Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 12(1): 164-175 (2024)

Drug Solubility Prediction: A Comparative Analysis of GNN, MLP, and Traditional Machine Learning Algorithms

Veysel GİDER[1*] Cafer BUDAK[2]

[1*]Batman University, Distance Education Application and Research Center, Batman, Turkey

[1,2]Dicle University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Diyarbakır, Turkey

Article Info

Research article
Received: 05/10/2023
Revision: 05/12/2023
Accepted: 05/12/2023

Keywords

Machine Learning
Drug Solubility
Graph Neural Networks
Regression Models

Makale Bilgisi

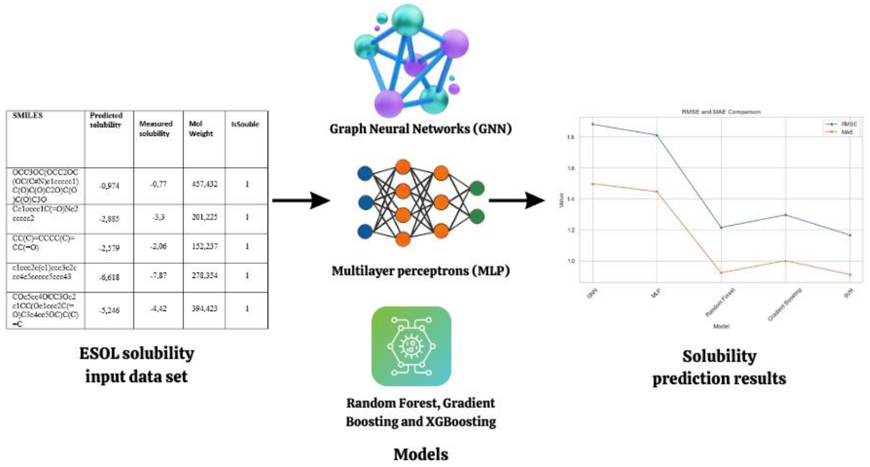
Araştırma makalesi
Başvuru: 05/10/2023
Düzeltme: 05/12/2023
Kabul: 05/12/2023

Anahtar Kelimeler

Makine Öğrenmesi
İlaç Çözünürlük
Grafik Sinir Ağları
Regresyon Modelleri

Graphical/Tabular Abstract (Grafik Özet)

This study evaluates drug solubility prediction models, highlighting Random Forest's superior efficacy compared to Graph Neural Networks. / Bu çalışma ilaç çözünürlük tahmin modellerini değerlendirirek, Rastgele Orman'ın Grafik Sinir Ağlarına göre üstün etkililiğini vurgulamaktadır.



SMILES	Predicted solubility	Measured solubility	Mol Weight	IsSouble
OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)C(O)C3O	-0,974	-0,77	457,432	1
Cc1occc1C(=O)Nc2ccccc2	-2,885	-3,3	201,225	1
CC(C)=CCCC(C)=CC(=O)	-2,579	-2,06	152,237	1
c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc43	-6,618	-7,87	278,354	1
COc5cc4OCC3Oc2c1CC(Oc1ccc2C(=O)C3c4cc5OC)C(C)=C	-5,246	-4,42	394,423	1

ESOL solubility input data set
Graph Neural Networks (GNN)
Multilayer perceptrons (MLP)
Random Forest, Gradient Boosting and XGBoosting
Models
RMSE and MAE Comparison
RMSE
MAE
Value
1.8
1.6
1.4
1.2
1.0
GNN
MLP
Random Forest
Gradient Boosting
SVR
Model
Solubility prediction results

***Figure A**: Illustrates the schematic representation of the proposed methodology through a flow diagram. / **Şekil A:** Önerilen metodolojinin şematik gösterimini bir akış diyagramı aracılığıyla göstermektedir.*

Highlights (Önemli noktalar)

- *The Random Forest model stands out in drug solubility prediction with minimal error rates and superior efficacy. / Rastgele Orman modeli, ilaç çözünürlüğü tahmininde minimal hata oranları ve üstün etkinlikle öne çıkıyor.*
- *The GNN model exhibits lower performance with higher error rates and lower explanatory power compared to other models. / GNN modeli, diğer modellere göre yüksek hata oranları ve düşük açıklama gücü ile daha düşük bir performans sergiliyor.*
- *Study emphasizes differences among modeling approaches, highlighting Random Forest's effectiveness. / Çalışma, modelleme yaklaşımları arasındaki farkları vurgular, Random Forest'ın etkinliğini öne çıkarır.*

***Aim (Amaç):** To analyze Random Forest, GNN and traditional ML models for drug resolution in detail and to identify the most effective model for pharmaceutical design. / İlaç çözünürlüğü için Random Forest, GNN ve geleneksel ML modellerini ayrıntılı bir şekilde analiz etmek ve farmasötik tasarım için en etkili modeli belirlemektir.*

***Originality (Özgünlük):** The study contributes by deeply assessing models' performances and comparing their accuracy and explanatory powers. / Çalışma, modellerin performanslarını derinlemesine değerlendirerek doğruluk ve açıklama güçlerini karşılaştırmasıyla literatüre katkı sağlamaktadır.*

***Results (Bulgular):** Results reveal Random Forest's superior efficacy (RMSE: 1.2145, MAE: 0.9221) compared to GNN (RMSE: 1.8389, MAE: 1.4684, R2: 0.2147). / Sonuçlar, Random Forest'ın üstün etkinliğini (RMSE: 1.2145, MAE: 0.9221) ve GNN'nin nispeten daha düşük etkinliğini (RMSE: 1.8389, MAE: 1.4684, R2: 0.2147) ortaya koyar.*

***Conclusion (Sonuç):** The Random Forest model showed superior efficiency with minimal error rates, whereas the GNN model showed inferior performance. / Random Forest modeli, minimal hata oranları ile üstün bir etkinlik gösterirken, GNN modeli daha düşük performans sergilemiştir.*

*Corresponding author, e-mail: veysel.gider@batman.edu.tr

DOI: 10.29109/gujsc.1371519

GU J Sci, Part C, 12(1): 164-175 (2024)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

Drug Solubility Prediction: A Comparative Analysis of GNN, MLP, and Traditional Machine Learning Algorithms

Veysel GİDER[1*] Cafer BUDAK[2]

[1*]*Batman University, Distance Education Application and Research Center, Batman, Turkey*

[1,2]*Dicle University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Diyarbakır, Turkey*

Article Info

Research article
Received: 05/10/2023
Revision: 05/12/2023
Accepted: 05/12/2023

Keywords

Machine Learning
Drug Solubility
Graph Neural Networks
Regression Models

Abstract

The effective development and design of pharmaceuticals hold fundamental importance in the fields of medicine and the pharmaceutical industry. In this process, the accurate prediction of drug molecule solubility is a critical factor influencing the bioavailability, pharmacokinetics, and toxicity of drugs. Traditionally, mathematical equations based on chemical and physical properties have been used for drug solubility prediction. However, in recent years, with the advancement of artificial intelligence and machine learning techniques, new approaches have been developed in this field. This study evaluated different modeling approaches consisting of Graph Neural Networks (GNN), Multilayer Perceptron (MLP), and traditional Machine Learning (ML) algorithms. The Random Forest (RF) model stands out as the optimal performer, manifesting superior efficacy through the attainment of minimal error rates. It attains a Root Mean Square Error (RMSE) value of 1.2145, a Mean Absolute Error (MAE) value of 0.9221, and an R-squared (R2) value of 0.6575. In contrast, GNN model displays comparatively suboptimal performance, as evidenced by an RMSE value of 1.8389, an MAE value of 1.4684, and an R2 value of 0.2147. These values suggest that the predictions of this model contain higher errors compared to other models, and its explanatory power is lower. These findings highlight the performance differences among different modeling approaches in drug solubility prediction. The RF model is shown to be more effective than other methods, while the GNN model performs less effectively. This information provides valuable insights into which model should be preferred in pharmaceutical design and development processes.

İlaç Çözünürlüğü Tahmini: GNN, MLP ve Geleneksel Makine Öğrenimi Algoritmalarının Karşılaştırmalı Analizi

Makale Bilgisi

Araştırma makalesi
Başvuru: 05/10/2023
Düzeltme: 05/12/2023
Kabul: 05/12/2023

Anahtar Kelimeler

Makine Öğrenmesi
İlaç Çözünürlük
Grafik Sinir Ağları
Regresyon Modelleri

Öz

İlaçların etkin bir şekilde geliştirilmesi ve tasarlanması, tıp ve ilaç endüstrisi alanlarında temel öneme sahiptir. Bu süreçte, ilaç molekülünün çözünürlüğünün doğru bir şekilde tahmin edilmesi, ilaçların biyoyararlanımını, farmakokinetiğini ve toksisitesini etkileyen kritik bir faktördür. Geleneksel olarak, ilaç çözünürlüğü tahmini için kimyasal ve fiziksel özelliklere dayalı matematiksel denklemler kullanılmıştır. Ancak son yıllarda yapay zekâ ve makine öğrenimi tekniklerinin ilerlemesiyle bu alanda yeni yaklaşımlar geliştirilmiştir. Bu çalışmada, Grafik Sinir Ağları (GNN), Çok Katmanlı Algılayıcı (MLP) ve geleneksel Makine Öğrenmesi (ML) algoritmalarından oluşan farklı modelleme yaklaşımları değerlendirilmiştir. Rastgele Orman (RF) modeli, minimum hata oranlarına ulaşarak üstün etkinlik gösteren en iyi performans gösteren model olarak öne çıkmaktadır. Kök Ortalama Kare Hata (RMSE) değeri 1,2145, Ortalama Mutlak Hata (MAE) değeri 0,9221 ve R-kare (R2) değeri 0,6575'tir. Buna karşılık GNN modeli, 1,8389 RMSE değeri, 1,4684 MAE değeri ve 0,2147 R2 değeri ile kanıtlandığı üzere nispeten düşük bir performans sergilemektedir. Bu değerler, bu modelin tahminlerinin diğer modellere kıyasla daha yüksek hata içerdiğini ve açıklayıcı gücünün daha düşük olduğunu göstermektedir. Bu bulgular, ilaç çözünürlüğü tahmininde farklı modelleme yaklaşımları arasındaki performans farklılıklarını vurgulamaktadır. RF modelinin diğer yöntemlere göre daha etkili olduğu, GNN modelinin ise daha az etkili performans gösterdiği görülmektedir. Bu bilgi, farmasötik tasarım ve geliştirme süreçlerinde hangi modelin tercih edilmesi gerektiği konusunda değerli bilgiler sağlamaktadır.

*Corresponding author, e-mail: veysel.gider@batman.edu.tr

DOI: 10.29109/gujsc.1371519

1. INTRODUCTION (GİRİŞ)

In today's world, drug discovery and development represent paramount domains in the pharmaceutical industry and medical research. In this intricate process, the accurate anticipation of drug molecule solubility assumes a pivotal significance. The solubility of a drug molecule is considered a fundamental parameter in drug design and formulation. Solubility determines how well a compound can dissolve in water or another solvent, which can impact the pharmacokinetics, bioavailability, and toxicity of the drug. Therefore, accurately predicting the solubility of a drug candidate during its development stage is essential for the early detection and resolution of potential issues [1, 2].

Over the years, the landscape of drug solubility prediction has transitioned from conventional methodologies to embrace advanced techniques rooted in artificial intelligence (AI) and machine learning (ML). These technological advancements have provided access to vast amounts of molecular data, enabling the development of new approaches in solubility prediction. While traditional methods attempt to predict drug solubility using mathematical equations based on the chemical and physical properties of the compound, AI and ML offer a more flexible and data-driven approach, capable of recognizing complex molecular interactions and patterns [2, 3].

In this context, Graph Neural Networks (GNNs) have attracted considerable attention. GNNs exhibit success in handling graph-structured datasets across diverse domains and under various learning paradigms, including supervised, semi-supervised, self-supervised, and unsupervised settings. The majority of graph-based methodologies fall within the domain of unsupervised learning, frequently relying on Auto-encoders, contrastive learning, or concepts related to random walks.

GNNs differ from traditional neural networks in that they are specifically designed to operate on graph structures rather than sequences. The use of graphs has experienced significant growth and recognition in recent years, primarily due to their remarkable ability to effectively represent complex real-world problems characterized by interconnections. These applications include structured data where information is used in unstructured formats such as test cases. Furthermore, graphs have proven valuable in modeling a variety of domains, including social networks, molecular structures, web link data, and more, enabling extensive analysis and interpretation. GNNs have proven their efficacy in image analysis tasks such as image segmentation and object detection by employing graphs as representations for images. In conclusion, the use of GNNs allows for a specialized approach to processing graph-structured data, facilitating improved performance and results in a variety of domains where interconnectedness and structured relationships are paramount considerations.

GNNs find application across a diverse spectrum of tasks and domains, including but not limited to network embedding, graph classification, node classification, spatial-temporal graph forecasting, and graph generation. Their utility spans a wide array of activities and fields. Their adaptability positions GNNs as pivotal tools for tackling intricate problems characterized by relational structures and dependencies. GNNs utilize a graph-based approach to represent and analyze molecular structures. This approach represents molecular structures as graphs and can make solubility predictions by analyzing these graphs. GNNs can contribute to a better understanding of molecular interactions and accelerate drug design [4, 5].

However, traditional regression models and conventional ML algorithms are still considered effective tools in this field. Traditional machine learning methodologies, particularly exemplified by Multilayer Perceptron (MLP), have exhibited their efficacy in the domain of solubility predictions and are widely adopted within the pharmaceutical industry [6]. Additionally, prominent ML algorithms such as XGBoost, Gradient Boosting (GB), and Random Forest (RF) have demonstrated robust predictive capabilities, particularly on extensive molecular datasets [7].

This study aims to compare AI-based models with traditional ML methods in drug solubility prediction. Specifically, we will evaluate the performance of GNNs and MLP in predicting drug solubility. We will also examine the role of popular ML algorithms like RF, GB, and XGBoost in this context. This research could guide which model performs best in drug design and development processes and contribute to future drug discovery studies.

1.1. Literature Review (Literatür Taraması)

The prediction of drug solubility is widely recognized as a longstanding challenge in the pharmaceutical industry, leading to extensive research in this field. Traditional methods utilize various mathematical equations and rules to predict drug solubility by considering factors such as chemical groups in molecular structure, bond lengths, and atomic charges. However, these methods often achieve limited success in predicting the solubility of individual drug molecules.

In recent epochs, substantial advancements have materialized within the purview of computational modeling, with a specialized emphasis on the intricate task of predicting the aqueous solubility of diverse substances. Due to the limitations of traditional methods, AI and machine learning techniques have become more appealing for drug solubility prediction. These techniques offer greater flexibility in analyzing large datasets and identifying complex patterns in molecular interactions. Specifically, AI methods such as deep learning and GNNs have been employed to better represent molecular structures and improve solubility predictions [8]. Various prediction techniques, including Multiple Linear Regression (MLR), Principal Component Analysis (PCA), Partial Least Squares (PLS), Artificial Neural Network (ANN), K-Nearest Neighbors (k-NN), Support Vector Machine (SVM), and RF, are utilized to forecast the properties of molecules. These predictions rely on descriptors that capture the characteristics of chemical structures [9-11].

GNNs represent molecular structures using a graph-based approach, modeling atoms as nodes and chemical bonds as edges. GNNs play an effective role in a range of applications, from predicting drug-protein binding values to analyzing drug similarities, predicting drug side effects without extracting drug scaffolds, and more. The use of GNNs has emerged as a significant tool to accelerate the drug discovery process, reduce costs, and discover new drug candidates by allowing for a more detailed analysis of molecular interactions [12]. In contrast, traditional regression models like MLP represent molecular features as vectors and make predictions using these feature vectors [13].

The potential of GNNs in drug solubility prediction has been explored in several recent studies. In one investigation, diverse deep learning models were developed for solubility prediction. Four discrete GNN models were postulated to encapsulate molecular representation, and among them, the AttentionFP model showcased noteworthy superiority in performance [14]. Simultaneously, an innovative Multi-Ordered Graphical Attention Network (MoGAT) was introduced as an advanced framework for the prediction of solubility. The primary objectives of this proposition were to enhance prediction performance and facilitate the explication of the predicted results. Findings indicated that MoGAT outperformed contemporary methods in terms of performance and demonstrated the compatibility of predicted results with established chemical knowledge [15]. In a different investigation, a novel graph framework was proposed to anticipate the water solubility of pharmaceutical compounds. This conceptual framework introduced a distinct GNN model named ALIGNN, explicitly tailored for the QM9 dataset [16].

The collective findings from these studies suggest that GNNs have the potential to enhance solubility predictions by offering a more nuanced understanding of molecular interactions. Nevertheless, traditional regression models such as MLP continue to be acknowledged as effective tools in this domain and are widely utilized by various pharmaceutical companies. For example, one research project employed multiple machine learning algorithms (MLR, ANN, RF, ET, and SVM) to predict drug solubility in different solvents [17]. In a separate investigation, Kernel Ridge Regression (KRR), Least Angle Regression (LAR), and MLP were applied to forecast the solubility of Lenalidomide, a drug used in the treatment of specific bone marrow-related conditions in adults [18]. Another study focused on developing an AI-based model using a SVM to examine the solubility data of the drug Busulfan [19]. Additionally, three distinct machine learning approaches, namely k-NN, MLP, and KRR, were employed to predict the solubility of the drug Nystatin [20].

Additionally, popular machine learning algorithms such as RF, GB, and XGBoost have been employed for drug solubility prediction, yielding successful results. These algorithms are considered valuable tools for making predictions on large molecular datasets and are deemed necessary for drug solubility prediction [21].

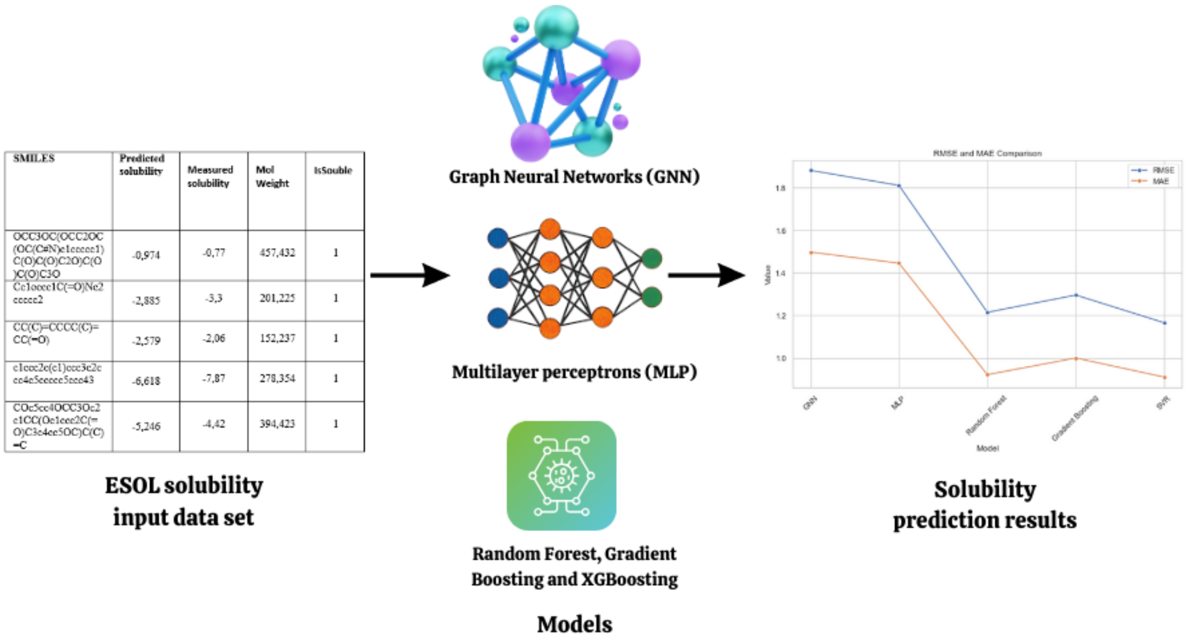
2.MATERIALS AND METHODS (MATERYAL VE METOD)

This research endeavors to undertake a comparative analysis of diverse ML and AI methodologies concerning the prediction of drug solubility. In pursuit of this objective, an assembly of drug data is initially amassed, and various modeling strategies are scrutinized through the lens of this dataset.

The drug dataset includes the chemical properties, structures, and solubility values of various drug molecules. This dataset is obtained from a database widely used in drug design and development processes and encompasses a variety of drug molecules.

Various methodologies are employed in the prediction of drug molecule solubility, encompassing models such as GNNs, MLP, RF, GB, and XGBoost. These models employ different mathematical and graphical approaches to represent the molecular features and structures of drugs.

The methods of the study encompass the preparation and preprocessing of the drug dataset as the initial steps. Subsequently, the training of various models and their utilization for solubility predictions are executed. The results are then employed to compare the performance of different models, ultimately determining which model is the most effective for drug solubility prediction. Figure 1 diagram shows the flowchart of the proposed method.



SMILES	Predicted solubility	Measured solubility	Mol Weight	IsSouble
OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)C(O)C3O	-0,974	-0,77	457,432	1
Cc1occc1C(=O)Nc2ccccc2	-2,885	-3,3	201,225	1
CC(C)=CCCC(C)=CC(=O)	-2,579	-2,06	152,237	1
c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc43	-6,618	-7,87	278,354	1
COc5cc4OCC3Oc2c1CC(Oc1ccc2C(=O)C3c4cc5OC)C(C)=C	-5,246	-4,42	394,423	1

ESOL solubility input data set
Graph Neural Networks (GNN)
Multilayer perceptrons (MLP)
Random Forest, Gradient Boosting and XGBoosting
Models
RMSE and MAE Comparison
RMSE
MAE
1.8
1.6
1.4
1.2
1.0
Value
GNN
MLP
Random Forest
Gradient Boosting
SVR
Model
Solubility prediction results

Figure 1. Illustrates the schematic representation of the proposed methodology through a flow diagram. (Önerilen metodolojinin şematik gösterimini bir akış diyagramı aracılığıyla göstermektedir)

2.1. Dataset (Veri Seti)

This study utilizes the "ESOL (Estimated Aqueous Solubility)" dataset, which is employed for predicting the solubility of drugs. The ESOL dataset serves as a data source created with the purpose of measuring and estimating the aqueous solubility of various chemical compounds. This dataset finds application in drug design, chemical analysis, pharmacokinetic investigations, and molecular modeling studies [22, 23].

The ESOL dataset comprises 1,125 chemical compounds, each with its solubility provided as a logarithmically transformed value. These values represent solubility predictions in micromolars. Additionally, it includes independent variables representing the chemical structure and molecular properties of each compound. These molecular properties constitute the fundamental data used for solubility predictions [22, 23].

This dataset is employed to contribute to the understanding of critical pharmaceutical parameters such as the bioavailability, pharmacokinetics, and toxicity of drugs. Furthermore, it is extensively investigated to comprehend how artificial intelligence-based regression models and deep learning techniques can be utilized to expedite drug design processes.

Regarded as a pivotal asset within the pharmaceutical industry, chemical research, and molecular modeling domains, this dataset holds substantial importance. The objective of this study is to predict solubility utilizing the ESOL dataset, with the anticipation that these predictions will provide invaluable insights for the design and development of pharmaceutical compounds.

The esol dataset consists of 10 columns and 1128 rows. Table 1 below shows the relevant attribute columns for 19 drugs for this data.

Table 1. ESOL dataset (ESOL veriseti)

Compound - ID	Minimum Degree	Molecular Weight	H-Bond Donors	Rings	Rotatable Bonds	Polar Surface Area	measured log solubility in mols per liter	SMILES
Amigdalin	1	457,432	7	3	7	202,32	-0,77	OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)C(O)C3O
Fenfuram	1	201,225	1	2	2	42,24	-3,3	Cc1occc1C(=O)Nc2ccccc2
citral	1	152,237	0	0	4	17,07	-2,06	CC(C)=CCCC(C)=CC(=O)
Picene	2	278,354	0	5	0	0	-7,87	c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc43
Thiophene	2	84,143	0	1	0	0	-1,33	c1ccsc1
benzothiazole	2	135,191	0	2	0	12,89	-1,5	c2ccc1scnc1c2
2,2,4,6,6'-PCB	1	326,437	0	2	1	0	-7,32	Clc1cc(Cl)c(c(Cl)c1)c2c(Cl)cccc2Cl
Estradiol	1	272,388	2	4	0	40,46	-5,03	CC12CCC3C(CCc4cc(O)ccc34)C2CCC1O
Dieldrin	1	380,913	0	5	0	12,53	-6,29	ClC4=C(Cl)C5(Cl)C3C1CC(C2OC12)C3C4(Cl)C5(Cl)Cl
Rotenone	1	394,423	0	5	3	63,22	-4,42	COc5cc4OCC3Oc2c1CC(Oc1ccc2C(=O)C3c4cc5OC)C(C)=C
2-pyrrolidone	1	85,106	1	1	0	29,1	1,07	O=C1CCCN1
2-Chloronapthalene	1	162,619	0	2	0	0	-4,14	Clc1ccc2ccccc2c1
1-Pentene	1	70,135	0	0	2	0	-2,68	CCCC=C
Primidone	1	218,256	2	2	2	58,2	-2,64	CCC1(C(=O)NCNC1=O)c2ccccc2
Tetradecane	1	198,394	0	0	11	0	-7,96	CCCCCCCCCCCCCC
2-Chloropropane	1	78,542	0	0	0	0	-1,41	CC(C)Cl
2-Methylbutanol	1	88,15	1	0	2	20,23	-0,47	CCC(C)CO
Benzonitrile	1	103,124	0	1	0	23,79	-1	N#Cc1ccccc1
Diazinon	1	304,352	0	1	7	53,47	-3,64	CCOP(=S)(OCC)Oc1cc(C)nc(n1)C(C)C

2.2. Graph Neural Networks (Grafik Sinir Ağları)

GNNs represent an artificial intelligence approach that graphically depicts the molecular structures of chemical compounds and utilizes deep learning techniques to analyze these graphical structures. The molecular graph representations of chemical compounds encompass the bonds between atoms and molecular properties. GNNs possess the ability to make solubility predictions by processing this graphical structure [24].

GNNs offer several crucial features that assist in a better understanding of the molecular properties and chemical structures of chemical compounds. In particular, GNNs can make solubility predictions by considering the local structure of the molecular graph. This is highly valuable in evaluating the interactions and bonds between different atoms in chemical compounds. Additionally, they can take into account the topology and structural characteristics of chemical compounds.

Within the framework of this investigation, the pivotal function of GNNs is to augment and refine the regression model deployed for the anticipation of chemical compound solubility within the confines of the ESOL dataset. This model generates solubility predictions by utilizing the molecular graph representations of chemical compounds.

These predictions can contribute to the comprehension and improvement of critical parameters in drug design processes, such as the bioavailability and pharmacokinetics of drugs.

The integration of GNNs in this study serves the overarching purpose of catalyzing the evolution of pioneering solubility prediction methodologies within the spheres of pharmaceuticals and chemistry. This strategic inclusion aims to expedite and enhance the intricate processes involved in drug design. Consequently, a GNN-based solubility prediction model can serve as a valuable tool in drug development processes and assist in the more effective and safe design of drugs. Figure 2 shows the working structure of a GNN.

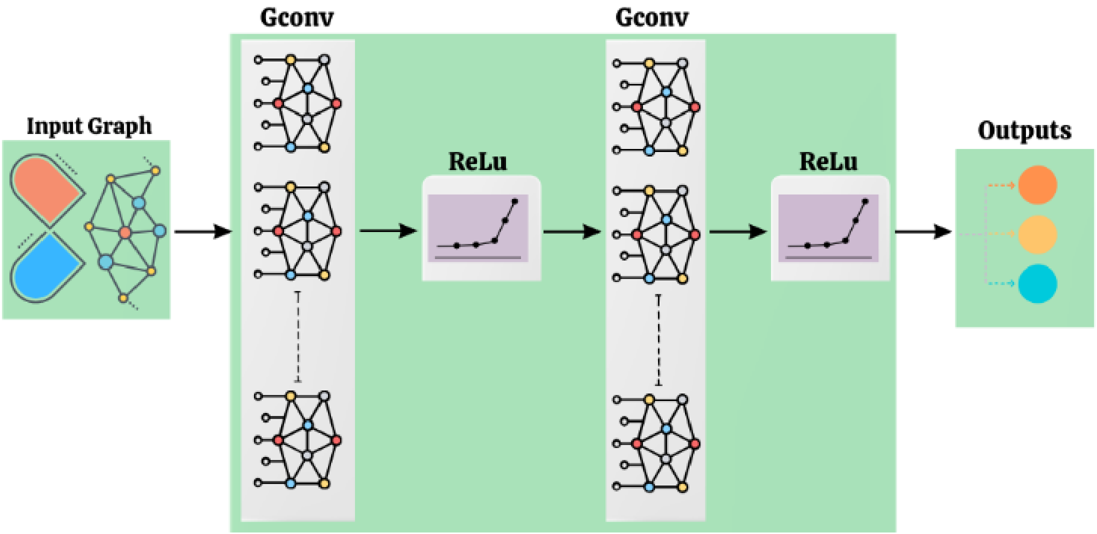


Figure 2. GNNs work structure (GNN’lerin çalışma yapısı)

2.3. Multilayer Perceptron (MLP) (Çok Katmanlı Algılayıcı)

Within the scope of this study, the MLP is employed as a deep learning model derived from artificial neural networks. MLP is a widely used regression and classification model in the fields of data mining and pattern recognition. It is known for its capability to learn complex functions and is particularly successful in prediction problems involving structured and unstructured data types.

The multi-layered architecture is a distinctive feature of MLP. MLP architecture comprises no fewer than three essential layers: an initial input-layer, followed by one or more hidden-layers, and culminating in output-layer. Neurons within these layers are fully interconnected, establishing connections with every neuron in both the preceding and succeeding layers. This structural characteristic augments the model's capacity to adeptly represent and learn intricate functions [6].

Every neuron undertakes the processing of incoming data through the application of an activation function. Frequently employed activation functions encompass sigmoid, Rectified Linear Unit (ReLU), and tanh. The selection of activation functions plays a crucial role in determining the outputs of neurons, facilitating the model's capacity to acquire and comprehend non-linear relationships.

Training MLP is typically performed using a process called backpropagation. This process involves comparing the model's predictions to actual values and propagating errors arising from this comparison backward between layers. This allows for the updating of weights and the training of the model. Figure 3 illustrates the structure of an MLP.

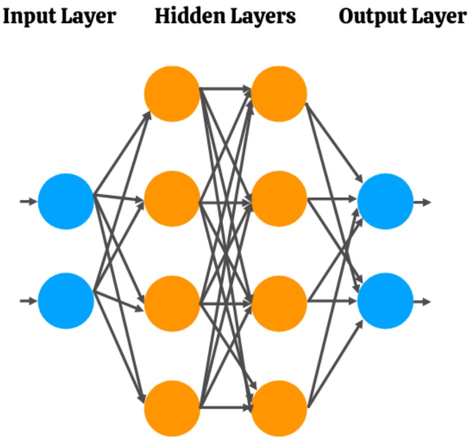


Figure 3. An MLP structure (Bir MLP yapısı)

2.1. Machine Learning (ML) (Makine Öğrenmesi)

This study encompasses a diverse array of ML algorithms utilized in the prediction of drug molecule solubility. These algorithms are employed in the fields of data mining and pattern recognition with the purpose of accomplishing the critical task of solubility prediction, which plays a significant role in drug development processes and drug design. Here is a brief description of some fundamental ML algorithms utilized in this study:

a. **Random Forest (RF)**: RF is an ensemble learning algorithm where multiple decision trees are constructed and combined. Each decision tree evaluates data samples independently and aggregates their results. This approach enhances the overall performance of the model while reducing the risk of overfitting [25].

b. **Gradient Boosting (GB)**: GB is a technique of ensemble learning where simple models, referred to as weak learners, are combined to create a powerful model. This method employs an iterative process where each new model attempts to correct the errors of the previous ones. GB can create a robust regression model capable of achieving high accuracy [26].

c. **XGBoost (Extreme Gradient Boosting):** XGBoost is a variation of GB and is often known for its high-performance and scalability. XGBoost is frequently preferred, especially in structured datasets and regression tasks like solubility prediction in tabular data [27].

These ML algorithms are integral to this study's objective of predicting drug solubility, aiding in the advancement of drug design processes, and contributing to pharmaceutical research.

3. PERFORMANCE METRICS AND THE RESEARCH FINDINGS (Performans Metrikleri ve Araştırma Bulguları)

Performance metrics are criteria used to assess how close a model's predictions are to the actual values. The metrics utilized in this study include:

1. **Root Mean Square Error (RMSE):** RMSE is a measure of how much a model's predictions deviate from the actual values. When calculating RMSE, the difference between each prediction and the actual value is computed, the squares of these differences are averaged, and finally, the square root of this value is taken. RMSE shares the same units as the predicted variable, making it directly interpretable. The mathematical formula is shown in Equation 1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i^{pre} - y_i^{exp})^2}{n}} \qquad (1)$$

2. **Mean Absolute Error (MAE):** MAE serves as a metric indicating the extent to which a model's predictions deviate from the actual values by utilizing the absolute values of these differences. It involves computing the average of the absolute differences between each prediction and its corresponding actual value. MAE provides insight into the average magnitude of deviations between predictions and actual values. The mathematical expression for MAE is presented in Equation 2.

$$MAE = \frac{\sum_{i=1}^{n} \left| y_i^{pre} - y_i^{exp} \right|}{n} \tag{2}$$

3. **R-squared (R^2):** R^2 measures how well a model explains the variance in the dataset. R^2, ranging between 0 and 1, signifies the goodness of fit of the model to the data, where higher values denote a superior fit. As R^2 approaches 1, the model explains the data well. However, R^2 can also be negative, indicating the model's failure to predict the data. The mathematical formula is shown in Equation 3.

$$R^2 = \sqrt{\frac{\sum_{i=1}^{n} (y_i^{pre} - y_i^{exp})^2}{\sum_{i=1}^{n} (y_i^{exp} - \bar{y})^2}} \tag{3}$$

These metrics are commonly used in evaluating the performance of regression models. RMSE and MAE determine the accuracy of predictions, while R^2 indicates how well the model fits the dataset. Utilizing these metrics is essential for assessing a model's effectiveness and accuracy, providing fundamental tools for understanding and comparing results.

The analyses were performed in PYTHON environment on a computer with 32 GB RAM and GPU RTX3060 graphics processor. The parameters of the models were chosen as epcoh=50, optimizer= Adam, learning rate= 0.01. The dataset was partitioned into distinct training and test subsets, with the training set encompassing 80% of the data and the test set constituting the remaining 20%. The subsequent assessment and measurement of outcomes were exclusively conducted on the designated test set. The findings of the study show the impact of different modeling approaches on drug resolution prediction. The results show that GNN and MLP perform poorly compared to other traditional machine learning methods. In particular, performance measures such as RMSE, MAE and R^2 reveal that the predictions made by GNN and MLP contain more errors compared to other models.

ML algorithms such as RF, GB, and XGBoost achieved better results in this regression task. These outcomes suggest that traditional machine learning methods are more effective for drug solubility prediction compared to AI-based approaches like GNN and MLP.

However, it's essential to note that results can vary depending on numerous factors, including dataset quality, feature engineering, hyperparameter tuning, and other considerations. Therefore, further research and consideration of different model structures and data characteristics may be necessary for improving drug solubility prediction.

Table 2. The solubility prediction error amounts of the models (Modellerin çözünürlük tahmin hata miktarları)

Models	RMSE	MAE	R^2
Random Forest	**1.2145**	0.9221	0.6575
Gradient Boosting	1.2955	1.0002	0.6102
MLP	1.1473	**0.8586**	0.6943
GNN	1.8389	1.4684	**0.2147**

The results in Table 2 indicate that RF and MLP exhibited superior performance in drug solubility prediction compared to other models. Lower RMSE and MAE values along with higher R^2 scores signify the better performance of these models in predicting drug solubility. Specifically, MLP achieved the lowest RMSE and the highest R^2 score.

In contrast, the GNN model showed lower performance compared to other models. High RMSE and MAE values along with a low R^2 score suggest that the GNN model contained more errors in drug solubility prediction and was less successful compared to other models.

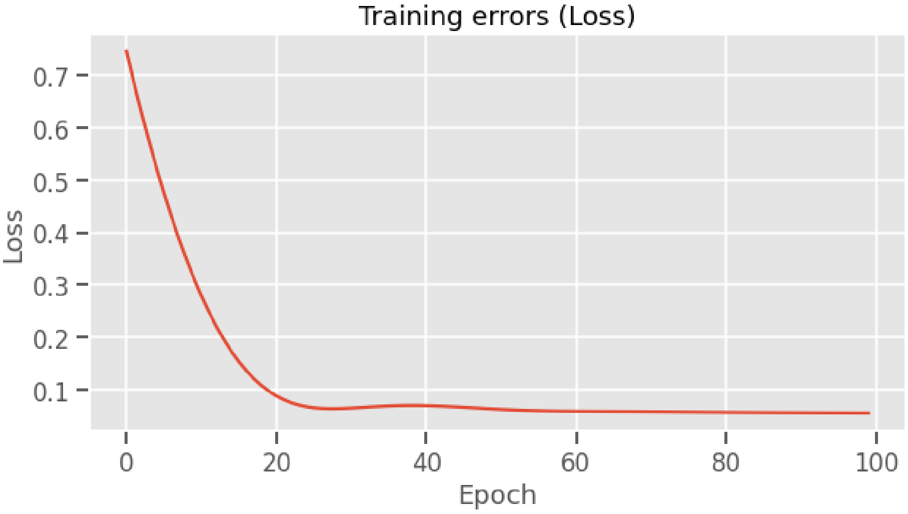


Figure 4. Loss values during training (Eğitim sırasındaki kayıp değerleri)

Figure 4 illustrates the loss values during training. After 100 iterations, the training loss value was completed at 0.082. A trained GNN model can be utilized to predict the solubility of new molecules.

Table 2 demonstrate that the RF model had the lowest error rates compared to other models. This indicates that the RF model predicted solubility more accurately and provided results closer to the actual values. On the other hand, the GNN model exhibited higher error rates compared to other models, suggesting that it made less accurate predictions for this specific dataset and needs improvement.

This graph aids in a clear comparison of the performance of each model and helps us understand which model excelled in drug solubility prediction and which one has more room for improvement. This information guides the selection of the appropriate model in drug design and development processes.

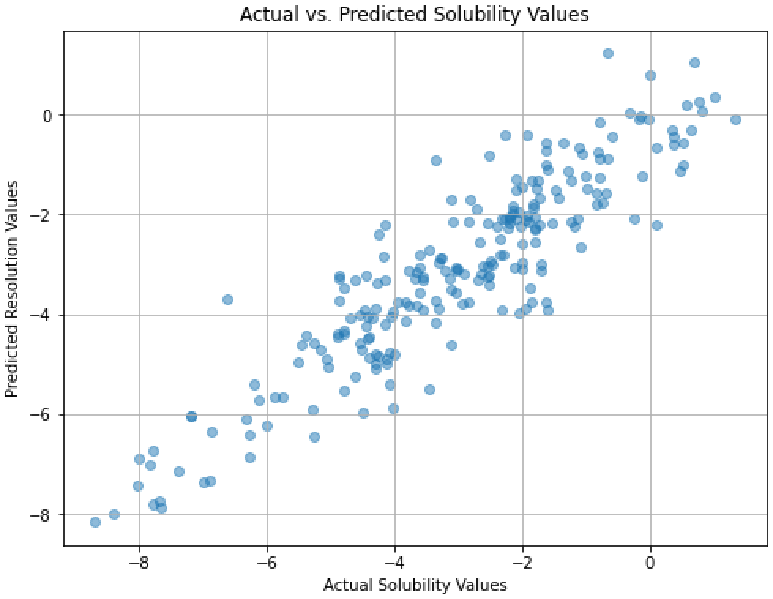


Figure 5. Graphical representation of actual and predicted solubility values (Gerçek ve tahmin edilen çözünürlük değerlerinin grafiksel gösterimi)

The real and predicted solubility graphs are essential tools for evaluating the performance of different models used in drug solubility prediction. Figure 5 visually represents how well each model predicts the actual solubility and the extent of deviation between their predictions and the actual values.

4.DISCUSSION AND CONCLUSION (Tartışma ve Sonuç)

This study was conducted to evaluate the performance of different ML algorithms in predicting drug solubility. Solubility prediction holds critical importance in drug development processes and pharmaceutical design. Therefore, accurate and reliable solubility prediction models play a significant role in the success of the pharmaceutical industry.

At the outset of the study, the ESOL dataset was employed. This dataset encompasses experimentally measured solubility values of various drug molecules. In the preprocessing and preparation phase of the dataset, molecular graph features were extracted, and the data was partitioned into training and test sets.

In the first stage of the study, the GNN model was utilized. GNN is a AI-based approach used for drug solubility prediction. However, when the performance of the GNN model was compared to other ML models in this study, it was found to be inferior. Upon examination of performance metrics such as RMSE, MAE, and R^2, it was observed that the GNN model's predictions contained more errors compared to other models. These results indicate that the GNN model was not effective for solubility prediction in this specific dataset.

In the second stage, the MLP model and traditional ML algorithms (RF, GB, XGBoost) were employed. These models are commonly used ML approaches for regression tasks. The results show that these traditional ML algorithms achieved higher success in solubility prediction. Particularly, algorithms like RF, GB, and XGBoost performed better with lower RMSE and MAE values and higher R^2 scores, indicating more accurate predictions.

These findings suggest that traditional ML algorithms may be more effective than AI-based approaches (such as GNN) in regression tasks like drug solubility prediction. However, these results can vary depending on factors like the dataset, feature engineering, hyperparameter tuning, and model selection. Therefore, further research and consideration of different model structures and data characteristics may be necessary for drug solubility prediction.

In conclusion, this study provides a comparative analysis of ML and artificial intelligence models used in drug solubility prediction. It emphasizes that traditional ML algorithms may perform better for a specific dataset and should be considered in model selection. Accurate solubility predictions can offer a significant advantage in the design and discovery of new drugs in pharmaceutical development processes.

ACKNOWLEDGMENTS (TEŞEKKÜR)

If the study was supported by a research institution or fund, the name of the supporting institution/organization and the project number should be written. / Çalışma araştırma kurumu veya fonu tarafından desteklenmiş ise destek sağlayan kurum/kuruluş adı ve proje numarası yazılmalıdır.

DECLARATION OF ETHICAL STANDARDS (ETİK STANDARTLARIN BEYANI)

The author of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

Bu makalenin yazarı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

AUTHORS’ CONTRIBUTIONS (YAZARLARIN KATKILARI)

***Veysel GİDER*:** He contributed to the implementation of the study, analysis of the results and writing of the manuscript.

Araştırmanın uygulanmasına, sonuçların analizine ve makalenin yazılmasına katkıda bulunmuştur.

***Cafer BUDAK*:** He contributed to the design and implementation of the study

Araştırmanın tasarımına ve uygulanmasına katkıda bulunmuştur.

CONFLICT OF INTEREST (ÇIKAR ÇATIŞMASI)

There is no conflict of interest in this study.

Bu çalışmada herhangi bir çıkar çatışması yoktur.

REFERENCES (KAYNAKLAR)

[1] Prieto-Martínez, F. D., López-López, E., Juárez-Mercado, K. E., & Medina-Franco, J. L. (2019). Computational drug design methods—current and future perspectives. *In silico drug design*, 19-44.

[2] Barrett, Jaclyn A., et al. "Discovery solubility measurement and assessment of small molecules with drug development in mind." Drug Discovery Today 27.5 (2022): 1315-1325.

[3] Vora, Lalitkumar K., et al. "Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design." Pharmaceutics 15.7 (2023): 1916.

[4] Budak, Cafer, Vasfiye Mençik, and Veysel Gider. "Determining similarities of COVID-19–lung cancer drugs and affinity binding mode analysis by graph neural network-based GEFA method." Journal of Biomolecular Structure and Dynamics 41.2 (2023): 659-671.

[5] Gider, Veysel, and Cafer Budak. "Instruction of molecular structure similarity and scaffolds of drugs under investigation in ebola virus treatment by atom-pair and graph network: A combination of favipiravir and molnupiravir." Computational biology and chemistry 101 (2022): 107778.

[6] Gardner, Matt W., and S. R. Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." Atmospheric environment 32.14-15 (1998): 2627-2636.

[7] Hu, Pingfan, et al. "Development of solubility prediction models with ensemble learning." Industrial & Engineering Chemistry Research 60.30 (2021): 11627-11635.

[8] Selvaraj, Chandrabose, Ishwar Chandra, and Sanjeev Kumar Singh. "Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries." Molecular diversity (2021): 1-21.

[9] Kherouf, Soumaya, et al. "Modeling of linear and nonlinear quantitative structure property relationships of the aqueous solubility of phenol derivatives." *Journal of the Serbian Chemical Society* 84.6 (2019): 575-590.

[10] Eros, Daniel, et al. "Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS and ANN methods." *Mini Reviews in Medicinal Chemistry* 4.2 (2004): 167-177.

[11] Sinha, Priyanka, et al. "Integrating Machine Learning and Molecular Simulation for Material Design and Discovery." *Transactions of the Indian National Academy of Engineering* 8.3 (2023): 325-340.

[12] Reiser, Patrick, et al. "Graph neural networks for materials science and chemistry." Communications Materials 3.1 (2022): 93.

[13] Qin, Yongfei, et al. "MLP-based regression prediction model for compound bioactivity." Frontiers in Bioengineering and Biotechnology 10 (2022): 946329.

[14] Ahmad, Waqar, Hilal Tayara, and Kil To Chong. "Attention-Based Graph Neural Network for Molecular Solubility Prediction." *ACS omega* 8.3 (2023): 3236-3244.

[15] Lee, Sangho, et al. "Multi-order graph attention network for water solubility prediction and interpretation." *Scientific Reports* 13.1 (2023): 957.

[16] Hamdi, Mohammad Erfan, et al. "Prediction of Aqueous Solubility of Drug Molecules by Embedding Spatial Conformers Using Graph Neural Networks." *2022 29th National and 7th International Iranian Conference on Biomedical Engineering (ICBME)*. IEEE, 2022.

[17] Ge, Kai, and Yuanhui Ji. "Novel computational approach by combining machine learning with molecular thermodynamics for predicting drug solubility in solvents." *Industrial & Engineering Chemistry Research* 60.25 (2021): 9259-9268.

[18] Alzhrani, Rami M., Atiah H. Almalki, and Sameer Alshehri. "Novel numerical simulation of drug solubility in supercritical CO2 using machine learning technique: Lenalidomide case study." *Arabian Journal of Chemistry* 15.11 (2022): 104180.

[19] Sadeghi, Arash, et al. "Machine learning simulation of pharmaceutical solubility in supercritical carbon dioxide: Prediction and experimental validation for busulfan drug." *Arabian Journal of Chemistry* 15.1 (2022): 103502.

[20] Meng, Di, and Zhenyu Liu. "Machine learning aided pharmaceutical engineering: Model development and validation for estimation of drug solubility in green solvent." *Journal of Molecular Liquids* 392 (2023): 123286.

[21] Li, Mengshan, et al. "Prediction of the aqueous solubility of compounds based on light gradient boosting machines with molecular fingerprints and the cuckoo search algorithm." ACS omega 7.46 (2022): 42027-42035.

[22] Sadybekov, Anastasiia V., and Vsevolod Katritch. "Computational approaches streamlining drug discovery." *Nature* 616.7958 (2023): 673-685.

[23] KAGGLE, Online (2023). https://www.kaggle.com/code/mmelahi/physical-chemistry-esol/input Access: 02.09.2023.

[24] Gong, Weiyi, and Qimin Yan. "Graph-based deep learning frameworks for molecules and solid-state materials." Computational Materials Science 195 (2021): 110332.

[25] Liu, Yanli, Yourong Wang, and Jian Zhang. "New machine learning algorithm: Random forest." *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*. Springer Berlin Heidelberg, 2012.

[26] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232.

[27] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." *Artificial Intelligence Review* 54 (2021): 1937-1967.