Journal of *Xenobiotics*

MDPI

Article

Computational Prediction of Endocrine-Disrupting Xenobiotics Associated with Hormone-Related Cancer Classification Using Molecular and Physicochemical Features

P. C. Karthik [1,*] and Divya Midhunchakkaravarthy [2]

[1] Department of Computer Science Engineering, Lincoln University College Malaysia, Wisma Lincoln, No. 12-18, Jalan SS 6/12, Petaling Jaya 47301, Malaysia
[2] Centre of Postgraduate Studies, Lincoln University College Malaysia, Wisma Lincoln, No. 12-18, Jalan SS 6/12, Petaling Jaya 47301, Malaysia
* Correspondence: pdf.karthikc@lincoln.edu.my

Abstract

Background: Chemicals associated with hormone-related cancer may interfere with endocrine signaling and are linked to biological processes involved in cancer development. Conventional toxicity assessment relies heavily on laboratory experiments and animal studies. Existing computational approaches are also limited by a lack of validated toxicity datasets and challenges in modeling complex biological mechanisms. Methods: This study presents an artificial intelligence-based framework for the binary classification of chemicals associated with hormone-related cancer. Chemical information was collected from PubChem and combined with biological information from the Comparative Toxicogenomics Database (CTD). CTD disease association records were used to construct the binary Hormone_Cancer_Label, while the disease association information was used for label generation. RDKit was used to process SMILES representations and generate molecular descriptors, while character-level TF-IDF features were extracted from the canonical SMILES. Results: The resulting Toxicity Cancer dataset contained 10,000 unique chemicals and was stratified into training (7000; 70%), validation (1500; 15%), and testing (1500; 15%) sets. Several machine learning and deep learning models, including the proposed CNN-BiLSTM architecture, were evaluated. The proposed model achieved an accuracy of 99.07%, a precision of 98.94%, a recall of 99.20%, an F1-score of 99.07%, and an ROC-AUC of 98.18%. Conclusions: The framework provides a computational classification of chemicals associated with hormone-related cancer and supports early-stage hazard screening and prioritization for toxicological investigation.

Keywords: chemical risk assessment; endocrine disruption; endocrine-disrupting chemicals (EDCs); hormone-related cancer; RDKit feature engineering

Check for updates

Academic Editor: François Gagné

Received: 8 September 2026
Revised: 12 October 2026
Accepted: 22 October 2026
Published: 11 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Chemicals associated with hormone-related cancer can intrude on the human endocrine system [1]. The interference of these chemicals has been related to many harmful social health consequences, like changed nervous, pulmonary, and lung systems, increased rates of metabolic disorders, cardiac syndromes, and tumors, and several other pathophysiological properties [2,3]. Endocrine-disrupting chemicals (EDCs) are present in everyday products as additives or contaminants, particularly in food packaging, cosmetics, cookware, toys and cleaning products [4–6]. EDCs can enter the body by ingestion, inhalation and

dermal contact [7,8]. Eco-friendly epigenetic alterations [9,10] describe molecular pathways by which varying quantities or risky elements affect hereditary conditions, resulting in the development of breast tumors. Some EDCs can be classified as poisonous, like dioxin and cadmium for breast and thyroid cancer, arsenic trioxide, asbestos, and dioxin for prostate tumor, and organochlorines for testicular tumor [11,12]. In addition to the extensive changes in EDCs and the limited ecological attention, outdated critical biochemical methods require multi-step pretreatment developments [13,14]. The toxicity of endocrine disruptors is given by the capacity that these elements have to interfere with hormonal signaling processes, mediated by nuclear receptors (NRs) [15]. Exposure to EDCs is more damaging to the fetus than to children, causing low birth weight, intrauterine growth delay, and early growth defects in infants [16,17]. EDCs in aquatic environments primarily originate from anthropogenic activities such as farming, livestock agriculture, aquaculture, businesses, clinics, and landfills [18–20]. The COVID-19 pandemic has also been a reason for the increased spread of EDC exposure [21]. The available evidence is not yet adequate to fully elucidate the involvement of EDCs in the complex molecular and cellular pathways involved in tumor progression and cancer development [22]. Microplastics are also emerging as key environmental pollutants that may have toxicological and reproductive effects. They are small in size and remain for long periods of time, enabling them to live in water and land environments and interact with living systems. Recent experimental evidence has demonstrated that exposure to polystyrene microplastics can induce molecular alterations in protamine-like proteins and affect their interaction with DNA in the mussel Mytilus galloprovincialis [23]. Still, there is a necessity for advanced technologies to eradicate EDCs more effectively [24].

Figure 1 illustrates the major genetic, environmental, biological, and lifestyle aspects that encourage cancer development and increase cancer risk. The integration of multi-omics datasets with ML techniques enables the identification of EDCs and their related genetic factors involved in clear-cell Renal Cell Carcinoma (ccRCC) [25]. Biological age-related chemicals associated with hormone-related cancer (BA-EDCs) accelerate biological aging, which has been identified, and it is suggested that increasing dietary intake of vitamins C and E may help to reduce BA-EDC-induced biological aging [26]. The key osteoporosis-related genes (FOXO3 and LUM) and chemicals associated with hormone-related cancer associated with osteoporosis are identified using network toxicology, bioinformatics, molecular docking, QRT-PCR experimental validation, and machine learning (ML) techniques [27]. An AI-based consensus Quantitative Structure–Activity Relationship (QSAR) model of estrogen receptors (ERs), pregnane X receptors (PXRs), and androgen receptors (ARs) is used for the accurate identification of EDCs and to assist environmental risk assessment [28,29]. However, there are some challenges in the computational classification of chemicals associated with hormone-related cancer. These include the need for effective representations of chemical structures, reliable classification methods, and transparent evaluation using clearly defined prediction targets. In addition, disease association databases provide valuable information for defining classification labels, but these labels must be distinguished from the features used to train predictive models [30–32]. To solve these problems, this study develops a deep learning (DL) framework that uses chemical structure-derived features to predict a binary classification label constructed from Comparative Toxicogenomics Database (CTD) disease association records. The framework is developed to support the computational screening and prioritization of chemicals for further investigation.

The proposed study uses chemical structure information obtained from PubChem and disease association records from the CTD. CTD disease association records are used exclusively to construct the binary Hormone_Cancer_Label, whereas predictive features are generated independently from the chemical structures. Specifically, canonical Simpli-

fied Molecular Input Line Entry System (SMILES) representations are transformed into 1000 character-level term frequency–inverse document frequency (TF-IDF) dimensions, which are combined with eight physicochemical and structural descriptors calculated using RDKit. When all dimensions are retained, these feature blocks form a 1008-dimensional representation for each chemical. This representation is utilized to train and evaluate ML and DL models, including the proposed CNN-BiLSTM model. The framework aims to support the computational classification and prioritization of chemicals according to their CTD-derived labels; it does not directly use chemical–gene interactions or gene–disease associations as predictive inputs.

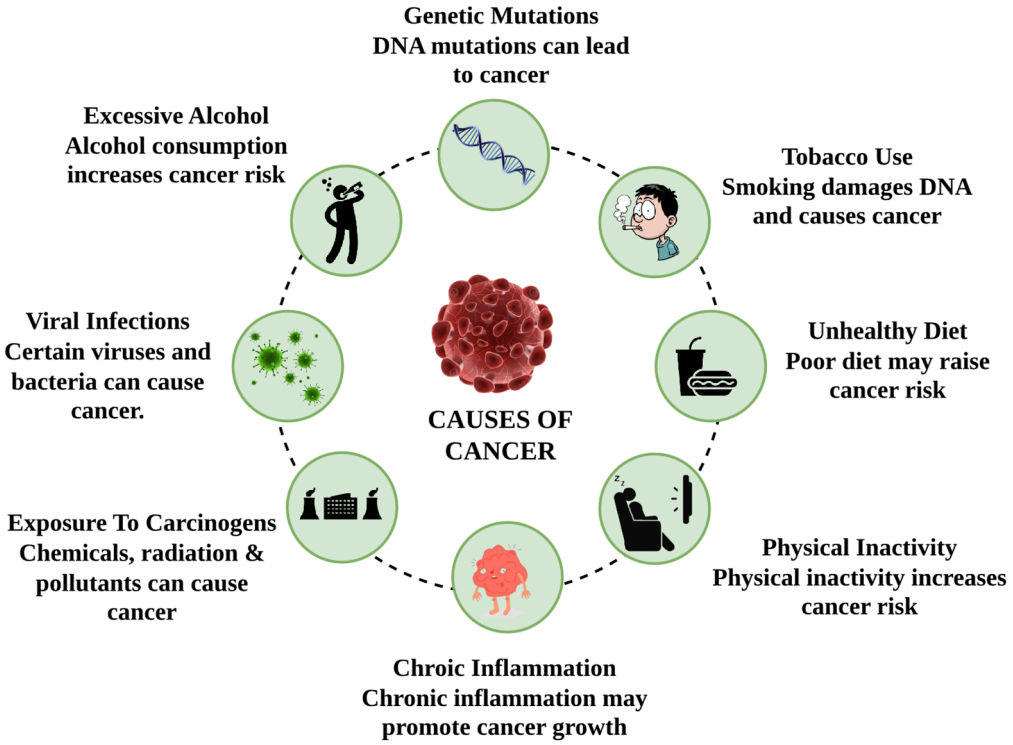


Figure 1. Illustration of common biological, environmental, and lifestyle factors related to cancer.

Research Contributions

This study aims to develop a novel AI-based framework for classifying chemicals associated with hormone-related cancer by building a new toxicity dataset. Since no publicly available dataset directly integrates chemical structures, endocrine disruption information, and carcinogenic association, this work focuses on building a Toxicity Cancer dataset by integrating multi-source chemical and biological information. The key contributions of this study are summarized as follows:

- Development of a New Integrated Toxicity Dataset: A dataset of chemicals associated with hormone-related cancer was constructed using chemical information from PubChem and CTD disease association records. The CTD records were used to define the binary Hormone_Cancer_Label, while the predictive representation was generated independently from chemical structures using canonical SMILES representations and RDKit-derived molecular descriptors. This separation was adopted to avoid direct target leakage.
- Chemical Structure-Based Feature Engineering: Canonical SMILES representations were converted into 1000 character-level TF-IDF dimensions and combined with eight RDKit-derived molecular descriptors to form the predictive representation, subject to retention of all dimensions during preprocessing.

- ML Model Development and Evaluation: We trained and evaluated several ML models, including XGBoost, Random Forest and ANN models, using the newly created toxicity dataset to assess their effectiveness at classifying chemicals associated with hormone-related cancer.
- Early-Stage Risk Identification of Chemicals Associated with Hormone-Related Cancer: The proposed framework enables early screening and prioritization for toxicological investigation.

The remainder of the manuscript is structured as follows: Section 2 reviews the related work on hormone-related cancer and chemicals associated with hormone-related cancer. Section 3 describes the proposed AI-based framework for classifying chemicals associated with hormone-related cancer. Experimental results are reported in Section 4, and Section 5 concludes the paper.

2. Related Work

This section reviews the ML models and molecular descriptors used to predict the toxicity and endocrine-disrupting effects of chemicals. Integration of chemical–gene interactions and disease association data has improved the identification of compounds associated with hormone-related cancer. Limbu et al. [33] developed an artificial intelligence-hybrid neural network (AI-HNN) and a pathophysiology-based chemo-phenotypic toxicity measurement (CPTM) to predict chemical-mixture toxicity and carcinogenicity. This approach not only enhanced the accuracy of the classification of hormone-related cancer but also improved the understanding of the molecular mechanisms associated with chemical toxicity. However, the model did not fully capture the toxicokinetic, toxicodynamic, and exposomic aspects of chemical mixtures. Limbu and Dakshanamurthy [34] proposed a novel framework (HNN-MixCancer) based on a hybrid neural network (HNN) for the evaluation of the carcinogenic risk of hazardous chemical mixtures. Carcinogenic chemical mixtures were preferred by the model even when using assumption-based mixtures. However, the model needed the molecular properties and biological interaction information necessary for accurate early-stage chemical risk prediction. Chen et al. [35] constructed a DL framework based on capsule networks and an attention mechanism (DCAMCP) to predict molecular carcinogenicity. The model improved cancer risk evaluation by capturing complex molecular features. However, the model required improved molecular feature representation and optimization techniques, and it mainly focused on structural information without considering chemical, gene and disease-related associations. Ollitrault et al. [36] focused on Qualitative Gene expression Activity Relationship (QGexAR) models to predict the likelihood of chemical-induced disruption. ML techniques, including the CatBoost, XGBoost, Random Forest, SVM, Logistic Regression, AutoKeras, and TPOT methods, were employed to predict disruptions in estrogen signaling, androgen receptor, and thyroid hormone pathways. The model identified important features that can help identify chemicals and develop safer chemical substances. Zhang et al. [37] proposed the CarcGC model, employing a CNN to predict the carcinogenicity of hazardous air pollutants. The model learned chemical structural patterns to enhance carcinogenicity prediction. However, the model lacked interpretability and ignored biological interaction data to explain the relation between chemical exposure and cancer generation. Hao et al. [38] introduced a forward multi-toxicity regression model of chemicals associated with hormone-related cancer based on the algorithm of XGBoost. Multiple toxicity endpoints were forecast by the model. But the model was complicated by the interactions of features, which made it less interpretable, and by not considering combined molecular fingerprints, chemical–gene interactions and gene–disease associations for full risk assessment. The existing AI-based models for the classification of chemicals associated with hormone-related cancer face challenges in iden-

tifying EDCs associated with hormone-related cancer because of the complex interplay between chemical structures, biological mechanisms, and genetic factors. Conventional toxicity assessment techniques are costly and time-consuming. Although several ML and DL methods have been developed, many existing approaches rely primarily on chemical features, lack integrated biological information, and are limited by the availability of well-structured toxicity datasets. Existing studies show the potential of AI-based approaches in the prediction of chemical hazards, which have differences in their objectives, datasets and prediction targets. Some methods are directed toward molecular toxic endpoints and others examine receptor activity or carcinogenicity-related outcomes. Most methods focus mainly on chemical structure features and lack integration of complementary biological association information. This study combines the molecular features extracted from the SMILES with curated biological association information for the computational classification of chemicals associated with hormone-related cancers.

3. Proposed Methodology

In this study, an AI-based framework is introduced for the classification of chemicals associated with hormone-related cancer using chemical structure-derived features.

Chemical information was obtained from PubChem, while CTD disease association records were used only to construct the binary Hormone_Cancer_Label. CTD-derived disease association information was not utilized as a predictive model feature. The chemical structures represented by canonical SMILES representations were processed to generate character-level TF-IDF features, and eight physicochemical descriptors were calculated using RDKit. These features formed the final 1008-dimensional input representation used by the ML and DL models. The created dataset was split into training, validation, and test splits, with the resultant features preprocessed using training-set-based procedures. The extracted features were utilized as inputs to the ML and DL models, including the XGBoost [39], Random Forest [40], and ANN [41] models and the proposed CNN-BiLSTM model. The developed dataset was utilized to train and test these models to learn the correlation between chemical features and their associated biological outcomes. The overall framework of the proposed model is visualized in Figure 2.

3.1. Dataset Creation

The Toxicity Cancer dataset was built by integrating chemical information from PubChem [42] and disease association information from the CTD [43]. The source dataset contains 19,064 distinct disease types. For label construction, the eligible disease categories were identified according to the predefined inclusion criteria. The main chemical identifiers were the PubChem Compound IDs (CIDs). The canonical SMILES representations were an additional level of structural identifier. Duplicate chemical records were removed using PubChem CIDs and the canonical SMILES before dataset construction. When multiple CTD disease association records were available for the same chemical, all records were consolidated at the chemical level before assigning the target label. Thus, each unique chemical compound contributed only one sample to the final dataset. The target variable, denoted as the Hormone_Cancer_Label, was generated from CTD disease association records. Class 1 (Hormone Cancer) represents chemicals with at least one CTD disease association record corresponding to the hormone-related cancer categories. Class 0 (Other Diseases) represents chemicals associated with diseases outside these hormone-related cancer categories. This is reliable given the lack of CTD association with the hormone-related cancer categories under the labeling criteria applied in this study. The dataset label construction process is presented in Table 1.

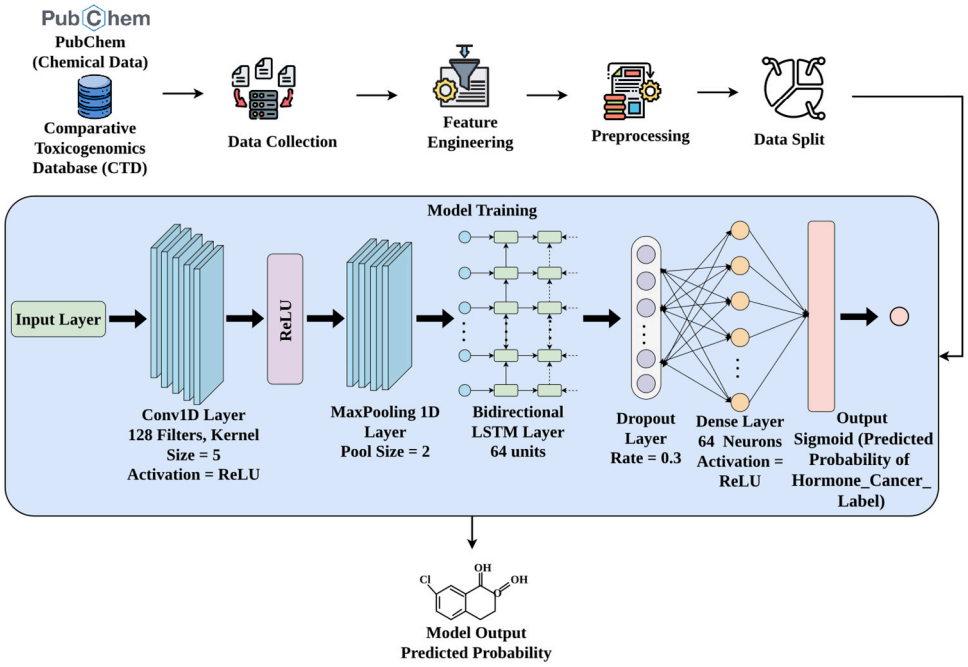


Figure 2. The architecture of the proposed framework. Chemical information is obtained from PubChem, and CTD disease association records are used only for construction of the Hormone_Cancer_Label. Canonical SMILES representations are processed to generate 1000 character-level TF-IDF features and eight RDKit molecular descriptors, which form the 1008-dimensional predictive input for model training.

Table 1. Definition and construction of Hormone_Cancer_Label.

Component	Description
Data source	CTD disease–PubChem association records
Data retrieval date	[July 2026]
Number of disease types	109,064
Prediction target	Hormone_Cancer_Label
Positive class (Class 1)	Chemical with at least one CTD disease association record matching a hormone-related cancer category
Negative class (Class 0)	Chemical with CTD disease association records outside the hormone-related cancer categories; termed "Other Diseases"
Label assignment rule	Class 1 was assigned if at least one associated disease matched the hormone-related cancer categories; otherwise, Class 0 was assigned.
Duplicate handling	Duplicate chemicals were removed using PubChem Compound IDs and canonical SMILES
Predictive CTD disease information	CTD disease identifiers, disease categories, hormone-related cancer annotations, and variables derived from these records are omitted from the predictive feature matrix.
Final dataset	10,000 chemicals: 5000, Class 1; 5000, Class 0

The combination of PubChem and CTD information was done at the chemical-identifier level. The chemical structure and molecular information were provided by PubChem, while disease association records were provided by the CTD to generate the binary target label, as represented in Figure 3. The CTD disease identifiers, disease categories, annotations of hormone-related cancer and any variables derived from these disease association records were not included in the predictive feature matrix. Only canonical-SMILES-based character-

level TF-IDF features and RDKit molecular descriptors were used to generate the model inputs from the target. The TF-IDF vectorizer was only fitted to the training SMILES strings. The resulting vocabulary and IDF weights were then fixed and used on the testing and validation SMILES strings without refitting. Therefore, disease association information was not directly used to obtain a target class from the prediction model. The initial chemical feature representation included 1000 character-level TF-IDF features derived from canonical SMILES representations with an n-gram range of 2–5 together with eight molecular descriptors calculated using RDKit: molecular weight (MolWt), LogP, topological polar surface area (TPSA), number of hydrogen bond donors (NumHDonors), number of hydrogen bond acceptors (NumHAcceptors), number of rotatable bonds (RotatableBonds), ring count (RingCount), and heavy atom count (HeavyAtomCount). The features generated an initial 1008-dimensional representation.

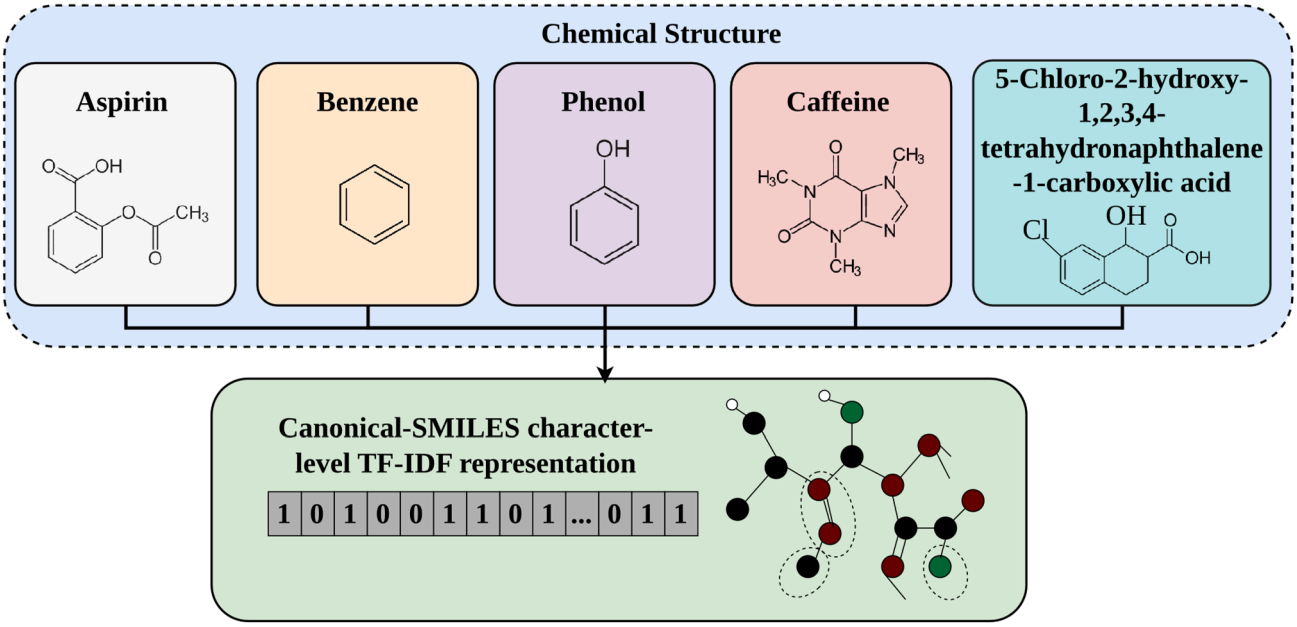


Figure 3. Molecular structure processing and predictive feature generation using canonical-SMILES character-level TF-IDF features and RDKit molecular descriptors. The 1000 TF-IDF features are combined with eight molecular descriptors to form a 1008-dimensional representation for model development. The dashes lines in the Canonical-SMILES character-level TF-IDF representation denotes the molecular features, while the different colors distinguish the atoms and components.

3.2. Data Preprocessing

The preprocessing procedures were applied to the final 1008-dimensional feature representation. Missing-value imputation was fitted on the training set, and the learned parameters were used on the validation and testing sets. Missing values in numerical features were imputed through the median value calculated from the training partition. The same training-derived median values were subsequently used for the validation and testing partitions. In the same way, StandardScaler [44] was trained on the training data only and used on the validation and test data without retraining. Mutual-information-based feature selection was also fitted only on the training set, and the selected features were subsequently used on the validation and testing sets. This procedure confirms that data contained in the validation and testing sets does not affect the model development. The initial chemical feature representation consists of 1000 character-level TF-IDF dimensions of canonical SMILES representations and eight RDKit-derived molecular descriptors, resulting in an input dimensionality of 1008. Missing numerical values were given the median value

that was computed only from the training partition. The TF-IDF vectorizer was also trained using only the training SMILES strings, and the vocabulary and inverse document frequency (IDF) weights were used unchanged for the validation and test sets. The final CNN-BiLSTM input retained all 1008 dimensions.

For model training, the training and validation datasets were used. The validation dataset was utilized for checking learning behavior, hyperparameter tuning and applying early stopping, while the independent testing dataset could not be accessed until the evaluation step. For the additional chemical-level separation of dataset partitions, overlap analysis was conducted among the training, validation, and testing datasets. PubChem CID codes and canonical SMILES representations were compared across dataset partitions. No common PubChem CIDs and canonical SMILES representations were found in the training, validation and testing subsets. Therefore, the evaluation datasets had truly independent chemical compounds. Lastly, the dataset was split into training, validation, and testing sets in a 70%/15%/15% stratified class distribution. The class counts were determined before balancing using the down-sampling method. To construct the final balanced dataset, 5000 chemicals were selected from each class. The majority class was randomly down-sampled to 5000 chemicals, while the minority class was retained if it contained exactly 5000 eligible chemicals. Therefore, the final dataset comprised 10,000 chemicals, with 5000 samples in each class. Therefore, there were 7000 training instances (3500 per class), 1500 validation instances (750 per class) and 1500 testing instances (750 per class). These are defined in Table 2. The 70%/15%/15% split was performed after the removal of duplicates and assignment of labels, with equal class distributions in the training, validation, and testing datasets.

Table 2. Distribution of hormone cancer and Other Diseases compounds across training, validation, and testing datasets.

Dataset Partition	Other Diseases (0)	Hormone Cancer (1)	Total Samples
Training Set (70%)	3500	3500	7000
Validation Set (15%)	750	750	1500
Testing Set (15%)	750	750	1500
Total	5000	5000	10,000

3.3. Feature Engineering and Dataset Preparation

Each chemical compound was represented using chemical structure-derived features. Character-level TF-IDF encoding of canonical SMILES representations generated a 1000-dimensional vector, which was concatenated with eight RDKit [45] molecular descriptors: MolWt, LogP, TPSA, NumHDonors, NumHAcceptors, RotatableBonds, RingCount, and HeavyAtomCount. The resulting representation had 1008 dimensions per chemical compound. These dimensions represent positions in the engineered feature vector. This separation prevents the predictive model from using label disease information directly in the training. For N chemical compounds, the -th chemical compound feature vector is represented in Equation (1):

$$x_i = [x_{i1}, x_{i2}, \ldots, x_{im}] \quad (1)$$

where x_i signifies the feature vector of the i^{th} chemical compound, m signifies the extracted features and x_{ij} represents the value of the j^{th} molecular descriptor and fingerprint feature. The whole set of data may be represented as in Equation (2):

$$D = \{(x_i, y_i)\}_{i-1}^{N} \quad (2)$$

where D represents the final dataset, and y_i denotes the binary target label. The target variable represents the CTD-derived hormone-related cancer disease association label. A chemical is assigned to Class 1 if at least one of its CTD disease association records matches a predefined hormone-related cancer category; otherwise, it is assigned to Class 0 under the labeling criteria. The definition of the label is demonstrated in Equation (3):

$$y_i = \begin{cases} 1, CTD - derived\ hormone - related\ cancer\ disease \\ 0, otherwise \end{cases} \quad (3)$$

To eliminate feature-scale differences, StandardScaler normalization was applied. Each feature was transformed using the standardization formula. This transformation produces standardized features with approximately zero mean and unit variance based on the training distribution.

3.4. Model Development and Training

The CNN-BiLSTM model received a 1008-dimensional feature vector for each chemical compound, comprising 1000 canonical-SMILES character-level TF-IDF dimensions followed by eight RDKit molecular descriptors. The vector was reshaped to a tensor of a 1008 × 1 size and fed to the one-dimensional CNN. The CNN was then applied to the ordered positions of features using convolutional filters, followed by a BiLSTM layer. In this architecture, the BiLSTM models the dependencies in the ordered engineered feature representation, but the positions are not a temporal sequence of chemical events. The learned representation was passed into the dense and sigmoid output layers to predict the binary classification label derived from the CTD. The -th chemical compound feature vector is represented in Equation (4):

$$X_i = [t_{i1}, t_{i2}, \ldots, t_{i1000}, d_{i1}, d_{i2}, \ldots, d_{i8}] \quad (4)$$

where t_{ij} signifies the -th TF-IDF feature and d_{ik} represents the -th RDKit descriptor. Thus, each chemical compound is represented by 1008 features. After the training-derived preprocessing and standardization, the feature vector is reshaped into a 1008 × 1 depiction before it is sent to the CNN-BiLSTM network, as denoted in Figure 4. Therefore, for a batch of B chemical compounds, the input tensor has the shape B × 1008 × 1. Here, the dimension 1008 represents the ordered feature positions, while the final dimension of one represents the single feature channel. The eight RDKit descriptors are appended directly to the TF-IDF representation before entering the CNN.

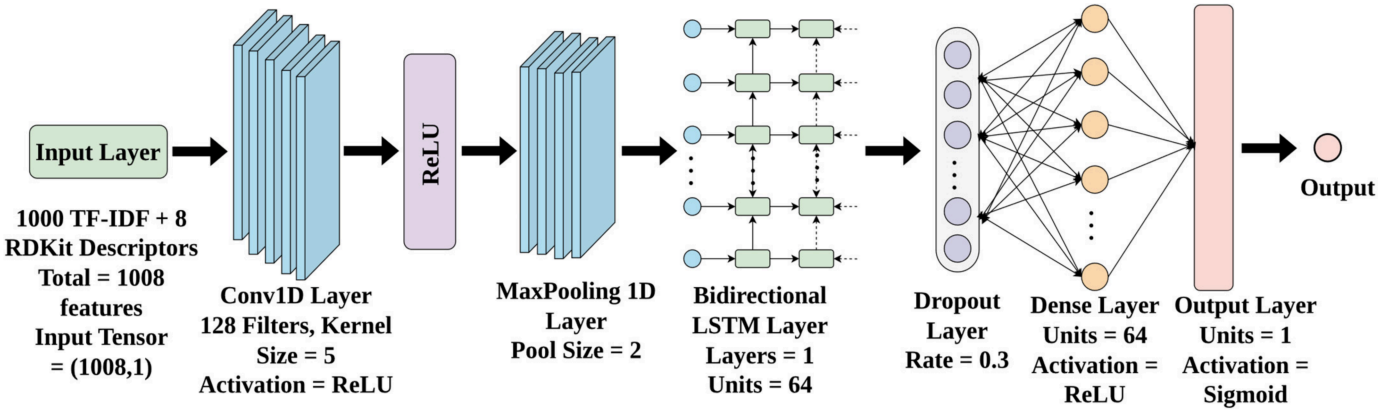


Figure 4. The architecture of the proposed CNN-BiLSTM model. The 1000-dimensional character-level TF-IDF representation extracted from canonical SMILES representations is concatenated with eight RDKit-derived molecular descriptors to form a 1008-dimensional input vector for each chemical compound. The vector is reshaped into a 1008 × 1 tensor, with 1008 representing the input length

and 1 representing the single input channel. The CNN extracts local patterns across the input dimensions, while the BiLSTM captures forward and backward contextual dependencies. The learned representation is passed through dense and sigmoid output layers to perform binary classification.

The CNN part applies a one-dimensional convolution operation to the ordered feature representation to learn local patterns and combinations between neighboring feature positions. The convolution layer consists of 64 filters with a kernel size of 3 and ReLU activation. The extracted feature representation is passed into one BiLSTM layer with 64 units. The BiLSTM progresses the learned representation in the forward and backward directions to focus dependencies across the feature-index sequence. Thus, the BiLSTM is used to model relationships among the learned molecular feature representations rather than temporal dependencies. The CNN and BiLSTM components provide complementary feature learning. The CNN is trained to learn local patterns from the engineered feature representation, while the BiLSTM is trained to learn more global dependencies from the resulting representation. The learned representation is subsequently fed using a dense layer followed through a sigmoid output layer for the binary classification of chemicals associated with hormone-related cancer. The prediction function is represented as in Equation (5):

$$\hat{y} = f_{\mathrm{CNN-BiLSTM}}(X) \tag{5}$$

where X represents the 1008-dimensional chemical feature representation and $f_{\mathrm{CNN-BiLSTM}}$ denotes the trained CNN-BiLSTM model.

The output probability of a chemical fitting to the hormone-related cancer-associated class is calculated through the sigmoid activation function, as denoted in Equation (6):

$$P(y = 1 \mid X) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

where z represents the output of the final dense layer. The sigmoid output represents the probability of the positive CTD-derived Hormone_Cancer_Label class. The model parameters are optimized through the binary cross-entropy (BCE) loss function, as denoted in Equation (7):

$$L = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\,] \tag{7}$$

where N signifies the training samples, y_i denotes the actual class label, and $\hat{y}_i$ signifies the predicted probability for sample i.

The CNN-BiLSTM model was tuned using the validation set by optimizing the convolutional filters, kernel size, BiLSTM units, dropout rate, learning rate, and training epochs. The experiments were performed on Python 3.10.13. The major software packages and their versions were NumPy 1.26.4, Pandas 2.2.2, scikit-learn 1.5.1, TensorFlow 2.17.0, Keras 3.5.0 and RDKit 2024.03.5. Datasets were partitioned, and the modeling procedure was always performed with a random seed of 42. In addition, the random seeds of NumPy and TensorFlow were both set to 42 to minimize variance resulting from stochastic processes that occur during neural network training. The same preprocessing configuration and fixed feature generation procedure were used throughout the corresponding training, validation, and testing experiments.

4. Results and Discussion

This study develops an AI-based framework for the early estimation of chemicals associated with hormone-related cancer before human exposure. The dataset was generated from chemical and biological interaction data and processed using RDKit and StandardScaler. Specific hyperparameter settings are represented in Table 3. The model performance is evaluated through standard metrics.

Table 3. Details of parameter settings.

Category	Parameter	Value
Dataset	Total samples	10,000
Target	Target variable	Hormone_Cancer_Label
SMILES Encoding	Method	TF-IDF
	Analyzer	Character-level
	N-gram range	(2, 5)
Feature Representation	RDKit descriptor count	8
	TF-IDF dimensionality	1000
	Initial input dimensionality	1008
Preprocessing	Scaling method	StandardScaler
Model	Algorithm	XGBoost Classifier
	n_estimators	800
	max_depth	10
	learning_rate	0.03
	subsample	0.9
	colsample_bytree	0.9
	objective	binary: logistic
	eval_metric	logloss
	random_state	42
CNN-BiLSTM	CNN conv layers	1
	Number of filters	64
	Kernel sizes	3
	BiLSTM layers	1
	BiLSTM units	64
	Dropout rate	0.30
	Activation functions	CNN: ReLU
	Activation functions	BiLSTM: tanh & sigmoid
	Dense layer	1
Training	Optimizer	Adam
	Learning rate	0.001
	Batch size	64
	Early-stopping criteria	val_loss, patience = 5
	Epochs	50
	Random seed	42
Output	Classification type	Binary classification
	Class 0	Other Diseases
	Class 1	Hormone-related cancer

4.1. Performance Analysis of Proposed Model

The proposed CNN-BiLSTM framework achieved a good performance for classifying chemicals associated with hormone-related cancer, with an accuracy of 99.07%. It also obtained 98.94% precision, 99.20% recall and a 99.07% F1-score, which means good classification with few false positives. Moreover, the 99.07% specificity indicates a good performance in correctly classifying the chemicals not related to hormone-related cancer. The MCC value of 98% is high, which further confirms the good agreement between predicted and actual classifications. The low FPR of 0.93% and FNR of 1.07% mean that the false-positive and false-negative classifications were constrained. These show that the proposed framework is effective for the early-stage chemical hazard screening and prioritization of chemicals related to hormone-associated cancer.

Figure 5 displays the confusion matrix for the classification of chemicals related to hormone-related cancer into hormone-related cancer-associated and other disease-associated categories. The model shows a high classification performance with very few misclassifications between the two classes. This proves reliable discrimination between the classes and supports the potential use of the framework for early-stage chemical hazard screening and prioritization. The proposed framework is useful for the prioritization of chemicals for further toxicological evaluation.

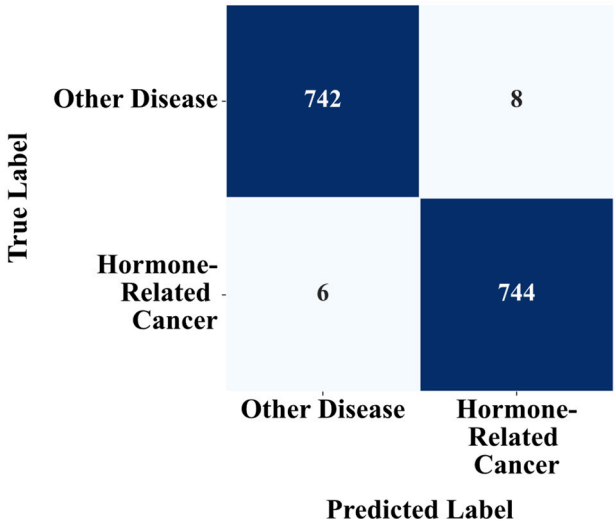


Figure 5. The confusion matrix of the proposed CNN-BiLSTM model evaluated on the testing dataset for the classification of chemicals associated with hormone-related cancer. The color intensity represents the number of samples in each class, with darker blue indicates correct classifications and lighter blue indicates misclassifications.

Figure 6 shows the PR curve of the proposed CNN-BiLSTM for the classification of chemicals associated with hormone-related cancer. For a wide range of recall values, this curve is close to the top-right corner. This suggests good accuracy and a good classification performance. The high average precision also indicates that the framework is capable of finding a good trade-off. The PR curve demonstrates the proposed framework's performance for the early-stage screening of chemical hazards and prioritization of chemicals associated with hormone-related cancer.

Figure 7 shows the comparison of the precision–recall (PR) curves of the proposed CNN-BiLSTM model with those of the ANN, XGBoost, CNN and BiLSTM models. The proposed CNN-BiLSTM model maintains higher precision across recall values, showing a superior classification performance compared with the existing models. The increase in the average precision indicates that the trade-off is effective and the classification of

chemicals related to hormone-associated cancer is robust. These findings prove the value of the proposed framework in early-stage chemical hazard screening and prioritization.

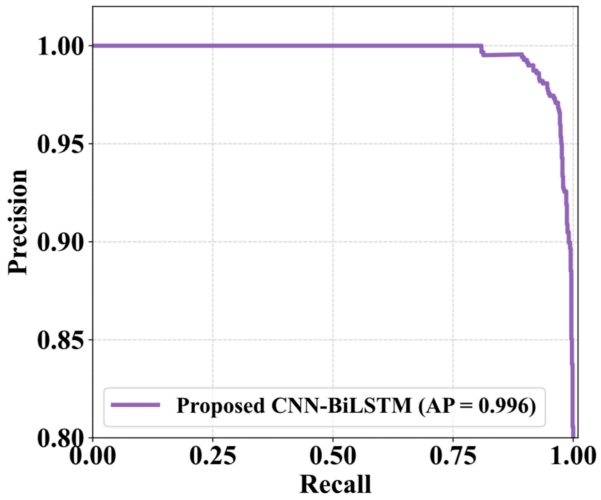


Figure 6. Precision–recall (PR) curve of proposed CNN-BiLSTM model for chemicals associated with hormone-related cancer.

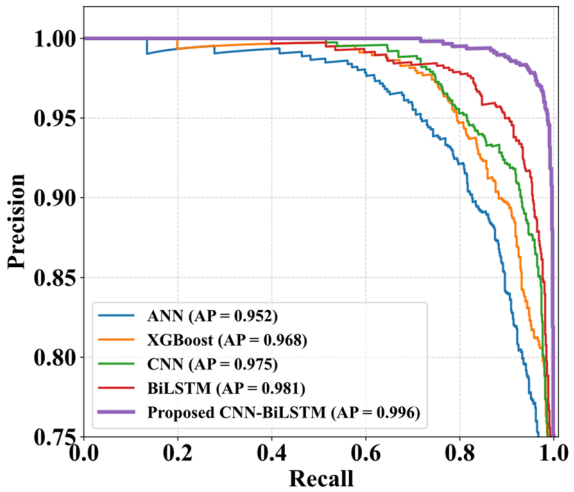


Figure 7. Precision–recall (PR) curve comparison of proposed CNN-BiLSTM model with existing ML and DL models for classifying chemicals associated with hormone-related cancer.

Figure 8 shows the Receiver Operating Characteristic (ROC) curve of the proposed CNN-BiLSTM. The ROC curve is close to the upper-left corner. This indicates that the two chemical classes are well discriminated. The high AUC value means that the model can reach a high true-positive rate with a low false-positive rate. These results help demonstrate the utility of the proposed framework for chemical hazard screening and prioritization at the early stage.

Figure 9 represents the learning behavior of the proposed CNN-BiLSTM with the training and validation datasets. The validation curves were monitored during model development to evaluate convergence and detect possible overfitting. Only the test set was used for the final model architecture and parameters. Figure 9a displays that the training and validation accuracy gradually increases across epochs. The curves converge to a high value, showing effective model learning. Figure 9b shows that the loss values of the training and validation tend to decrease, indicating that the optimization is stable and the model is converging. The training and validation curves are in close agreement, indicating little

overfitting in the model training. This demonstrates that the proposed framework can be used to classify stable chemicals related to hormone-related cancer.

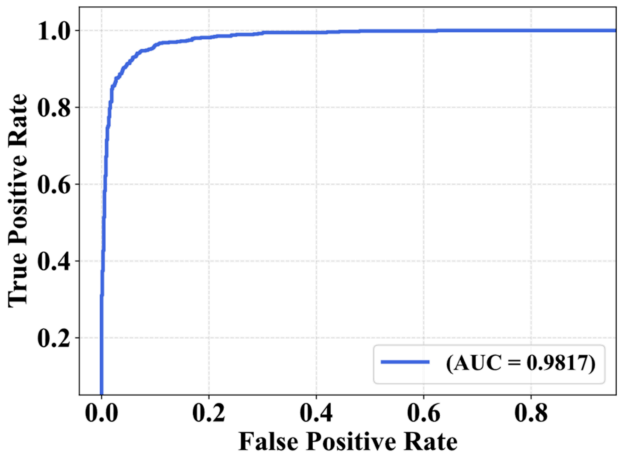
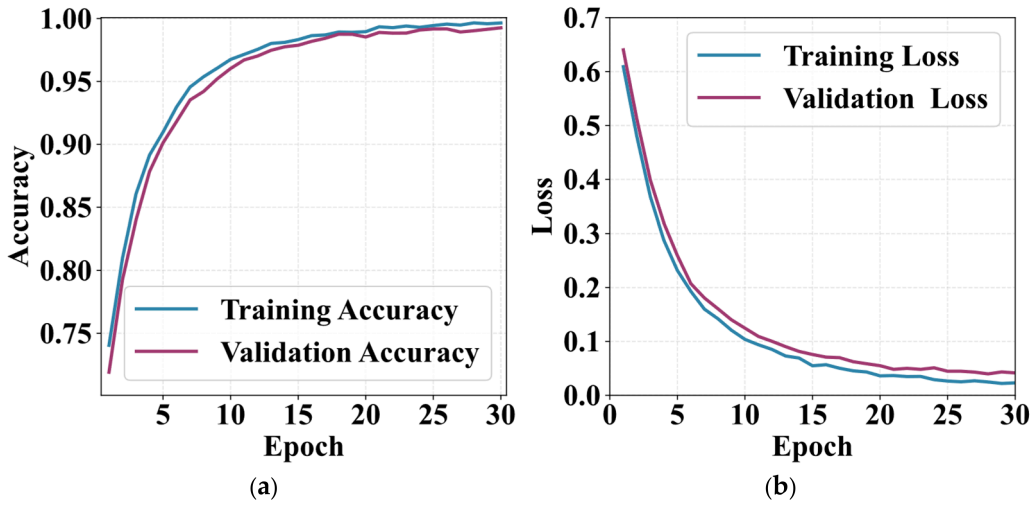


Figure 8. ROC curve of proposed CNN-BiLSTM model for classifying chemicals associated with hormone-related cancer.



(**a**) (**b**)

Figure 9. Training and validation performances of proposed CNN-BiLSTM: (**a**) training and validation accuracies versus epochs, and (**b**) training and validation losses versus epochs.

Figure 10 demonstrates the inference times of diverse models for varying batch sizes. The inference time increases for all methods as the batch size increases because of the higher workload. The proposed CNN-BiLSTM has less inference time compared to other models, which shows better computational efficiency and scalability. SVM has the maximum inference time for all batch sizes, whereas the Random Forest, AdaBoost, CatBoost, and GCN models are in the middle. Therefore, the proposed framework is more appropriate for real-time and large-scale prediction applications.

Table 4 shows the training and validation accuracies and loss curves of the CNN-BiLSTM model at different training epochs. The table illustrates the learning behavior and convergence during training. The training and validation accuracies increase steadily while the loss values decrease gradually with the increasing number of epochs, which indicates successful optimization and stable convergence. The training and validation metrics are in good agreement, suggesting that the model training did not overfit much. This proves that the framework can classify the chemicals related to hormone-related cancer with a stable learning performance.

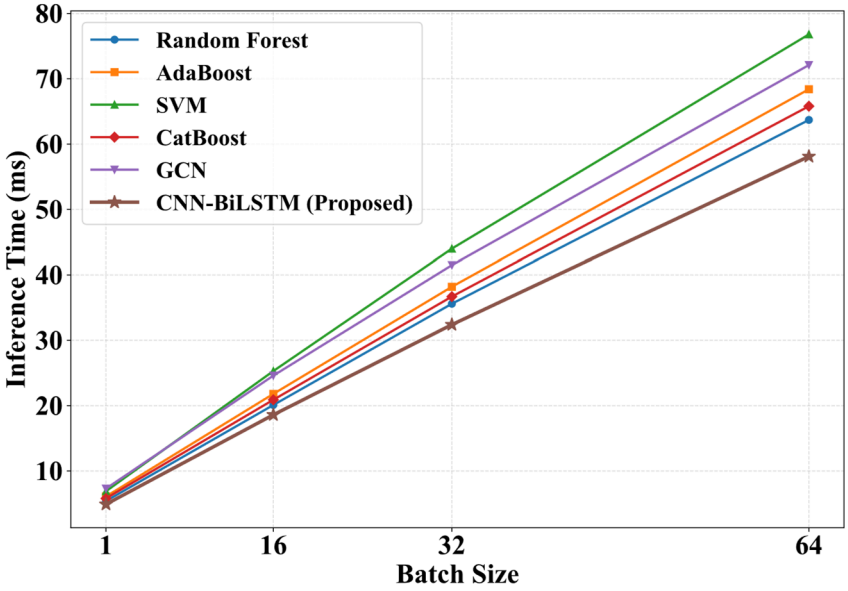


Figure 10. Inference time comparison of diverse ML and DL models across varying batch sizes.

Table 4. Training and validation accuracies and losses of proposed CNN-BiLSTM model across different epochs. Results were obtained with training and validation datasets during model optimization.

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss	Validation Loss
1	86.42	84.71	0.412	0.446
5	93.88	93.21	0.192	0.204
10	96.54	96.12	0.104	0.109
15	97.83	97.51	0.071	0.074
20	98.51	98.24	0.052	0.056
25	98.96	98.71	0.039	0.043
30	99.18	99.00	0.031	0.035

4.2. Ablation Study

Figure 11 shows the ablation study for the contributions of different input features and network architectures to the proposed framework.

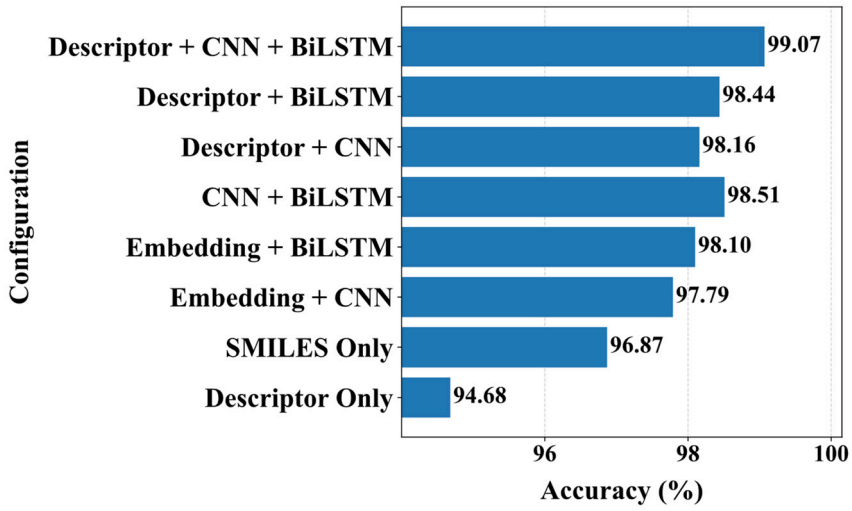


Figure 11. Ablation study of different feature representations and model configurations.

Individual Feature Representations: Models using only molecular descriptors or SMILES representations provide the baseline prediction performance. Molecular descriptors describe physicochemical properties, and SMILES representations keep molecular structural information. However, neither representation alone can learn complex molecular patterns, resulting in lower prediction accuracy.

Hybrid Feature Learning with Deep Networks: The CNN or BiLSTM + molecular descriptors achieve a significantly better prediction performance. The CNN can capture the local spatial patterns of features, and BiLSTM is good at modeling the sequential dependencies among molecular features. The hybrid configurations learn more discriminative feature embeddings and outperform individual feature representations.

Descriptor + CNN + BiLSTM Framework: The proposed model fuses the RDKit molecular descriptors with CNN and BiLSTM spatial and sequential learning. This complementary architecture captures the local and long-range molecular feature relationships and produces the most informative feature representation. Consequently, it obtained the highest classification accuracy among all the configurations tested for the detection of chemicals associated with hormone-related cancer.

4.3. Comparative Analysis

Table 5 shows the comparative performances of different ML and DL models trained and assessed on the proposed toxicity dataset for classifying chemicals associated with hormone-related cancer. To make a fair comparison, all baseline models were implemented, trained, and evaluated with the equivalent preprocessing pipeline, feature extraction process, and experimental settings on the proposed dataset. The results prove that the proposed CNN-BiLSTM performs better than baseline techniques such as the AdaBoost, SVM, CatBoost, and GCN methods. The proposed CNN-BiLSTM model achieved the highest accuracy score of 99.07%, while its precision score was 98.94% and its recall score was 99.20%, showing that the proposed CNN-BiLSTM is an accurate and balanced classifier for classifying chemicals related to hormone-associated cancer.

Table 5. Comparative analysis of proposed model and existing models.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)	Specificity (%)	MCC (%)	Inference Time (ms)
AdaBoost	95.82	95.6	95.41	95.5	96.87	96.15	91.54	5.8
SVM	96.14	96.01	95.84	95.92	97.12	96.48	92.36	6.6
GCN	97.84	97.79	97.7	97.74	98.05	98.08	96.41	7
Proposed CNN-BiLSTM	99.07	98.94	99.20	99.07	98.18	99.07	98	4.7

Table 6 illustrates the performances of the baseline models trained and tested on the proposed integrated dataset. All the models were tested with the same preprocessing pipeline and experimental settings for a fair comparison. The proposed CNN-BiLSTM attained the maximum accuracy (99.07%), precision (98.94%), recall (99.20%), F1-score (99.07%), ROC-AUC (98.18%) and specificity (99.07%) compared to all the other models. The superior accuracy, recall, F1-score, ROC-AUC and specificity consistently confirm that the developed CNN-BiLSTM model exhibits an advanced classification performance, well-balanced predictive results and strong discriminative ability.

Table 6. Baseline models trained on proposed Toxicity Cancer dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)	Specificity (%)
Random Forest	95.68	95.51	95.43	95.47	96.72	95.88
XGBoost	97.94	97.88	97.82	97.85	97.83	98.01
ANN	97.31	97.24	97.18	97.21	98.04	97.52
CNN	98.18	98.13	98.06	98.09	98.02	98.26
CNN-BiLSTM (Proposed)	99.07	98.94	99.20	99.07	98.18	99.07

Table 7 shows the comparison of the proposed CNN-BiLSTM and the existing models. Literature-based comparison was executed with similarity of the task and relevance to the method. The existing literature uses various datasets, endpoints and evaluation protocols. These comparisons are proposed to provide contextual positioning. The proposed CNN-BiLSTM achieved the best overall performance, with an accuracy of 99.07%, a precision of 98.94%, a recall of 99.20%, an F1-score of 99.07% and an ROC-AUC of 98.18%. CarcGC had the best performance among the existing methods, with a 98.23% accuracy and 98.07% ROC-AUC, followed by HNN-MixCancer, Multi-Toxicity XGBoost and DCAMCP. These results confirm that the proposed spatial feature extraction and sequential learning based on the CNN and BiLSTM improve the classification performance and discrimination ability over the existing methods.

Table 7. Contextual performance comparison of proposed CNN-BiLSTM with existing state-of-the-art methods.

Method	Architecture	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
DCAMCP	Capsule Network + Attention	97.86	97.81	97.74	97.77	98.74
HNN-MixCancer	Hybrid Neural Network	98.05	97.99	97.93	97.96	98.91
CarcGC	CNN	98.23	98.17	98.11	98.14	98.07
Multi-Toxicity XGBoost	Ensemble Learning	97.91	97.85	97.79	97.82	97.81
Proposed CNN-BiLSTM	CNN + BiLSTM	99.07	98.94	99.20	99.07	98.18

4.4. K-Fold Cross-Validation for Robustness Assessment

To assess the robustness of the proposed CNN-BiLSTM model, five-fold scaffold-based cross-validation was performed using Bemis–Murcko molecular scaffolds. The Bemis–Murcko scaffold of each chemical compound was obtained from its molecular structure. Compounds with the same Bemis–Murcko scaffolds are allocated to the same fold rather than being split into different folds. The scaffold groups were then split into five folds so that the validation fold was composed of scaffolds that were not part of the respective training folds. For the scaffold-based cross-validation, Bemis–Murcko scaffolds were assigned before the partitioning and preprocessing operations. These were independently fitted within each training fold before application to the corresponding validation fold. In each fold, the training and validation partitions contained chemically distinct scaffolds, and

the number of structurally similar compounds across folds was minimized. The complete pipeline was executed independently in each of the training folds during scaffold-based cross-validation. This fold-wise implementation prevents information transfer between training and validation samples and offers an unbiased estimation of model robustness. The model was trained in each fold, and the evaluation metrics were computed on the corresponding validation fold, as demonstrated in Table 8. The low standard deviations indicate a consistent performance across the five scaffold-separated validation folds. This scaffold-based cross-validation represents an internal generalization assessment within the proposed dataset.

Table 8. Five-fold scaffold-based cross-validation performance of proposed CNN-BiLSTM model.

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Fold 1	98.82	98.79	98.84	98.81	99.55
Fold 2	98.94	98.91	98.95	98.93	99.60
Fold 3	99.01	98.98	99.03	99.00	99.63
Fold 4	98.88	98.86	98.89	98.87	99.53
Fold 5	98.96	98.93	98.89	98.95	99.61
Mean ± SD	98.92 ± 0.07	98.89 ± 0.07	98.92 ± 0.07	98.91 ± 0.07	99.58 ± 0.04

In this section, all baseline and proposed models used the same 70/15/15 training, validation, and testing partitions and the same 1008-dimensional chemical feature representation consisting of 1000 character-level TF-IDF features derived from canonical SMILES representations and eight RDKit molecular descriptors. The preprocessing operations, including missing-value imputation and feature scaling, were fitted through the training partition and subsequently used for the validation and testing partitions. Baseline model hyperparameters were selected using the validation partition, while the test partition was not used for model selection or tuning.

Table 9 shows that the baseline ANN, CNN and BiLSTM models were implemented and estimated through the same dataset partition and leakage-free preprocessing procedure as the proposed CNN-BiLSTM model. The principal training configurations for all models are shown in Table 9. The models were trained through the specified Adam optimizer, learning rate, batch size and maximum training epochs, with early stopping applied to the CNN, BiLSTM, and proposed CNN-BiLSTM models. For the proposed CNN-BiLSTM model, the best-performing weights were restored after early stopping.

Table 9. Hyperparameter configurations of baseline and proposed models evaluated on proposed Toxicity Cancer dataset.

Model	Optimizer	Learning Rate	Batch Size	Maximum Epochs	Early Stopping
ANN (MLP)	Adam	0.001	64	50	-
CNN	Adam	0.001	64	50	Patience = 5
BiLSTM	Adam	0.001	64	50	Patience = 5
Proposed CNN-BiLSTM	Adam	0.001	64	50	Patience = 5, restore best weights

4.5. External Dataset Evaluation Using DEDuCT Dataset

To estimate the generalization capability of the proposed framework, an external validation experiment was performed with the Database of Endocrine Disrupting Chemicals

and their Toxicity Profiles (DEDUCT) dataset [46]. The DEDuCT dataset was treated as a completely independent external dataset and was not used during the construction of the proposed Toxicity Cancer dataset, model training, validation, testing, or hyperparameter optimization. Therefore, no information from the DEDuCT was available during model development. The chemical structures in the DEDuCT dataset were processed using the same pipeline established for the proposed framework.

The external validation performance of the proposed CNN-BiLSTM model was compared with those of the baseline Random Forest, XGBoost, CNN and BiLSTM methods. The results are illustrated in Table 10. For external validation, the TF-IDF vocabulary and IDF weights learned from the original training partition were retained and applied directly to the DEDuCT SMILES representations without refitting. The CNN-BiLSTM model attained an accuracy of 91.84%, a precision of 91.56%, a recall of 92.27% and an F1-score of 91.91% on the DEDuCT dataset. These results provide an assessment of the applicability of the learned structure-based representation to an independent chemical collection. However, the DEDuCT endpoint is related to endocrine-disrupting chemicals and is not necessarily identical to the CTD-derived hormone-related cancer association label used to construct the proposed Toxicity Cancer dataset. Therefore, the DEDuCT results are interpreted as an external dataset evaluation.

Table 10. External validation performance on DEDuCT dataset.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	87.42	86.95	88.17	87.55
XGBoost	89.76	89.31	90.08	89.69
CNN	90.18	89.84	90.76	90.30
BiLSTM	90.63	90.27	91.14	90.70
Proposed CNN-BiLSTM	91.84	91.56	92.27	91.91

4.6. Discussion

In this study, a CNN-BiLSTM-based model was developed for the prediction of chemicals associated with hormone-related cancer (EDCs) by integrating molecular descriptors, molecular fingerprints, chemical–gene interactions and gene–disease associations. The proposed framework provides a rapid approach for large-scale chemical screening. The addition of chemical and biological information permits the model to learn complex relationships associated with endocrine disruption and hormone-related carcinogenicity. The proposed CNN-BiLSTM achieved an accuracy of 99.07%, a precision of 98.94%, a recall of 99.20%, an F1-score of 99.07%, a specificity of 99.07% and an ROC-AUC of 98.18%, which signifies high-quality prediction. The high precision means that false-positive predictions are minimized. The model predicts the binary Hormone_Cancer_Label and represents the computational classification of chemical association with hormone-related cancer. The confusion matrix, precision–recall curve, and ROC curve validated the robustness of the proposed framework. This shows accurate classification with minimal misclassification and excellent discrimination between hormone-related cancer-associated and other disease-associated chemicals. The training and validation curves showed stable convergence with high accuracy and low loss. This specifies effective learning, minimal overfitting and effective learning and minimal overfitting within the evaluated Toxicity Cancer dataset. The proposed CNN-BiLSTM framework achieved a good classification performance for chemicals associated with hormone-related cancer.

Some limitations are acknowledged. The model was developed using a Toxicity Cancer dataset of 10,000 samples collected from PubChem and the CTD, which does not fully capture the structural and biological diversity of endocrine disruptors. The reported model performance was obtained using a 1008-dimensional input representation consisting of 1000 SMILES-based TF-IDF features and eight RDKit molecular descriptors; therefore, the current representation may not capture all relevant structural, physicochemical, and biological features of the chemicals. The proposed model does not incorporate an interpretability method. Descriptors from RDKit give an interpretable representation of the physical and chemical properties, and TF-IDF features represent the structural patterns from the SMILES strings. Moreover, at present, the proposed framework mainly focuses on molecular descriptors and biological interactions and does not consider other complementary data, such as gene expressions, protein interactions, or other multi-omics data. In addition, the model is a binary classifier and is not able to give mechanistic insights into the molecular pathways that underlie hormone-related cancer associations. The proposed framework is further extended by explainable artificial intelligence (XAI) methods to boost predictive accuracy. This enhances the interpretability and makes it more suitable for the large-scale chemical hazard screening and prioritization of chemicals associated with hormone-related cancer.

5. Conclusions

This study proposes a CNN-BiLSTM-based artificial intelligence framework for the computational classification of chemicals associated with hormone-related cancer. Since a suitable integrated dataset was not directly available, a Toxicity Cancer dataset was developed by integrating chemical information from PubChem with disease association information from the CTD. The CTD disease association records were used to construct the binary Hormone_Cancer_Label, while the disease identifiers, categories, and annotations used for label generation were excluded from the predictive feature matrix to minimize information leakage. The predictive representation was generated exclusively from canonical SMILES representations using 1000 character-level TF-IDF features and eight RDKit-derived molecular descriptors. The proposed CNN-BiLSTM model showed an accuracy of 99.07%, a precision of 98.94%, a recall of 99.20%, an F1-score of 99.07%, and an ROC-AUC of 98.18% on the evaluated test set. The model can be used as a tool for the early-stage computational and prioritization of chemical substances for additional research.

Author Contributions: Both authors contributed to the conception of the problem setting and overall design of the work. P.C.K. built the conceptualization and methodology. D.M. implemented the work and writing. Both authors were involved in the visualization of concepts. This version was approved by both authors, who also read and approved the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study are openly available at github at https://github.com/Project-1414-Cec/project_dt_124312.git. The URLs were accessed on 8 August 2026.

Acknowledgments: The Code availability in https://github.com/Project-1414-Cec/project_834101.git. The URLs were accessed on 8 August 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Macedo, S.; Teixeira, E.; Gaspar, T.B.; Boaventura, P.; Soares, M.A.; Miranda-Alves, L.; Soares, P. Endocrine-disrupting chemicals and endocrine neoplasia: A forty-year systematic review. *Environ. Res.* **2023**, *218*, 114869. [CrossRef] [PubMed]
2. Aghayev, Z.; Szafran, A.T.; Tran, A.; Ganesh, H.S.; Stossi, F.; Zhou, L.; Mancini, M.A.; Pistikopoulos, E.N.; Beykal, B. Machine learning methods for endocrine disrupting potential identification based on single-cell data. *Chem. Eng. Sci.* **2023**, *281*, 119086. [CrossRef] [PubMed]
3. Virtuoso, S.; Raggi, C.; Maugliani, A.; Baldi, F.; Gentili, D.; Narciso, L. Toxicological effects of naturally occurring endocrine disruptors on various human health targets: A rapid review. *Toxics* **2024**, *12*, 256. [CrossRef] [PubMed]
4. Marchiandi, J.; Alghamdi, W.; Dagnino, S.; Green, M.P.; Clarke, B.O. Exposure to endocrine disrupting chemicals from beverage packaging materials and risk assessment for consumers. *J. Hazard. Mater.* **2024**, *465*, 133314. [CrossRef] [PubMed]
5. Özel, F.; Rüegg, J. Exposure to endocrine-disrupting chemicals and implications for neurodevelopment. *Dev. Med. Child Neurol.* **2023**, *65*, 1005–1011. [CrossRef] [PubMed]
6. Jermin Jeaunita, T.C.; Ramesh, T.; Kalpana, A.V.; Shanthala, P.T. Multi-Task Deep Learning Framework for Anomaly Detection and Carbon Emission Prediction in Precision Agriculture. *IEEE Trans. Sustain. Comput.* **2025**, *10*, 1337–1350. [CrossRef]
7. Cai, Z.; Li, Z.; Wang, Q.; Wang, Z.; Wu, Q.; Wang, C. Synthesis of cyano and ionic dual-functional hypercrosslinked porous polymer for effective adsorption and detection of endocrine disrupting chemicals in milk matrix. *J. Hazard. Mater.* **2024**, *462*, 132746. [CrossRef] [PubMed]
8. Wan, M.L.Y.; Co, V.A.; El-Nezami, H. Endocrine disrupting chemicals and breast cancer: A systematic review of epidemiological studies. *Crit. Rev. Food Sci. Nutr.* **2022**, *62*, 6549–6576. [CrossRef] [PubMed]
9. Singh, D.D. Epigenetic mechanisms of endocrine-disrupting chemicals in breast cancer and their impact on dietary intake. *J. Xenobiotics* **2024**, *15*, 1. [CrossRef] [PubMed]
10. Lin, Y.; Zhang, Y.; She, J.; Zhao, R.; Lin, S.; Zhang, Y.; Zhang, L.; Wei, J.; Lin, Y.; Yang, Q. Novel insights into the causal relationship between endocrine-disrupting chemicals and breast cancer mediated by circulating metabolites. *Environ. Pollut.* **2025**, *375*, 126349. [CrossRef] [PubMed]
11. Modica, R.; Benevento, E.; Colao, A. Endocrine-disrupting chemicals (EDCs) and cancer: New perspectives on an old relationship. *J. Endocrinol. Investig.* **2023**, *46*, 667–677. [CrossRef] [PubMed]
12. Chauhan, S.S.; Garg, P.; Parthasarathi, R. Computational framework for identifying and evaluating mutagenic and xenoestrogenic potential of food additives. *J. Hazard. Mater.* **2024**, *470*, 134233. [CrossRef] [PubMed]
13. Yao, L.; Fu, Z.; Duan, Q.; Wu, M.; Song, F.; Wang, H.; Qin, Y.; Bai, Y.; Zhou, C.; Quan, X.; et al. An intelligent spectral identification approach for the simultaneous detection of endocrine-disrupting chemicals in aquatic environments. *Environ. Res.* **2025**, *264*, 120368. [CrossRef] [PubMed]
14. Vandenberg, L.N.; Williams, M.; Maharaj, A.; Matkowski, S.M. Testing and evaluations for endocrine disrupting chemicals: Identification, hazard and risk assessment. *Curr. Opin. Endocr. Metab. Res.* **2026**, *43*, 100599. [CrossRef]
15. Interdonato, L.; Siracusa, R.; Fusco, R.; Cuzzocrea, S.; Di Paola, R. Endocrine disruptor compounds in environment: Focus on women's reproductive health and endometriosis. *Int. J. Mol. Sci.* **2023**, *24*, 5682. [CrossRef] [PubMed]
16. Mitsuhashi, R.; Sato, K.; Kawakami, H. Novel Epigenetics Control (EpC) Nanocarrier for Cancer Therapy Through Dual-Targeting Approach to DNA Methyltransferase and Ten-Eleven Translocation Enzymes. *Epigenomes* **2025**, *9*, 6. [CrossRef] [PubMed]
17. Ding, T.; Yan, W.; Zhou, T.; Shen, W.; Wang, T.; Li, M.; Zhou, S.; Wu, M.; Dai, J.; Huang, K.; et al. Endocrine disrupting chemicals impact on ovarian aging: Evidence from epidemiological and experimental evidence. *Environ. Pollut.* **2022**, *305*, 119269. [CrossRef] [PubMed]
18. Zhang, H.; Ouyang, W.; He, K.; Wang, L.; Pei, J.; Lin, C.; Zhang, S.; Li, D.; He, M.; Liu, X. Developing water quality and land use surrogates to predict endocrine-disrupting chemical profiles in a highly urbanized river basin. *Environ. Pollut.* **2024**, *362*, 124951. [CrossRef] [PubMed]
19. Ismanto, A.; Hadibarata, T.; Kristanti, R.A.; Maslukah, L.; Safinatunnajah, N.; Kusumastuti, W. Endocrine disrupting chemicals (EDCs) in environmental matrices: Occurrence, fate, health impact, physio-chemical and bioremediation technology. *Environ. Pollut.* **2022**, *302*, 119061. [CrossRef] [PubMed]
20. Chen, Y.; Yang, J.; Yao, B.; Zhi, D.; Luo, L.; Zhou, Y. Endocrine disrupting chemicals in the environment: Environmental sources, biological effects, remediation techniques, and perspective. *Environ. Pollut.* **2022**, *310*, 119918. [CrossRef] [PubMed]
21. Di Pietro, G.; Forcucci, F.; Chiarelli, F. Endocrine disruptor chemicals and children's health. *Int. J. Mol. Sci.* **2023**, *24*, 2671. [CrossRef] [PubMed]

22. Maksimova, V.; Popova, V.; Kholodova, A.; Makus, J.; Usalka, O.; Lylova, E.; Kudriashov, A.; Belitsky, G.; Yakubovskaya, M.; Kirsanov, K. Epigenetic Activity of Cancer Therapy Drugs Revealed by HeLa TI Cell-Based Assay. *Epigenomes* **2026**, *10*, 14. [CrossRef] [PubMed]
23. Marinaro, C.; Scarciello, G.; Bianchi, A.R.; Berman, B.; Chianese, T.; Scudiero, R.; Rosati, L.; De Maio, A.; Lettieri, G.; Piscopo, M. Toxicological effects and potential reproductive risk of microplastic-induced molecular changes in protamine-like proteins and their DNA binding. *Chem.-Biol. Interact.* **2025**, *405*, 111309. [CrossRef] [PubMed]
24. Shah, S.; Oh, J.; Bang, Y.; Jung, S.; Kim, H.C.; Jeong, K.S.; Park, M.H.; Lee, K.A.; Ryoo, J.H.; Kim, Y.J. Pregnant women's lifestyles and exposure to endocrine-disrupting chemicals: A machine learning approach. *Environ. Pollut.* **2025**, *366*, 125309. [CrossRef] [PubMed]
25. Zheng, M.; Wang, D.; Guan, B.; Cao, L. Integrating multi-omics data and machine learning to identify endocrine disrupting chemicals targeting key ccRCC-related genes. *Discov. Oncol.* **2025**, *17*, 197. [CrossRef] [PubMed]
26. Huang, W.; Zhang, Z.; Colucci, M.; Deng, L.; Yang, M.; Huang, X.; Zhou, X.; Jin, Y.; Lazzarini, E.; Balbi, C.; et al. The mixed effect of Endocrine-Disrupting chemicals on biological age Acceleration: Unveiling the mechanism and potential intervention target. *Environ. Int.* **2024**, *184*, 108447. [CrossRef] [PubMed]
27. Huang, X.Z.; Huang, H.; Chen, H.; Wei, Y.K. Identification of endocrine-disrupting chemicals targeting key OP-associated genes via bioinformatics and machine learning. *Ecotoxicol. Environ. Saf.* **2024**, *286*, 117155. [CrossRef] [PubMed]
28. Zhang, R.; Wang, B.; Li, L.; Li, S.; Guo, H.; Zhang, P.; Hua, Y.; Cui, X.; Li, Y.; Mu, Y.; et al. Modeling and insights into the structural characteristics of endocrine-disrupting chemicals. *Ecotoxicol. Environ. Saf.* **2023**, *263*, 115251. [CrossRef] [PubMed]
29. Iwobi, N.; Sparks, N.R. Endocrine disruptor-induced bone damage due to hormone dysregulation: A review. *Int. J. Mol. Sci.* **2023**, *24*, 8263. [CrossRef] [PubMed]
30. Pan, J.; Liu, P.; Yu, X.; Zhang, Z.; Liu, J. The adverse role of endocrine disrupting chemicals in the reproductive system. *Front. Endocrinol.* **2024**, *14*, 1324993. [CrossRef] [PubMed]
31. Chaturvedi, M.; Joy, S.; Gupta, R.D.; Pandey, S.; Sharma, S. Endocrine disrupting chemicals (EDCs): Chemical fate, distribution, analytical methods and promising remediation strategies—A critical review. *Environ. Technol. Rev.* **2023**, *12*, 286–315. [CrossRef]
32. Cai, W.; Jiang, B.; Yin, Y.; Ma, L.; Li, T.; Chen, J. Identification of STAT3 phosphorylation inhibitors using generative deep learning, virtual screening, molecular dynamics simulations, and biological evaluation for non-small cell lung cancer therapy. *Mol. Divers.* **2025**, *29*, 3189–3205. [CrossRef] [PubMed]
33. Limbu, S.; Glasgow, E.; Block, T.; Dakshanamurthy, S. A machine-learning-driven pathophysiology-based new approach method for the dose-dependent assessment of hazardous chemical mixtures and experimental validations. *Toxics* **2024**, *12*, 481. [CrossRef] [PubMed]
34. Limbu, S.; Dakshanamurthy, S. Predicting dose-dependent carcinogenicity of chemical mixtures using a novel hybrid neural network framework and mathematical approach. *Toxics* **2023**, *11*, 605. [CrossRef] [PubMed]
35. Chen, Z.; Zhang, L.; Sun, J.; Meng, R.; Yin, S.; Zhao, Q. DCAMCP: A deep learning model based on capsule network and attention mechanism for molecular carcinogenicity prediction. *J. Cell. Mol. Med.* **2023**, *27*, 3117–3126. [CrossRef] [PubMed]
36. Ollitrault, G.; Marzo, M.; Roncaglioni, A.; Benfenati, E.; Mombelli, E.; Taboureau, O. Prediction of endocrine-disrupting chemicals related to estrogen, androgen, and thyroid hormone (EAT) modalities using transcriptomics data and machine learning. *Toxics* **2024**, *12*, 541. [CrossRef] [PubMed]
37. Zhang, H.; Yi, H.; Hao, Y.; Zhao, L.; Pan, W.; Xue, Q.; Liu, X.; Fu, J.; Zhang, A. Deciphering exogenous chemical carcinogenicity through interpretable deep learning: A novel approach for evaluating atmospheric pollutant hazards. *J. Hazard. Mater.* **2024**, *465*, 133092. [CrossRef] [PubMed]
38. Hao, N.; Zhao, Y.; Sun, P.; Deng, Z.; Cui, X.; Liu, J.; Zhao, W. Endocrine disruptor identification and multitoxicity level assessment of organic chemicals: An example of multiple machine learning models. *J. Hazard. Mater.* **2025**, *485*, 136896. [CrossRef] [PubMed]
39. Nachouki, M.; Mohamed, E.A.; Mehdi, R.; Abou Naaj, M. Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends Neurosci. Educ.* **2023**, *33*, 100214. [CrossRef] [PubMed]
40. Hakkal, S.; Ait Lahcen, A. XGBoost to enhance learner performance prediction. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100254. [CrossRef]
41. Shah, A.; Shah, M.; Pandya, A.; Sushra, R.; Sushra, R.; Mehta, M.; Patel, K.; Patel, K. A comprehensive study on skin cancer detection using artificial neural network (ANN) and convolutional neural network (CNN). *Clin. eHealth* **2023**, *6*, 76–84. [CrossRef]
42. PubChem. Available online: https://pubchem.ncbi.nlm.nih.gov/ (accessed on 6 June 2026).
43. Comparative Toxicogenomics Database. Available online: https://ctdbase.org/ (accessed on 6 June 2026).
44. Kumar, M.; Bhardwaj, V. Evaluating label encoding and preprocessing techniques for breast cancer prediction using machine learning algorithms. *Int. J. Comput. Intell. Syst.* **2025**, *18*, 218. [CrossRef]

45. Kunnakkattu, I.R.; Choudhary, P.; Pravda, L.; Nadzirin, N.; Smart, O.S.; Yuan, Q.; Anyango, S.; Nair, S.; Varadi, M.; Velankar, S. PDBe CCDUtils: An RDKit-based toolkit for handling and analysing small molecules in the Protein Data Bank. *J. Cheminform.* **2023**, *15*, 117. [CrossRef] [PubMed]
46. DEDuCT. Available online: https://cb.imsc.res.in/deduct/download (accessed on 6 June 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.