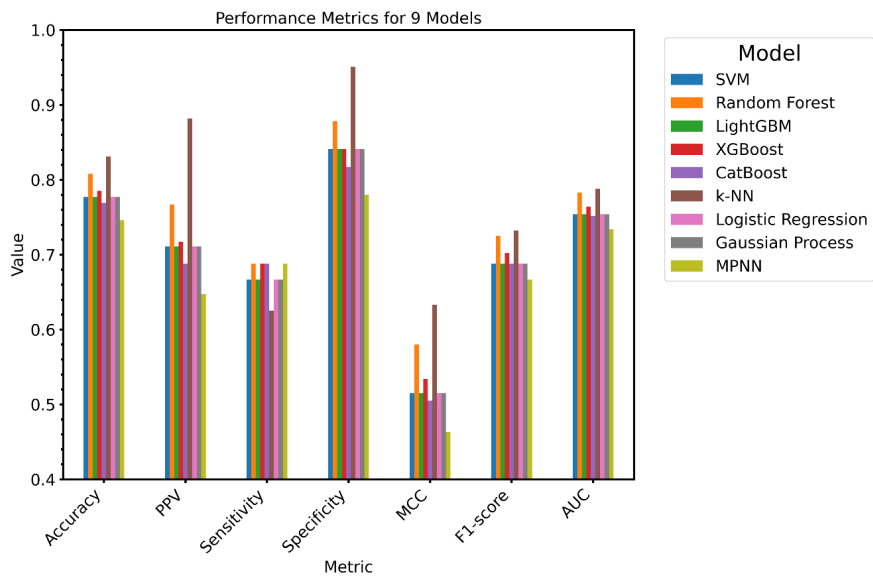


Fig. (4). Performance metrics of various machine learning algorithms for Yan *et al.* [8] test data. (A higher resolution / colour version of this figure is available in the electronic copy of the article).

Table 1. Performance metrics for Yan *et al.* test data compared to our previous work. The numbers highlighted in bold indicate the better model for each metric.

	Accuracy	PPV	Sensitivity	Specificity	MCC	F1-score	AUC
Ref. [18]	0.800	0.882	0.211	0.991	0.370	0.341	0.783
PredART	0.831	0.882	0.625	0.951	0.633	0.732	0.788

Note: [8].

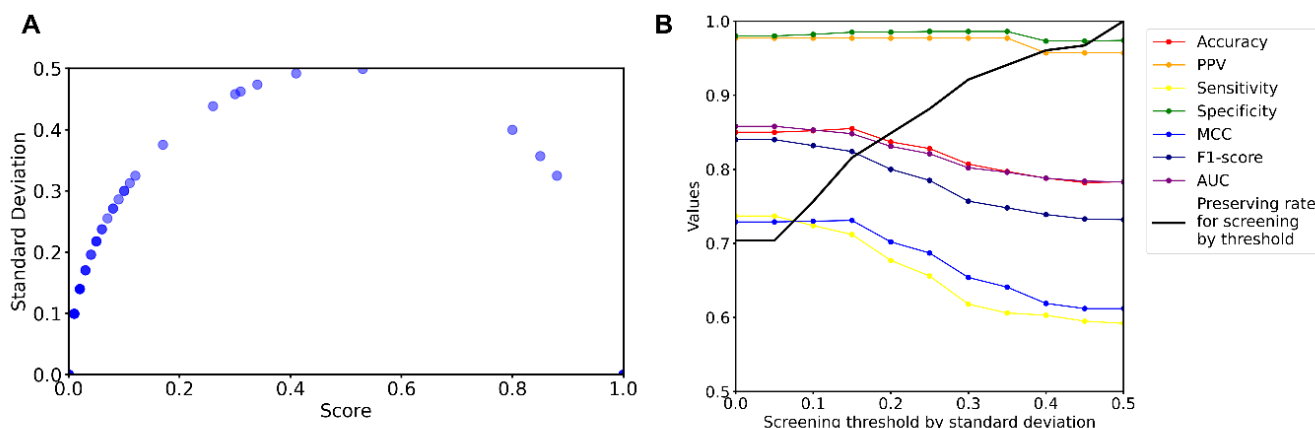


Fig. (5). **A)** Standard deviation for prediction using bagging algorithms for 152 holdout test data. **B)** Model performance removing uncertain data by threshold of standard deviation. (A higher resolution / colour version of this figure is available in the electronic copy of the article).

sensitivity while maintaining PPV and specificity, with nearly doubled MCC and F1 scores, indicating balanced and reliable predictions. Finally, we choose the k-NN as the optimal model for AR-agonist toxicity prediction.

3.3. Uncertainty Quantification

Quantifying uncertainty for ML-assisted decision-making is important to its reliability, especially in medical situations related to patient safety [25]. The Bagging algorithm inherently quantifies prediction uncertainty through the standard deviation of prediction values from multiple trained models as follows:

$$\sigma^2 = \frac{1}{T} \sum_{t=1}^T (\hat{s}_t^*)^2 - \left(\frac{1}{T} \sum_{t=1}^T \hat{s}_t^* \right)^2 \quad (7)$$

where T is the number of models for Bagging ($T=100$) and $\hat{s}_t^*$ means prediction values from these classifiers [47].

This standard deviation reflects the epistemic uncertainty, and variation in predictions during model ensemble which arises from a lack of knowledge about which model may be optimal (see Equation 7) [48]. As shown in Fig. (5A), high uncertainty indicates significant variation, leading to values near 0.5 which are considered unreliable prediction results. Conversely, low uncertainty, with predictions close to 0 or 1, indicates consistent predictions across the model ensemble. Regardless of the training data being undersampled for negative data, the prediction results remain stable. Thus, screening out test set data with low standard deviation can enhance the performance metrics for the remaining data (Fig. 5B). About 70% of the data showed consistent prediction results across all models in the bagging ensemble process, without variance. When uncertain results were screened out based on standard deviation, the preserving rate increased with higher threshold levels (see black line in Fig. 5B). For instance, by screening data with a standard deviation of less than 0.15 and calculating the model's performance, 81.6% of the target test data can be preserved while achieving accuracy of 0.855, PPV of 0.977, sensitivity of 0.712, specificity of 0.985,

MCC of 0.731, F1-score of 0.824, and AUC of 0.848. Utilizing the standard deviation of predicted values as an additional criterion for large-scale screening could facilitate the selection of candidate compounds based on more accurate model predictions.

3.4. Importance of Balanced Training and Model Ensemble

Table S4 compared the performance metrics of the optimal k-NN model trained on imbalanced data and with the Bagging algorithm. When trained on the imbalanced data, the k-NN model exhibited extreme values for PPV and specificity. However, accuracy, sensitivity, MCC, F1-score, and AUC were significantly lower. This bias towards the majority class (negative data) resulted in sensitivity close to random guessing, a crucial metric in drug screening as it reflects the model's ability to identify toxic compounds. On the other hand, the k-NN model trained with the Bagging algorithm addressed the imbalance by creating a balanced dataset. This resulted in enhanced sensitivity while maintaining good performance across other metrics like accuracy, MCC, F1-score, and AUC. Table S4 also includes the performance of a single undersampled model without an ensemble. While sensitivity improved compared to the imbalanced k-NN model, the overall performance was slightly lower than the Bagging due to the limited dataset. The metrics after uncertainty-quantified screening showed highly robust performances with a reasonable preserving rate of data.

CONCLUSION

We developed a binary classification model to predict agonists of AR and employed the bagging algorithm to address the challenge of an imbalanced training dataset. Uniform undersampling with sliding window generated balanced training subsets for the model ensemble. By encoding the compounds using substructure-based fingerprints consisting of 2048-size binary bits that capture their structural information, we tested a total of 9 machine learning and deep

learning algorithms. Our results showed that the k-NN classification model combined with the bagging algorithm yielded the best predictive performance as MCC. Furthermore, this model demonstrated excellent performances on an external test set from another data source, significantly improving the previously low sensitivity observed in other studies, with nearly a three-fold increase.

Additionally, we proposed another screening metric through the quantification of prediction uncertainty, suggesting a diversified model that enhances performance with more reliable predictions using their epistemic uncertainty. After screening out highly uncertain data whose standard deviation is higher than 0.15, our model can achieve an accuracy of 0.855, PPV of 0.977, sensitivity of 0.712, specificity of 0.985, MCC of 0.731, F1-score of 0.824, and AUC of 0.848 with preserving 81.6% of the test data.

Although we applied an algorithmic approach to mitigate the effects of data imbalance, the small and limited dataset fundamentally restricts the model's predictive capacity. Future work should focus on advanced ML techniques to overcome limitations posed by small datasets. Transfer learning [49] using foundation models pre-trained on large-scale datasets could be applied, and multimodal learning from alternative modalities of AR data with feature optimization may potentially address the limitation of small data availability [50]. In addition, while our model focuses on quantifying uncertainties arising from the ML model only, future studies could consider a mathematical approach using some external factors [51] relevant to AR agonistic mechanisms, potentially providing richer predictive insights for AR agonism.

LIST OF ABBREVIATIONS

TP	=	True Positive
TN	=	True Negative
FP	=	False Positive
FN	=	False Negative
PPV	=	Positive Predictive Value
MCC	=	Mathews Correlation Coefficient
AUC	=	Area Under receiver operating Characteristic curve
AR	=	Androgen Receptor
DHT	=	5 α -dihydrotestosterone
AI	=	Artificial Intelligence
ADMET	=	Absorption, Distribution, Metabolism, Excretion, and Toxicity
ML	=	Machine Learning
Bagging	=	Bootstrap aggregating
SVM	=	Support Vector Machine
k-NN	=	k-Nearest Neighbor
MPNN	=	Message Passing Neural Network

EC ₅₀	=	Half maximal Effective Concentration
SMILES	=	Simplified Molecular-Input Line-Entry System.

AUTHORS' CONTRIBUTION

The authors confirm their contribution to the paper as follows: study conception and design: JJ., KSO.; analysis and interpretation of results: JJ., DN., and KSO.; model construction: JJ. All authors reviewed the results and approved the final version of the manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The datasets, source codes, and their detailed descriptions are available for download through (<https://github.com/joshua-416/PredART>).

FUNDING

This work was supported by a grant from the National Research Foundation of Korea (NRF), funded by the Korean government (Grant no. 2022M3E5F3085688). This work was also funded by the Korea Research Institute of Chemical Technology (Grant no. KK2431-20).

CONFLICT OF INTEREST

The authors declare no conflict of interest financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

- [1] Brinkmann AO. Molecular basis of androgen insensitivity. *Mol Cell Endocrinol* 2001; 179(1-2): 105-9. [http://dx.doi.org/10.1016/S0303-7207\(01\)00466-X](http://dx.doi.org/10.1016/S0303-7207(01)00466-X) PMID: 11420135
- [2] McPhaul MJ, Marcelli M, Tilley WD, Griffin JE, Wilson JD. Androgen resistance caused by mutations in the androgen receptor gene. *FASEB J* 1991; 5(14): 2910-5. <http://dx.doi.org/10.1096/fasebj.5.14.1752359> PMID: 1752359
- [3] Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. *CA Cancer J Clin* 2013; 63(1): 11-30.

- <http://dx.doi.org/10.3322/caac.21166> PMID: 23335087
- [4] Heinlein CA, Chang C. Androgen receptor in prostate cancer. *Endocr Rev* 2004; 25(2): 276-308. <http://dx.doi.org/10.1210/er.2002-0032> PMID: 15082523
- [5] Tan MHE, Li J, Xu HE, Melcher K, Yong E. Androgen receptor: Structure, role in prostate cancer and drug discovery. *Acta Pharmacol Sin* 2015; 36(1): 3-23. <http://dx.doi.org/10.1038/aps.2014.18> PMID: 24909511
- [6] Lynch C, Sakamuru S, Huang R, et al. Identifying environmental chemicals as agonists of the androgen receptor by using a quantitative high-throughput screening platform. *Toxicology* 2017; 385: 48-58. <http://dx.doi.org/10.1016/j.tox.2017.05.001> PMID: 28478275
- [7] Ng HW, Zhang W, Shu M, et al. Competitive molecular docking approach for predicting estrogen receptor subtype α agonists and antagonists. *BMC bioinformatics* 2014; 15(11): 1-15. <http://dx.doi.org/10.1186/1471-2105-15-S11-S4>
- [8] Yan L, Zhang Q, Huang F, et al. Ternary classification models for predicting hormonal activities of chemicals *via* nuclear receptors. *Chem Phys Lett* 2018; 706: 360-6. <http://dx.doi.org/10.1016/j.cplett.2018.06.022>
- [9] Manganello S, Roncaglioni A, Mansouri K, et al. Development, validation and integration of *in silico* models to identify androgen active chemicals. *Chemosphere* 2019; 220: 204-15. <http://dx.doi.org/10.1016/j.chemosphere.2018.12.131> PMID: 30584954
- [10] Mansouri K, Kleinstreuer N, Abdelaziz AM, et al. CoMPARA: Collaborative modeling project for androgen receptor activity. *Environ Health Perspect* 2020; 128(2): 027002. <http://dx.doi.org/10.1289/EHP5580> PMID: 32074470
- [11] Cáceres EL, Tudor M, Cheng AC. Deep learning approaches in predicting ADMET properties. *Future Med Chem* 2020; 12(22): 1995-9. <http://dx.doi.org/10.4155/fmc-2020-0259> PMID: 33124448
- [12] Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery. *Drug Discov Today* 2019; 24(5): 1157-65. <http://dx.doi.org/10.1016/j.drudis.2019.03.015> PMID: 30890362
- [13] Wong L, Wang L, You ZH, Yuan CA, Huang YA, Cao MY. GKLOMLI: A link prediction model for inferring miRNA-lncRNA interactions by using Gaussian kernel-based method on network profile and linear optimization algorithm. *BMC Bioinformatics* 2023; 24(1): 188. <http://dx.doi.org/10.1186/s12859-023-05309-w> PMID: 37158823
- [14] Wei M, Wang L, Li Y, et al. BioKG-CMI: A multi-source feature fusion model based on biological knowledge graph for predicting circRNA-miRNA interactions. *Sci China Inf Sci* 2024; 67(8): 189104. <http://dx.doi.org/10.1007/s11432-024-4098-3>
- [15] Guo LX, Wang L, You ZH, et al. Likelihood-based feature representation learning combined with neighborhood information for predicting circRNA-miRNA associations. *Brief Bioinform* 2024; 25(2): bbae020. <http://dx.doi.org/10.1093/bib/bbae020> PMID: 38324624
- [16] Dulsat J, López-Nieto B, Estrada-Tejedor R, Borrell JI. Evaluation of free online ADMET tools for academic or small biotech environments. *Molecules* 2023; 28(2): 776. <http://dx.doi.org/10.3390/molecules28020776> PMID: 36677832
- [17] Xiong G, Wu Z, Yi J, et al. ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res* 2021; 49(W1): W5-W14. <http://dx.doi.org/10.1093/nar/gkab255> PMID: 33893803
- [18] Yu MS, Lee J, Lee Y, Na D. 2-D chemical structure image-based *in silico* model to predict agonist activity for androgen receptor. *BMC Bioinformatics* 2020; 21(S5)(Suppl. 5): 245. <http://dx.doi.org/10.1186/s12859-020-03588-1> PMID: 33106158
- [19] Idakwo G, Thangapandian S, Luttrell J, et al. Structure-activity relationship-based chemical classification of highly imbalanced Tox21 datasets. *J Cheminform* 2020; 12(1): 66. <http://dx.doi.org/10.1186/s13321-020-00468-x> PMID: 33372637
- [20] Cai X, Lai H, Wang X, et al. Comprehensive evaluation of molecule property prediction with ChatGPT. *Methods* 2024; 222: 133-41. <http://dx.doi.org/10.1016/j.ymeth.2024.01.004> PMID: 38242382
- [21] Snow O, Lallous N, Ester M, Cherkasov A. Deep learning modeling of androgen receptor responses to prostate cancer therapies. *Int J Mol Sci* 2020; 21(16): 5847. <http://dx.doi.org/10.3390/ijms21165847> PMID: 32823970
- [22] Huang R, Xia M, Nguyen DT, et al. Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front Environ Sci* 2016; 3: 85. <http://dx.doi.org/10.3389/fenvs.2015.00085>
- [23] Piir G, Sild S, Maran U. Binary and multi-class classification for androgen receptor agonists, antagonists and binders. *Chemosphere* 2021; 262: 128313. <http://dx.doi.org/10.1016/j.chemosphere.2020.128313> PMID: 33182081
- [24] Breiman L. Bagging predictors. *Mach Learn* 1996; 24(2): 123-40. <http://dx.doi.org/10.1007/BF00058655>
- [25] Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell* 2019; 1(1): 20-3. <http://dx.doi.org/10.1038/s42256-018-0004-1>
- [26] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. *Front Environ Sci* 2016; 3: 80. <http://dx.doi.org/10.3389/fenvs.2015.00080>
- [27] Clarivate. Cortellis drug discovery intelligence. <https://www.cortellis.com/drugdiscovery> (Accessed March 21, 2024).
- [28] Mansouri K, Abdelaziz A, Rybacka A, et al. CERAPP: Collaborative estrogen receptor activity prediction project. *Environ Health Perspect* 2016; 124(7): 1023-33. <http://dx.doi.org/10.1289/ehp.1510267> PMID: 26908244
- [29] Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010; 50(5): 742-54. <http://dx.doi.org/10.1021/ci100050t> PMID: 20426451
- [30] Landrum G. RDKit Documentation 2013. Available from: https://ftp.ccp4.ac.uk/ccp4/7.0/unpacked/checkout/rdkit-Release_2015_03_1/Docs/Book/RDKit.pdf
- [31] Braverman V, Ostrovsky R, Zaniolo C. Optimal sampling from sliding windows. Twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Providence, Rhode Island, USA, 2009. pp. 147-156. <http://dx.doi.org/10.1145/1559795.1559818>
- [32] Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. *IEEE Intell Syst Their Appl* 1998; 13(4): 18-28. <http://dx.doi.org/10.1109/5254.708428>
- [33] Breiman L. Random forests. *Mach Learn* 2001; 45(1): 5-32. <http://dx.doi.org/10.1023/A:1010933404324>
- [34] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 2017; 3149-57.
- [35] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. 22nd acm sigkdd international conference on knowledge discovery and data mining. San Francisco, California, USA, 2016. pp. 785-794
- [36] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: Unbiased boosting with categorical features. In: *Advances in Neural Information Processing Systems*. MIT Press 2017.
- [37] Peterson L. K-nearest neighbor. *Scholarpedia J* 2009; 4(2): 1883. <http://dx.doi.org/10.4249/scholarpedia.1883>
- [38] Menard S. *Applied logistic regression analysis*. Sage Publications, Inc. 2002. <http://dx.doi.org/10.4135/9781412983433>
- [39] MacKay DJ. *Introduction to Gaussian processes*. NATO ASI series F comput syst sci 1998; 168: 133-68.
- [40] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. 34th International Conference on Machine Learning. Sydney, NSW, Australia, :2017. PMLR: pp. 1263-1272
- [41] Song Y, Zheng S, Niu Z, Fu Z-H, Lu Y, Yang Y. Communicative Representation Learning on Attributed Molecular Graphs. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. 2020; pp. 2831-8. <http://dx.doi.org/10.24963/ijcai.2020/392>
- [42] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyperparameter optimization. *Adv Neural Inf Process Syst* 2011; 24

- [43] Wei Q, Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One 2013; 8(7): e67863.
<http://dx.doi.org/10.1371/journal.pone.0067863> PMID: 23874456
- [44] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020; 21(1): 6.
<http://dx.doi.org/10.1186/s12864-019-6413-7> PMID: 31898477
- [45] Li F, Xie Q, Li X, *et al.* Hormone activity of hydroxylated polybrominated diphenyl ethers on human thyroid receptor- β : *In vitro* and *in silico* investigations. Environ Health Perspect 2010; 118(5): 602-6.
<http://dx.doi.org/10.1289/ehp.0901457> PMID: 20439171
- [46] Ding D, Xu L, Fang H, *et al.* The EDKB: An established knowledge base for endocrine disrupting chemicals. BMC bioinformatics 2010; 11: 1-7.
<http://dx.doi.org/10.1186/1471-2105-11-S6-S5>
- [47] Ryu S, Kwon Y, Kim WY. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. Chem Sci (Camb) 2019; 10(36): 8438-46.
<http://dx.doi.org/10.1039/C9SC01992H> PMID: 31803423
- [48] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach Learn 2021; 110(3): 457-506.
<http://dx.doi.org/10.1007/s10994-021-05946-3>
- [49] Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. J Big Data 2016; 3(1): 9.
<http://dx.doi.org/10.1186/s40537-016-0043-6>
- [50] Ma M, Ren J, Zhao L, Tulyakov S, Wu C, Peng X. Smil: Multimodal learning with severely missing modality. Proceedings of the AAAI Conference on Artificial Intelligence. May 2021, pp. 2302-2310
<http://dx.doi.org/10.1609/aaai.v35i3.16330>
- [51] Tutsoy O, Balikci K, Ozdil NF. Unknown uncertainties in the COVID-19 pandemic: Multi-dimensional identification and mathematical modelling for the analysis and estimation of the casualties. Digit Signal Process 2021; 114: 103058.
<http://dx.doi.org/10.1016/j.dsp.2021.103058> PMID: 33879984

DISCLAIMER: The above article has been published, as is, ahead-of-print, to provide early visibility but is not the final version. Major publication processes like copyediting, proofing, typesetting and further review are still to be done and may lead to changes in the final published version, if it is eventually published. All legal disclaimers that apply to the final published article also apply to this ahead-of-print version.