INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

Web Scraping tool for extracting the abstracts from Research Papers available on Google Scholar and PubMed

Group 16
Suraj Kumar Jha - 2021209
Rajat jaiswal - 2021184
Tarun Bansal - 2021210

Various ML/DL models reported in the literature for the property : Permeability Prediction and Blood Barrier Permeability

- Title
- Author
- Year of publication
- Dataset size
- Available for download or not
- Which machine learning or deep learning model
- In the case of multiple models which is the one top performing
- Source reference and URL

Aim: To analyze scientific literature from PubMed for machine learning and deep learning models used in permeability prediction and other properties. The significance of permeability prediction and the need for web scraping.

Methodology

- Web scraping PubMed and Google Scholar
- Extracting information: Title, Year, Authors,Dataset size, Models Used, Download availability, Top performing Model
- Code snippet explanation
- Data preparation for analysis

Approach

- Utilizing Python with Selenium and BeautifulSoup libraries
- Model keywords for identification
- Iterating through search results
- Navigating to article pages
- Organizing data by year

The code snippet scrapes PubMed for articles on permeability prediction from 2010 to 2023, extracting data for machine learning and deep learning model mentions. It iterates through search results, simulates scrolling, and parses the HTML for analysis.

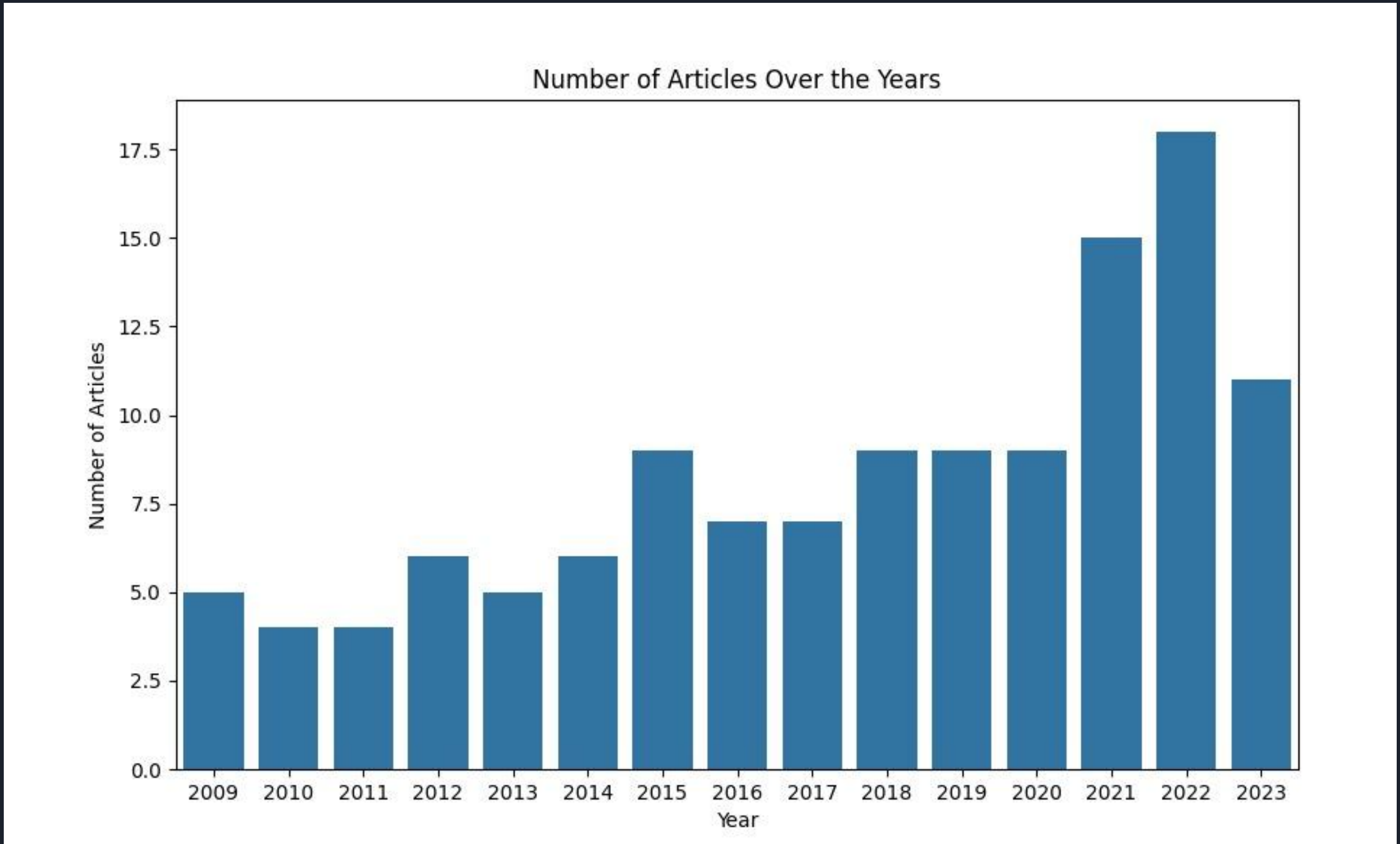
```python
# Loop through each year from 2010 to 2023
for year in range(2010, 2024):
    year_data = []

    # Inside the loop for each page
    for page in range(1, 5):  # Extract data from the first 10 pages
        # Open PubMed
        driver.get(f"https://pubmed.ncbi.nlm.nih.gov/?term={search_query}&filter=years.{year}-{year}&page={pa

        # Wait for the results to load
        time.sleep(5)

        # Scroll down to load more results (you can adjust the number of scrolls)
        for _ in range(3):
            driver.execute_script("window.scrollTo(0, document.body.scrollHeight);")
            time.sleep(2)

        # Get the page source and parse it with BeautifulSoup
        page_source = driver.page_source
        soup = BeautifulSoup(page_source, 'html.parser')
```

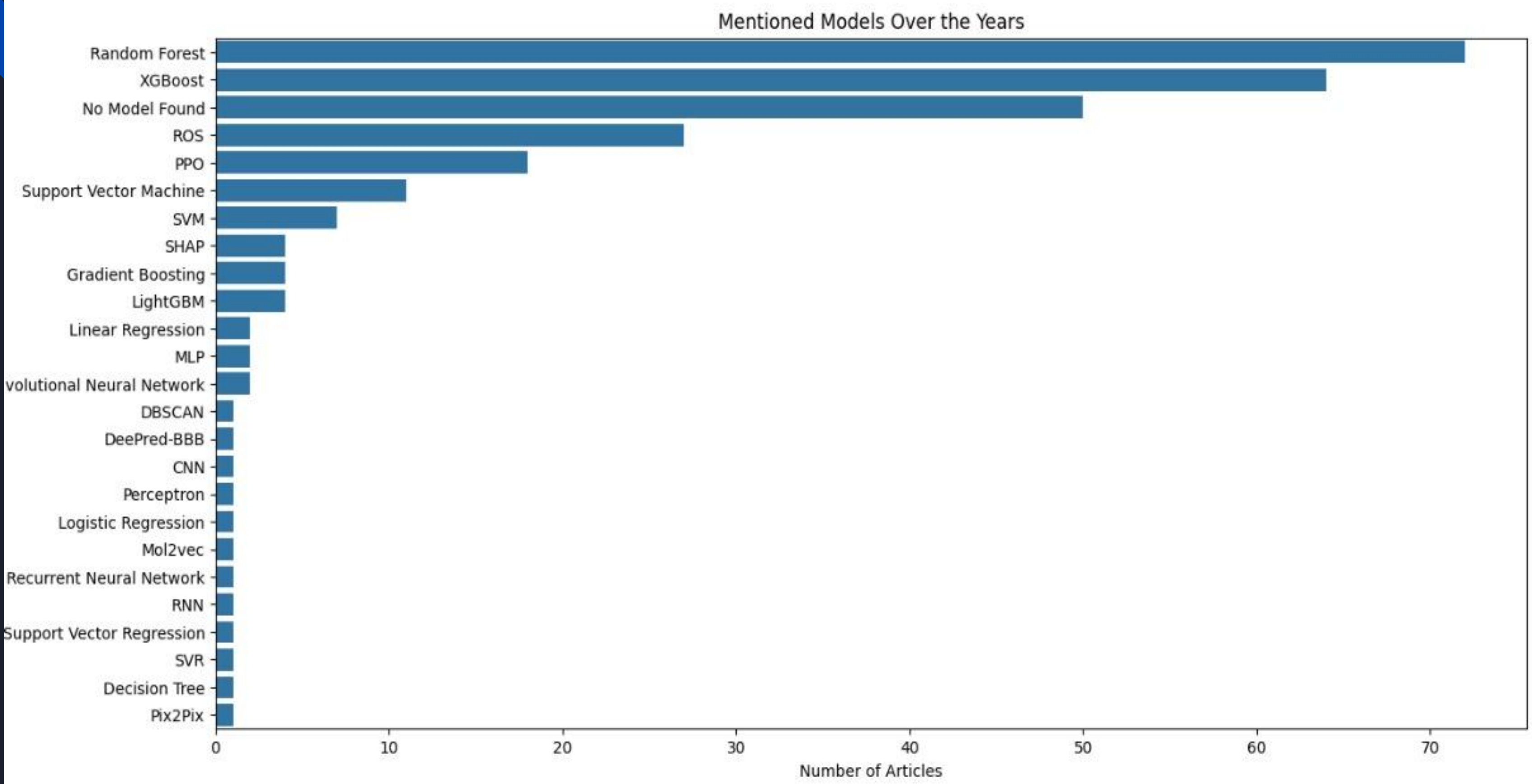
The code segment extracts article details and searches for downloadable articles. It checks if specific machine learning models are mentioned in the abstract or text, facilitating model identification in PubMed articles.

```
for result in results:
    title = result.find("a", {"class": "docsum-title"}).text
    authors = result.find("span", {"class": "docsum-authors"}).text
    link = "https://pubmed.ncbi.nlm.nih.gov" + result.find("a", {"class": "docsum-title"})["href"]
    pmid = result.find("span", {"class": "docsum-pmid"}).text

    # Open the link to the article
    driver.get(link)
    time.sleep(5)  # Wait for the article page to load

    # Get the article page source and parse it with BeautifulSoup
    article_page_source = driver.page_source
    article_soup = BeautifulSoup(article_page_source, 'html.parser')

    # Check if the "Save" button element is present inside the article
    save_button = article_soup.find("button", {"id": "save-results-panel-trigger"})
    downloadability = "Yes" if save_button else "No"
    abstract = result.find("div", {"class": "abstract-content"}).text

    # Extract the entire text content of the article page
    article_text = article_soup.get_text()
    article_upper = article_text.upper()

    abstract_upper = abstract.upper()
    models_mentioned = []
```

Function to Find the dataset used

```python
def find_data_size(text):
    # Split the text into lines
    lines = text.split('. ')

    max_value = None
    line_with_max_value = None

    for line in lines:
        # Use regular expressions to find all numeric values in the line
        numeric_values = re.findall(r'\d+(?:\.\d+)?', line)

        if numeric_values:
            # Convert the numeric values to float and find the maximum
            line_max_value = max(float(value) for value in numeric_values)

            if max_value is None or line_max_value > max_value:
                max_value = line_max_value
                line_with_max_value = line

    return line_with_max_value
```


Number of Articles Over the Years
Number of Articles
Year
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023



Mentioned Models Over the Years
Random Forest
XGBoost
No Model Found
ROS
PPO
Support Vector Machine
SVM
SHAP
Gradient Boosting
LightGBM
Linear Regression
MLP
volutional Neural Network
DBSCAN
DeePred-BBB
CNN
Perceptron
Logistic Regression
Mol2vec
Recurrent Neural Network
RNN
Support Vector Regression
SVR
Decision Tree
Pix2Pix
0
10
20
30
40
50
60
70
Number of Articles

Permeability Prediction Results (Extracted to Excel File)

1	Title	Year	Authors	Link	Model	Downloadability
2	Reliable Prediction of Caco-2 Permeability by Supervised Recu	2022	Falcón-Cano G, Molina C, Cal	https://pubmed.ncbi.nlm.nih.gov/36297432/	Random Forest, KNIME, CLIP, ROS, Random Forest	Yes
3	DeePred-BBB: A Blood Brain Barrier Permeability Prediction M	2022	Kumar R, Sharma A, Alexiou A	https://pubmed.ncbi.nlm.nih.gov/35592264/	DeePred-BBB, CNN, Convolutional Neural Network, Recurrent Neural Ne	Yes
4	DeepBBBP: High Accuracy Blood-brain-barrier Permeability Pre	2022	Cherian Parakkal S, Datta R, I	https://pubmed.ncbi.nlm.nih.gov/35393777/	Mol2vec, MLP, Convolutional Neural Network, Perceptron, CLIP, PPO, RC	Yes
5	Chloride Permeability Coefficient Prediction of Rubber Concre	2022	Huang X, Wang S, Lu T, Li H, V	https://pubmed.ncbi.nlm.nih.gov/36679189/	Random Forest, Linear Regression, Decision Tree, Extreme Learning Mac	Yes
6	Binary classification model of machine learning detected alter	2022	Rahman Z, Pasam T, Rishab, I	https://pubmed.ncbi.nlm.nih.gov/35758006/	SVM, VGG, CLIP, SAC, PPO, ROS, Cortex	Yes
7	A merged molecular representation deep learning method for	2022	Tang Q, Nie F, Zhao Q, Chen V	https://pubmed.ncbi.nlm.nih.gov/36002937/	DeePred-BBB, Support Vector Machine, LightGBM, CLIP, PPO, ROS, Light(	Yes
8	Trivariate Linear Regression and Machine Learning Prediction	2022	Shimizu M, Hayasaka R, Kami	https://pubmed.ncbi.nlm.nih.gov/35644566/	Linear Regression, Gradient Boosting, LightGBM, CLIP, ROS, LightGBM, Gr	Yes
9	Ensemble modeling with machine learning and deep learning t	2022	Yu TH, Su BH, Battalora LC, Li	https://pubmed.ncbi.nlm.nih.gov/34530437/	SVM, Support Vector Machine, Orange, CLIP, PPO, ROS	Yes
10	Quantifying face mask comfort.	2022	Koh E, Ambatipudi M, Boone	https://pubmed.ncbi.nlm.nih.gov/34747682/	Linear Regression, CLIP, SAC, ROS	Yes
11	Revolutionizing Membrane Design Using Machine Learning-Ba	2022	Gao H, Zhong S, Zhang W, Igo	https://pubmed.ncbi.nlm.nih.gov/34968041/	Gradient Boosting, LIME, CLIP, PPO, ROS, Gradient Boosting, LIME	Yes
12	In Silico Prediction of Skin Permeability Using a Two-QSAR App	2022	Wu YW, Ta GH, Lung YC, Wer	https://pubmed.ncbi.nlm.nih.gov/35631545/	BERT, SVR, Support Vector Regression, Bert, CLIP, PPO	Yes
13	Ensemble learning for predicting ex vivo human placental barri	2022	Chou CY, Lin P, Kim J, Wang S	https://pubmed.ncbi.nlm.nih.gov/36138350/	Random Forest, Linear Regression, CLIP, PPO, ROS, Random Forest	Yes
14	Biological Membrane-Penetrating Peptides: Computational Pr	2022	de Oliveira ECL, da Costa KS,	https://pubmed.ncbi.nlm.nih.gov/35402305/	Support Vector Machine, CLIP, PPO, ROS	Yes
15	Machine learning-based models for predicting gas breakthrough pressure of po	2022	Gao C, Lu PH, Ye WM, Liu ZR,	https://pubmed.ncbi.nlm.nih.gov/36538229/	BERT, Random Forest, Bert, SHAP, CLIP, ROS, Random Forest, SHAP	Yes
16	Prediction of organic contaminant rejection by nanofiltration and reverse osm	2022	Zhu T, Zhang Y, Tao C, Chen V	https://pubmed.ncbi.nlm.nih.gov/36228787/	SVM, XGBoost, LightGBM, CLIP, LightGBM, XGBoost	Yes
17	Blood-brain barrier penetration prediction enhanced by uncer	2022	Tong X, Wang D, Ding X, Tan X	https://pubmed.ncbi.nlm.nih.gov/35799215/	MLP, CLIP	Yes
18	Membrane Permeating Macrocycles: Design Guidelines from M	2022	Williams-Noonan BJ, Speer M	https://pubmed.ncbi.nlm.nih.gov/36178379/	Random Forest, Linear Regression, CLIP, PPO, ROS, Random Forest	Yes
19	Implications of Additivity and Nonadditivity for Machine Learn	2022	Kwapien K, Nittinger E, He J, I	https://pubmed.ncbi.nlm.nih.gov/35936431/	Orange, CLIP, PPO	Yes
20	Physics-informed machine learning with differentiable progran	2022	Pachalieva A, O'Malley D, Ha	https://pubmed.ncbi.nlm.nih.gov/36333378/	Convolutional Neural Network, BERT, Bert, CLIP, PPO	Yes
21	The Role of Different Retinal Imaging Modalities in Predicting I	2022	Elsharkawy M, Elrazzaz M, Sh	https://pubmed.ncbi.nlm.nih.gov/35591182/	Transformer, CLIP, ROS	Yes
22	Machine learning enables interpretable discovery of innovative polymers for ga	2022	Yang J, Tao L, He J, McCutche	https://pubmed.ncbi.nlm.nih.gov/35857839/	SHAP, CLIP, SHAP	Yes
23	Using in vitro ADME data for lead compound selection: An em	2022	Williams J, Siramshetty V, Ng	https://pubmed.ncbi.nlm.nih.gov/35030421/	CNN, Convolutional Neural Network, Random Forest, Decision Tree, Gra	Yes
24	Decoding river pollution trends and their landscape determina	2022	Xu G, Fan H, Oliver DM, Dai Y	https://pubmed.ncbi.nlm.nih.gov/35931190/	Random Forest, XGBoost, Gradient Boosting, SHAP, CLIP, PPO, XGBoost,	Yes
25	A general optimization protocol for molecular property predic	2022	Chen JH, Tseng YJ.	https://pubmed.ncbi.nlm.nih.gov/34498673/	CNN, Convolutional Neural Network, RNN, LSTM, CLIP, PPO	Yes
26	Prediction of irrigation groundwater quality parameters using ANN, LSTM, and	2022	Kouadri S, Pande CB, Pannee	https://pubmed.ncbi.nlm.nih.gov/34748181/	LSTM, Long Short-Term Memory, Linear Regression, CLIP	Yes
27	Better Performance with Transformer: CPPFormer in the Preci	2022	Xue Y, Ye X, Wei L, Zhang X, S	https://pubmed.ncbi.nlm.nih.gov/34544332/	Decision Tree, Transformer, CLIP	Yes
28	Machine Learning-Based Accelerated Approaches to Infer Breakdown Pressure	2022	Tariq Z, Yan B, Sun S, Gudala	https://pubmed.ncbi.nlm.nih.gov/36406508/	Random Forest, Decision Tree, CLIP, ROS, Random Forest	Yes
29	Conformational Effects on the Passive Membrane Permeabilit	2022	Rzepiela AA, Viarengo-Baker	https://pubmed.ncbi.nlm.nih.gov/35861996/	CLIP, SAC, ROS	Yes

... 2014 2015 2016 2017 2018 2019 2020 2021 **2022** 2023

Ready Accessibility: Good to go 100%

Blood Barrier Permeability Results
(Year wise Data Extracted to Excel File)

	A	B	C	D	E	F	G	H	I
1	Title	PMID	Year	Authors	Link	Model	Downloadability	Abstract	Dataset Si:
2	Machine learning based dynamic consen	37137267	2022	Mazumdar	https://pubmed.ncbi.nlm.nih.gov	Random Forest, XGBoost, XGBoost, Random Forest	Yes	The blood-brain barrier (BBB) is an important defence	A dataset
3	A machine learning-based quantitative m	37713469	2022	Shaker B, e	https://pubmed.ncbi.nlm.nih.gov	No Model Found	Yes	Motivation: Efficient assessment of the bloo	The mode
4	DeepBBBP: High Accuracy E	35393777	2022	Cherian Pa	https://pubmed.ncbi.nlm.nih.gov	Mol2vec, MLP, Convolutional Neural Network, Perceptro	Yes	Blood-brain-barrier permeability (BBBP) is an importar	In this wo
5	DeePred-BBB: A Blood Brai	35592264	2022	Kumar R, S	https://pubmed.ncbi.nlm.nih.gov	DeePred-BBB, Convolutional Neural Network	Yes	The blood-brain barrier (BBB) is a selective and semipe	Each com
6	A merged molecular repres	36002937	2022	Tang Q, Ni	https://pubmed.ncbi.nlm.nih.gov	ROS	Yes	The ability of a compound to permeate across the blo	To compl
7	Blood-brain barrier penetration prediction e	35799215	2022	Tong X, Wa	https://pubmed.ncbi.nlm.nih.gov	No Model Found	Yes	Blood-brain barrier is a pivotal factor to be considered	In particu
8	Biological Membrane-Pene	35402305	2022	de Oliveira	https://pubmed.ncbi.nlm.nih.gov	ROS	Yes	Peptides comprise a versatile class of biomolecules th	Cell-pene
9	Alvascience: A New Softwa	36361669	2022	Mauri A, B	https://pubmed.ncbi.nlm.nih.gov	No Model Found	Yes	Quantitative structure-activity relationship (QSAR) and	The result
10	Relational graph convolutic	35561199	2022	Ding Y, Jiar	https://pubmed.ncbi.nlm.nih.gov	LightGBM, LightGBM	Yes	Motivation: Evaluating the blood-brain barri	Our mode
11	Proteomic biomarkers of Kl	35859339	2022	Hédou J, C	https://pubmed.ncbi.nlm.nih.gov	ROS	Yes	Study objectives: Kleine-Levin syndrome (KLS	We quant
12	Ensemble modeling with m	34530437	2022	Yu TH, Su E	https://pubmed.ncbi.nlm.nih.gov	Support Vector Machine, PPO	Yes	The trade-off between a machine learning (ML) and de	A data se
13	MORPHIOUS: an unsupervi	35093113	2022	Silburt J, A	https://pubmed.ncbi.nlm.nih.gov	Support Vector Machine, DBSCAN, PPO	Yes	Background: In conditions of brain injury and	MORPHIC
14	Machine learning based dynamic consen	37137267	2022	Mazumdar	https://pubmed.ncbi.nlm.nih.gov	Random Forest, XGBoost, XGBoost, Random Forest	Yes	The blood-brain barrier (BBB) is an important defence	A dataset
15	A machine learning-based quantitative m	37713469	2022	Shaker B, e	https://pubmed.ncbi.nlm.nih.gov	No Model Found	Yes	Motivation: Efficient assessment of the bloo	The mode
16	A general optimization prot	34498673	2022	Chen JH, T	https://pubmed.ncbi.nlm.nih.gov	CNN	Yes	The key to generating the best deep learning model for predictin	
17	deepGraphh: AI-driven web	35868454	2022	Gautam V,	https://pubmed.ncbi.nlm.nih.gov	PPO, ROS	Yes	Artificial intelligence (AI)-based computational techniques allow	
18	Implication of type 4 NADP	34922273	2022	Luengo E,	https://pubmed.ncbi.nlm.nih.gov	ROS	Yes	Aggregates of the microtubule-associated protein tau	Our result
19	Comparison of Descriptor-	35755260	2022	Orosz Á, H	https://pubmed.ncbi.nlm.nih.gov	MLP, XGBoost, XGBoost	Yes	The screening of compounds for ADME-Tox targets pla	In this stu
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									

2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | **2022** | 2023

Observation

Web scraping revealed insights about permeability prediction models, highlighting:

- Dataset size information, extracted through careful reading of abstract.
- All Articles are not downloadable.
- Diverse ML and DL models, some more popular with some of them are best performing.
- Year-wise data on articles, authors, models, and download status.

Conclusion

Our web scraping and analysis have provided valuable insights into the utilization of machine learning and deep learning models for permeability prediction in scientific literature. While dataset sizes were not directly available, but managed to extract data from reading abstract. we observed variable article download availability and a diverse array of models mentioned. Some models were more prevalent than others. The report's year-wise breakdown offers a comprehensive overview. Future studies could further explore top-performing models, investigate dataset sizes' impact, and extend the analysis to properties beyond permeability, expanding our understanding of model applications in pharmaceutical research.

THANK YOU

GROUP 16

SURAJ KUMAR JHA - suraj21209@iiitd.ac.in

RAJAT JAISWAL - rajat21184@iiitd.ac.in

TARUN BANSAL - tarun21210@iiitd.ac.in