ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 05, May 2025

Classification of Oncogenic Compounds in Consumer Products using ChemBERTa & OCR

[1] Kundanika Pradhan, [2] Muhammed Anish, [3] V. Bibin Christopher

[1] [2] [3] Department of Computing Technologies, SRM University, Kattankulathur, Tamil Nadu, India
Emails ID: [1] pradhankundanika@gmail.com, [2] muhammedanish3464@gmail.com, [3] bibinchv@srmist.edu.in

***Abstract*— Guaranteeing consumer safety and regulatory compliance for the cosmetics and personal care market involves thoroughly examining product ingredients. Conventional methods proved inadequate in detecting hazardous materials accurately and with high efficiency. A new method of detecting oncogenic compounds from cosmetic and personal care products using OCR and deep learning is introduced in this paper. The OCR retrieves ingredient information from product labels and packaging and then analyzes it through ChemBERTa, a transformer model trained on chemical representations. A specially developed module retrieves the SMILES representation of each ingredient extracted via an API-based mechanism. ChemBERTa is compared with the usual machine learning classifiers, such as Support Vector Machine (SVM), Random Forest, Decision Trees, Bagging, and XGBoost, in this work. ChemBERTa is a better option than traditional classifiers, and it has better accuracy when predicting the carcinogenicity of chemical compounds. An interface that is easy-to-use has been deployed using Streamlit that combines ChemBERTa and Llama 3.2 to present an informative experience for users. These outcomes show us the promise of deep learning in enhancing harmful chemical detection and classification, offering a powerful tool for safer consumer goods.**

***Index Terms*— ChemBERTa, deep learning, oncogenic chemicals, optical character recognition.**

I. INTRODUCTION

The widespread adoption of packaged consumer goods has provided reasons for concerns about oncogenic ingredients contained in them, which may lead to cancer. Detection of these toxic compounds is therefore crucial to ensure public health protection. Conventional methods of chemical analysis need large amounts of time and resources, which are impractical for mass monitoring. Machine learning (ML) models especially ChemBERTa present cancer causing chemicals, optical character recognition, chemical informatics, Streamlit, and Llama 3.2 as an effective solution, automating detection and enhancing efficiency and accuracy.

Nonetheless, conventional ML models like Random Forest and Support Vector Machines are based on pre-defined molecular descriptors that might not adequately reflect the chemical interaction complexities. Contrarily, transformer-based deep learning models like ChemBERTa provide a sophisticated alternative by using self-attention mechanisms to learn complex molecular representations from SMILES strings directly. As the dataset does not store SMILES itself, we incorporate an API-based method to dynamically retrieve the SMILES representations. This work proposes an integrated framework consisting of OCR for text retrieval, an API-based SMILES retrieval module, and ChemBERTa for the classification of oncogenic compounds, presenting a scalable and precise solution to chemical safety evaluation.

II. RELATED WORK

In Paper [1], the authors explore the creation of carcinogen prediction models employing a mongrel neural network approach, HNN-Cancer, to identify chemical oncogenes. A new SMILES point representation is integrated into the model. The RandomForest and Bagging styles, grounded on HNN- Cancer, achieved an oncogene delicacy of 74 and an AUC of 0.81, establishing the prognostications as largely dependable. The HNN-Cancer model itself achieved a good micro AUC, accuracy and micro accuracy. In addition, the models were created to estimate the pTD50 of chemicals, with agreement validation yielding an overall R^2 of 0.40 by comprising the results across styles. Despite encompassing different chemical orders and data sources, the models successfully prognosticate binary, multiclass, and quantitative oncogenicity, aligning with other models in the literature that concentrated on lower, more homogenous datasets. This paper shows us that HNN-Cancer is suitable for prognosticating implicit carcinogenic traits in a wide variety of chemicals.

In Paper [2], prognosticating medicine seeker toxin is emphasized as a pivotal aspect of medicine discovery, impacting costs, late-stage failures, and medicine recessions. While machine literacy(ML) models have limitations, they offer a promising approach for the early discovery of poisonous composites in medicine discovery. As high-quality data scarcity increases and the connection of ML styles broadens, their integration into the medicine discovery channel will improve. Recent advancements have enhanced the understanding of ML model literacy, particularly in assessing neural unit activation and crucial features, but there's still room for enhancement, especially in combining

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 05, May 2025

network pharmacology with ML to address toxin estimation challenges.

Paper [3] focuses on the operation of machine literacy in the food industry, relating to the generally used algorithms and their separate pros and cons. The study's end is unique in its comparison of multiple ML algorithms across colorful studies, enabling the identification of top-performing styles. The Random Forest Classifier and Random Forest algorithms are stressed as the most effective for working problems in the food industry. Overall, ML holds significant importance in addressing a wide range of issues in the food industry, from product to marketing, improving effectiveness, and fostering innovation.

Paper [4] discusses how recent advances in AI algorithm design have opened up opportunities to break down problems across multiple disciplines. In cheminformatics and medicine discovery, machine learning models have greatly served the pharmaceutical industry, particularly in relating strong relations for new therapeutic targets, aligning with perfect drug pretensions.

Paper [5] reviews the use of machine learning models in prognosticating colorful medicine toxin endpoints, including oncogenicity, mutagenicity, hepatotoxicity, acute oral toxin, and hERG inhibition. Popular algorithms such as SVM, KNN, and neural networks are generally used due to their established propositions and ease of implementation. Newer algorithms, including deep neural networks and ensemble styles, have better vaccination accuracy. The development of standard datasets for all poisonous endpoints and the refinement of molecular features through better point selection algorithms are critical for future progress. Graph complications, in particular, show promise for developing further effective toxin prediction models.

In Paper [6], the authors emphasize the integration of ML algorithms into fungicide toxin prediction, a pivotal step for addressing the challenges of fungicide use in husbandry. Machine literacy provides experimenters, controllers, and stakeholders with important tools for making informed opinions about fungicide toxins. k-NN, SVM, CNN, DQA, LDA, and Random forest-grounded regression models are recommended for more accurate and effective toxin prognostications. ML models can also guide sustainable pest operation strategies, reducing reliance on dangerous fungicides and promoting safer agrarian practices.

III. PROPOSED WORK

The architecture consists of four main components: a text extraction module based on OCR, a SMILES retrieval module, a chemical classification module, and a user interface. An image of the product label is processed by the OCR module to extract text ingredient lists. After extraction, the chemical names are looked up via an external API to get their respective SMILES representations. This obviates the necessity of SMILES data to be manually collected within the dataset, where actual-time and current chemical structures are employed for classification. The SMILES representations are subsequently fed into ChemBERTa, a transformer deep learning model that is trained to examine chemical structures. ChemBERTa learns molecular properties from SMILES inputs directly, using self-attention mechanisms to identify oncogenic compounds. ChemBERTa learns the feature representations automatically, unlike conventional machine learning models that use predefined descriptors, to enhance classification accuracy.

Streamlit was used to create an easy-to-use interface so that non-technical users can interact with the system. The users can upload product label images, which are subjected to OCR to obtain the names of ingredients. These names are then translated into SMILES and subjected to oncogenic property analysis. Llama 3.2 is also incorporated as an LLM for providing explanations for classification outputs and giving insights into possible regulatory issues.

To clearly understand the end-to-end classification process, the following sequence outlines each component of the proposed system.

A. Proposed system

A. OCR-Based Ingredient Extraction

Input: Product label image
Output: Extracted text containing ingredient names
Tool: Tesseract OCR
Description: The product image is scanned to identify and extract textual ingredient information.

B. SMILES Retrieval

Input: Extracted ingredient names
Output: SMILES strings
Tool: PubChem REST API
Description: Each ingredient name is queried through PubChem to retrieve the corresponding SMILES representation dynamically.

C. Oncogenicity Classification using ChemBERTa

Input: SMILES strings
Output: Oncogenicity prediction (label: 1 = oncogenic, 0 = non-oncogenic)
Model: ChemBERTa Transformer
Description: The SMILES string is passed to the ChemBERTa model, which interprets molecular structure via contextual embeddings to classify the compound's carcinogenic potential.

D. User Interface and Output Visualization

Input: Classified prediction results
Output: On-screen visual result with model explanation
Tool: Streamlit + LLaMA 3.2

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 05, May 2025

Description: The interface displays each compound's classification and uses LLaMA to provide a natural-language explanation of why a compound may be oncogenic.

For the assessment of performance, ChemBERTa is contrasted with baseline ML classifiers such as Random Forest, Decision Trees, XGBoost, and SVM. The metrics, like accuracy, recall, precision, ROC-AUC, and F1-score, are used to identify the top-performing model. The results show that ChemBERTa considerably outperforms base classifiers in the detection of oncogenic compounds.

Combining OCR, real-time SMILES retrieval, deep learning, and a friendly UI makes the system adaptive and scalable. Dynamically retrieving chemical structures using advanced deep learning methods and enhancing usability via a UI makes the proposed framework more efficient and accurate in chemical safety assessment

B. Design

This model is built to predict whether a chemical compound could be cancer-causing by analyzing its molecular features. The design process is as follows:

A. Data Acquisition and Preprocessing

The dataset titled '**dataset.csv**' is imported into the environment through the **Pandas** library. The dataset has chemical compounds with their oncogenicity status labeled. To guarantee data integrity, basic preprocessing steps are executed, which include management of missing values, normalization of molecular feature values, and structural consistency verification. The **SMILES** representations of the compounds are fetched and inserted later in the code from **PubChem** for precise molecular structure inclusion for analysis purposes. As the dataset pertains to molecular properties and not textual data, no preprocessing based on languages is needed.

B. Feature Extraction

The relevant molecular features are extracted using the RDKit Python library for prediction, including:

A. Molecular Weight

Molecular weight is the total mass of a molecule, found by adding up the atomic weights of all the atoms it contains. It's usually measured in Daltons.

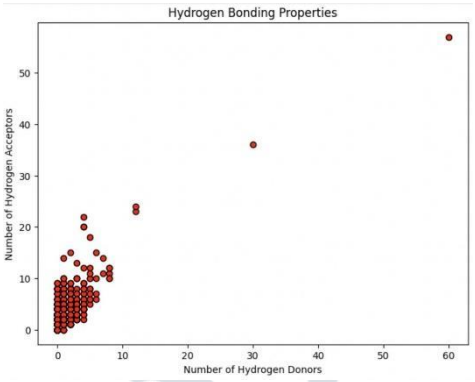

Fig. 1. Hydrogen Bonding Properties

B. Number of Hydrogen Acceptors

Hydrogen acceptors are atoms or groups in a molecule that can form hydrogen bonds by accepting hydrogen atoms.

C. Number of Hydrogen Donors

Hydrogen donors are parts of a molecule, like hydroxyl or amine groups, that can give up hydrogen atoms to form hydrogen bonds.

D. Topological Polar Surface Area (TPSA)

TPSA is the sum of the surface areas of polar atoms (usually oxygen and nitrogen) in a molecule, including their attached hydrogens.

E. MolLogP

The n-octanol-water partition coefficient shows how a chemical splits between water and oil (n-octanol). It helps show if the chemical likes water more or oily stuff more. It represents the compound's hydrophobicity, where higher values indicate greater lipophilicity (fat solubility).

These features provide significant insights into the physicochemical properties of the compounds, which are critical in determining their potential oncogenicity.

C. Target Variable Definition

The target variable, indicating the oncogenicity status of the chemical compounds, is represented by the 'label' column in the dataset. A binary encoding is used, with **'0'** representing non-oncogene and **'1'** representing oncogene.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 05, May 2025

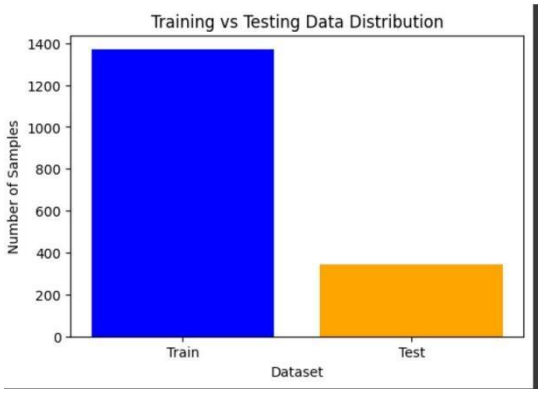

Fig. 2. Training vs Testing Data

D. Data Splitting

The train_test_split function from scikit-learn is used to break the dataset into two parts: one for training the model and the other for testing it. This split makes sure that the model is evaluated on unseen data to gauge its generalization capabilities. An 80% training and 20% testing split is typically implemented.

E. Algorithms

Several machine learning models were tested to determine the best-performing classifier for predicting oncogenicity:

A. ChemBERTa

ChemBERTa is a transformer-based deep-learning model designed for chemical informatics. It leverages self-attention mechanisms to learn molecular representations directly from SMILES strings, eliminating the need for predefined molecular descriptors. This improves accuracy in classifying the carcinogenicity of chemical compounds.

B. Random Forest Classifier

RandomForest is a type of ensemble algorithm that trains several decision tree models and merges their outcomes to make more accurate and reliable predictions. It improves accuracy and reduces overfitting by averaging the predictions of many trees.

C. Support Vector Machine (SVM)

SVM is a type of supervised learning algorithm used for tasks like classification and regression. It works by finding the best hyperplane that gives a good separation to the data points of different classes in a high-dimensional space.

D. Decision Trees

Decision trees utilize a tree-like framework to make decisions by analyzing the features of the data. Each internal node relates to a particular use, while branches show the possible outputs of those decisions.

E. XGBoost Classifier

XGBoost is a very scalable and useful implementation of gradient boosting, which is an ensemble learning technique. It constructs trees in a sequence, with each tree used to fix the errors of all the previous trees.

All models were trained on the extracted molecular features. ChemBERTa, unlike standard classifiers, utilized its transformer-based model to learn molecular representations from SMILES strings directly, without the requirement for predefined molecular descriptors. Hyperparameter tuning was used to optimize performance, and ChemBERTa proved to be better at predicting the carcinogenicity of chemical compounds.

F. Model Evaluation

All models were evaluated using the testing dataset. To make sure the classification model is reliable and performs well on new data, a rigorous evaluation process was followed using data that was strictly separated during the training and testing phases:

A. Hold-out Test Set

The dataset was divided in an 80/20 ratio, with 80% allocated for training and 20% set aside for testing. The testing data remained unseen by the model during training, hyperparameter tuning, and cross-validation.

B. Stratified Sampling

To preserve the distribution of oncogenic and non-oncogenic compounds in both training and testing sets, we applied stratified splitting using StratifiedShuffleSplit from Scikit-learn. This prevents class imbalance and ensures fair evaluation metrics across classes.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \times 100$$

$$Sensitivity\ (TPR) = \frac{TP}{TP + FN} \times 100$$

$$Specificity\ (TNR) = \frac{TN}{TN + FP} \times 100$$

Fig. 3. Evaluation Formulae

To evaluate how well our classification model identifies oncogenic compounds, we used a confusion matrix and an ROC curve.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 05, May 2025

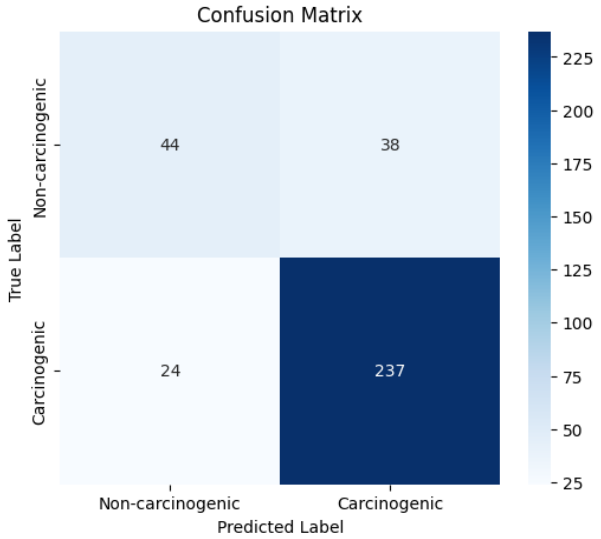

Fig. 4. Confusion Matrix

The confusion matrix (Figure 4) properly shows us the model's accuracy. It has identified 237 carcinogenic compounds (true positives) and 44 non-carcinogenic compounds (true negatives). It has also misidentified 38 non-oncogenes as oncogenic (false positives) and misclassified 24 oncogenes as non-oncogenic (false negatives).

The ROC curve shows how well the model performs at various categorization criteria. We observe a moderate capacity to differentiate between carcinogenic and non-carcinogenic compounds with an AUC of 0.7223.

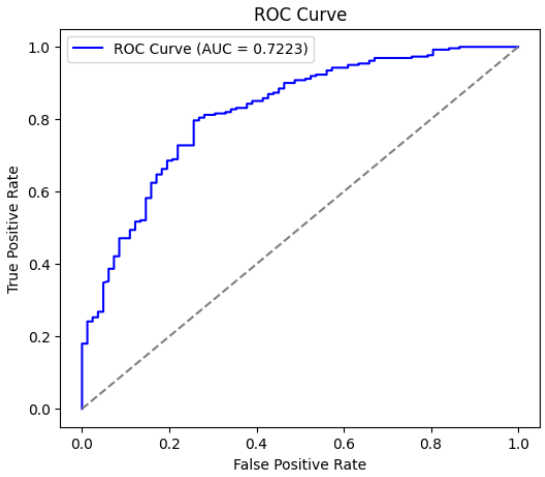

Fig. 6. ROC Curve

More importantly, these results not only show the amount of potential our proposed solution has but also the ways in which we can refine and fine-tune the model to enhance accuracy in further iterations of this project.

IV. RESULT ANALYSIS

These are the accuracy scores that we have received for each model:

Table I: Accuracy comparison for each model

MODEL	ACCURACY
ChemBERTa	0.81
XGBoost Classifier	0.80
Random Forest Classifier	0.78
SVM classifier	0.76
Decision Tree Classifier	0.73

ChemBERTa illustrated higher accuracy relative to conventional classifiers, such as Random Forest, SVM, and Decision Trees. The transformer architecture permits the model to learn intricate molecular relationships without relying on handcrafted molecular descriptors or traditional feature engineering. Instead, ChemBERTa leverages the raw SMILES structures, which it tokenizes into meaningful substructures through self-attention mechanisms and contextual embeddings. This allows the model to capture fine-grained molecular patterns and chemical dependencies that are often overlooked by conventional algorithms.

ChemBERTA had better performance over baseline machine learning models in forecasting oncogenic compounds with 81.92% accuracy, excellent precision of 86.18%, and a recall of 90.80%. The excellent F1-score of 88.43% attests to its good capacity in balancing precision and recall, demonstrating the power of transformer-based contextual embeddings in retrieving molecular feature representations important for oncogenicity.

Conversely, Decision Tree, Random Forest, XGBoost, and SVM performed poorly with accuracy levels of 72.89%, 78.13%, 80.47%, and 76.38%, respectively. These models failed to cope with the intricacies of molecular data, especially class imbalances and structural variations. To obtain these results, we ran the oncogenes.csv dataset on Google Colab, fetching SMILES representations through PubChem and molecular features through RDKit and DeepChem. The ChemBERTA model was trained on these features to generate better oncogenicity predictions.

A key advantage lies in ChemBERTa's ability to process and segment SMILES sequences similarly to how language models parse sentences, identifying structural motifs, functional groups, and stereochemical contexts within molecules. This linguistically similar comprehension of chemical representations allows the model to create a rich understanding of oncogenic characteristics, increasing its predictive ability. As a result, ChemBERTa generalizes more effectively across a wide range of chemical compounds, minimizing both false positives and false negatives. This profound contextual understanding, combined with its end-to-end learning pipeline, ultimately renders ChemBERTa a

.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 12, Issue 05, May 2025

better option for classifying the carcinogenic potential of molecules.

Fig. 7. Training vs Validation Loss

The diagram illustrates the Training vs Validation Loss over 8 epochs for the ChemBERTa model. The training loss, represented by the blue line, shows a steady and significant decline across epochs, indicating that the model is effectively learning from the training data. Starting from around 0.55, it drops sharply and plateaus below 0.30 by the 7th epoch.

In contrast, the validation loss, represented by the orange line, decreases initially but begins to level off after the 3rd epoch, stabilizing around 0.44. This trend suggests that while the model continues to improve on the training data, its generalization performance on unseen data reaches a plateau relatively early in training.

The widening gap between the training and validation loss in later epochs may suggest the early signs of overfitting, where the model becomes increasingly tailored to the training data and less effective on new inputs. However, the relatively stable validation loss also indicates that the model maintains consistent performance and does not degrade with further training. This graph properly shows the learning dynamics of the ChemBERTa model during fine-tuning and helps show its usefulness in the work of predicting carcinogenicity.

V. CONCLUSION

This research developed a ChemBERTA-based system designed to detect oncogenic compounds. It cleverly combines OCR, SMILES retrieval, and deep learning to achieve accurate classification. We improved molecular data processing using PubChem for molecular structure retrieval and RDKit & DeepChem for feature extraction. The fine-tuned ChemBERTA model proved better than conventional classifiers, with better oncogenicity prediction.

Our live Python-based user interface facilitates effective chemical safety evaluations, minimizing dependence on precomputed molecular descriptors and enhancing scalability. Upcoming improvements involve enhancing OCR accuracy, combining several SMILES databases, and tuning ChemBERTA with regulatory data and live information on newly discovered carcinogens.

.

A. Impact of OCR on Ingredient Extraction

The OCR module serves a vital purpose in the system by extracting data on ingredients from product packaging. The accuracy of OCR-based extraction, however, is image-dependent, font-variance-dependent, and text-encryption-dependent. Although the implementation at hand possesses high recognition precision, misinterpretation of abbreviations and chemical names is a problem that is yet to be solved. Improved OCR models focused on chemical nomenclature, fine-tuned to improve extraction precision, will be forthcoming.

B. SMILES Retrieval and Data Processing Challenges

The dynamic retrieval of SMILES enhances real-time classification by using the latest chemical representations. However, missing/incomplete API data can reduce precision due to failed lookups or incorrect matches. The API-based SMILES retrieval module makes it possible for the system to stay updated with current chemical information. There are challenges brought about by restrictions such as incomplete entries, discrepancies in chemical databases, and API rate limits. Overcoming such challenges calls for the use of multiple data sources and caching methods to increase the reliability of the system.

C. Usability and Interface Accessibility

The addition of Streamlit for the UI greatly enhances convenience. Users can easily add images or type ingredient names to receive real-time oncogenicity analysis. Llama 3.2 also maximizes user experience by providing elaborate explanations of model predictions. Future updates will include optimizing the UI design, enhancing the speed of response, and incorporating multilingual support to reach a global user base.

This scalable AI-based framework enhances chemical safety analysis, facilitating public health and regulation compliance through ongoing model accuracy, data merging, and accessibility improvements.

REFERENCES

[1] S. Limbu and S. Dakshanamurthy, “Predicting Chemical Carcinogens Using a Hybrid Neural Network Deep Learning Method,” *Sensors (Basel)*, vol. 22, no. 21, Oct. 2022.

[2] C. N. Cavasotto and V. Scardino, “Machine Learning Toxicity Prediction: Latest Advances by Toxicity End Point,” *ACS Omega*, vol. 7, no. 51, Dec. 2022. [Online].

[3] F. Açıkgöz, L. Verçin, and G. Erdoğan, “A Literature Review on Machine Learning in The Food Industry,” *Alphanumeric Journal*, vol. 11, no. 2, Sep. 2023. [Online].

[4] P. Carracedo-Reboredo, J. Liñares-Blanco, N. Rodríguez-Fernández, F. Cedrón, F. J. Novoa, A. Carballal, V. Maojo, and A. Pazos, "A review on machine learning approaches and trends in drug discovery," *Computational and Structural Biotechnology Journal*, Aug. 2021. [Online].

[5] L. Zhang, H. Zhang, H. Ai, and H. Hu, "Applications of Machine Learning Methods in Drug Toxicity Prediction," *Current Topics in Medicinal Chemistry*, vol. 18, no. 12, Jul. 2018. [Online].

[6] G. Anandhi and M. Iyapparaja, "Systematic Approaches to Machine Learning Models for Predicting Pesticide Toxicity," *Heliyon*, vol. 10, no. 3, Mar. 2024. [Online].

[7] L. Zhang, H. Ai, W. Chen, Z. Yin, H. Hu, J. Zhu, J. Zhao, Q. Zhao, and H. Liu, "CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods," 2023.

[8] E. Chung, D. P. Russo, H. L. Ciallella, Y.-T. Wang, M. Wu, L. M. Aleksunes, and H. Zhu, "Data-Driven Quantitative Structure-Activity Relationship Modeling for Human Carcinogenicity by Chronic Oral Exposure" 2023.

[9] Fradkin, P., Young, A., Atanackovic, L., Frey, B., Lee, L. J., & Wang, B. "A graph neural network approach for molecule carcinogenicity prediction," *Bioinformatics*, Volume 38, Issue Supplement_1, July 2022, Pages i84–i91 [Online].

[10] Y.-W. Wang, L. Huang, S.-W. Jiang, K. Li, J. Zou, and S.-Y. Yang, "CapsCarcino: A novel sparse data deep learning tool for predicting carcinogens," *Food and Chemical Toxicology*, vol. 135, p. 110921, Jan. 2020. [Online].

[11] S. Limbu and S. Dakshanamurthy, "Predicting Chemical Carcinogens Using a Hybrid Neural Network Deep Learning Method," *Sensors (Basel)*, vol. 22, no. 21, p. 8185, Oct. 2022. [Online].

[12] C. N. Cavasotto and V. Scardino, "Machine Learning Toxicity Prediction: Latest Advances by Toxicity End Point," *ACS Omega*, vol. 7, no. 51, 2022. [Online].

[13] L. Zhang, H. Zhang, H. Ai, and H. Hu, "Applications of Machine Learning Methods in Drug Toxicity Prediction," *Current Topics in Medicinal Chemistry*, vol. 18, no. 12, Jul. 2018. [Online].

[14] S. Dara, S. Dhamercherla, S.S. Jadav, C.H. M. Babu, and M.J. Ahsan, "Machine Learning in Drug Discovery: A Review," *Artificial Intelligence Review*, vol. 55, pp. 1947–1999, 2022. [Online].

[15] H. Askr, E. Elgeldawi, H. Aboul Ella, Y.A.M.M. Elshaier, M.M. Gomaa, and A.E. Hassanien, "Deep Learning in Drug Discovery: An Integrative Review and Future Challenges," *Artificial Intelligence Review*, vol. 56, pp. 5975–6037, 2023. [Online].