International Journal of
Molecular Sciences

MDPI

Article

Prediction of Drug-Induced Liver Toxicity Using SVM and Optimal Descriptor Sets

Keerthana Jaganathan [1], Hilal Tayara [2,*] and Kil To Chong [1,3,*]

[1] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju 54896, Korea; keerthanairtt@gmail.com
[2] School of International Engineering and Science, Jeonbuk National University, Jeonju 54896, Korea
[3] Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju 54896, Korea
* Correspondence: hilaltayara@jbnu.ac.kr (H.T.); kitchong@jbnu.ac.kr (K.T.C.)

Abstract: Drug-induced liver toxicity is one of the significant safety challenges for the patient's health and the pharmaceutical industry. It causes termination of drug candidates in clinical trials and also the retractions of approved drugs from the market. Thus, it is essential to identify hepatotoxic compounds in the initial stages of drug development process. The purpose of this study is to construct quantitative structure activity relationship models using machine learning algorithms and systematical feature selection methods for molecular descriptor sets. The models were built from a large and diverse set of 1253 drug compounds and were validated internally with 10-fold cross-validation. In this study, we applied a variety of feature selection techniques to extract the optimal subset of descriptors as modeling features to improve the prediction performance. Experimental results suggested that the support vector machine-based classifier had achieved a better classification accuracy with reduced molecular descriptors. The final optimal model provides an accuracy of 0.811, a sensitivity of 0.840, a specificity of 0.783 and Mathew's correlation coefficient of 0.623 with an internal validation set. Furthermore, this model outperformed the prior studies while evaluated in both the internal and external test sets. The utilization of distinct optimal molecular descriptors as modeling features produce an in silico model with a superior performance.

Keywords: drug-induced liver toxicity; feature selection; support vector machine; prediction; molecular descriptors

Citation: Jaganathan, K.; Tayara, H.; Chong, K.T. Prediction of Drug-Induced Liver Toxicity Using SVM and Optimal Descriptor Sets. *Int. J. Mol. Sci.* **2021**, *22*, 8073. https://doi.org/10.3390/ijms22158073

Academic Editor: Bono Lucic

Received: 29 June 2021
Accepted: 23 July 2021
Published: 28 July 2021

1. Introduction

The liver is an indispensable organ of the body due to its crucial contribution in metabolizing xenobiotics [1]. Drug-induced liver toxicity is one of the primary reasons for drug failure in clinical cases and also leads to termination of approved drugs from the market. Most commonly, drugs, herbals and other dietary products are responsible for the uncertain adverse liver injury [2–5]. The idiosyncratic behavior of the drugs not only caused by the dose level prescribed but also depends on the patient's metabolic, genetic and immunological factors [6]. Due to the unpredictable adverse hepatic effects on patient's health, drug-induced liver injury (DILI) risk assessment has become the most important concern for safe drug development [7–10]. Hence, it is required to concentrate more on identifying the potential hepatotoxic compounds in advance.

Animal studies for predicting DILI concerns in the preclinical assessment are not reliable, as it provides a low correlation results in clinical trials and also in post-marketing treatment [11,12]. In vitro and in vivo experiments for detecting DILI of large number of substances are time-consuming and expensive. Additionally, most of the compounds induce peculiar toxicity effects in human liver which cannot be discovered by the regulatory system for new drugs [13–16]. To address the limitations of experimental approaches, predictive computational modeling is taken into consideration for evaluating the DILI risk of drug candidates. Moreover, computational studies are reasonably cheaper, allows rapid

prediction in virtual screening of huge compounds and evade ethical challenges linked to animal methods [17,18].

In recent years, computational predictive modeling approaches have been recognized as an alternative by many research groups. Despite various data types, for example, chemical structure and gene expression data, more number of computational models using the molecular structure of the compounds have been reported [19–22]. In particular, in silico studies are beneficial for filtering out molecular structures causing hepatotoxicity in the early stages of drug discovery. However, expert-based models using structural alerts are not successful predictors as they are defined according to experts' experience and knowledge about the drug toxicity mechanisms [23–25]. So, various machine learning algorithms based on statistical Quantitative Structure Activity Relationship (QSAR) models have been developed by using the molecular structure features with the known hepatotoxicity endpoint datasets. Ekins et al. developed a Bayesian model using extended connectivity molecular fingerprints and interpretable descriptors based on a training set composed of 295 compounds and a test set of 237 compounds. This Bayesian model had a prediction accuracy of about 60% in external validation data [26]. Zhang et al. presented a naive Bayes classifier, which yielded an accuracy of 72.6% for the external test set [27]. Although many machine learning-based statistical models have been reported with sufficiently high accuracy, these models suffered from either imbalanced or small datasets with unsatisfactory prediction performances [28–30]. Mulliner et al. published Support Vector Machine (SVM) models combined with a genetic algorithm trained on a large dataset of 3712 compounds related to human and animal liver toxicity data [31]. Ai et al. reported an ensemble learning model using molecular fingerprints based on 1241 diverse compounds [32]. Recently, He et al. built a large and chemically diverse balanced training set of 1254 unique compounds as a result of system literature retrieval and constructed an ensemble model by integrating eight base classifiers to enhance prediction performance using molecular descriptors given by Marvin [33]. They achieved an average accuracy (ACC) of 0.783, sensitivity (SEN) of 0.818 and specificity (SPE) of 0.748 within a 10-fold cross-validation. Altogether, the prediction performance of the proposed models are not satisfactory, and there is substantial room for enhancing drug-induced liver toxicity predictions.

In this present study, we propose a drug-induced liver toxicity prediction model using an SVM classifier with optimal subset of numerically represented molecular structure features. We worked on a variety of machine learning methods and feature selection techniques to improve the liver toxicity prediction performance using molecular descriptors. We computed different molecular descriptor sets from compounds' Simplified Molecular Input Entry System (SMILES) format using various open software such as PaDEL, Chemopy, CDK and RDKit. We employed feature reduction techniques to remove redundant and irrelevant features from high dimensional molecular descriptor sets. Then, we applied feature selection techniques F-score algorithm for feature ranking followed by SVM linear kernel-based Recursive Feature Elimination with Cross-Validation (RFECV) method to select the optimal subset of features. Initially, we analyzed the performance of the SVM classifier with the optimal features of each individual molecular descriptor sets. Next, we investigated the prediction performance with different combinations of individual descriptor sets. Finally, the combination of all descriptor sets was used to build binary machine learning classifiers. The SVM-based classifier with reduced molecular descriptors showed improved prediction performance within 10-fold cross-validation and external validation set compared to the recently reported prior study. The overall workflow of the proposed model is shown in Figure 1.

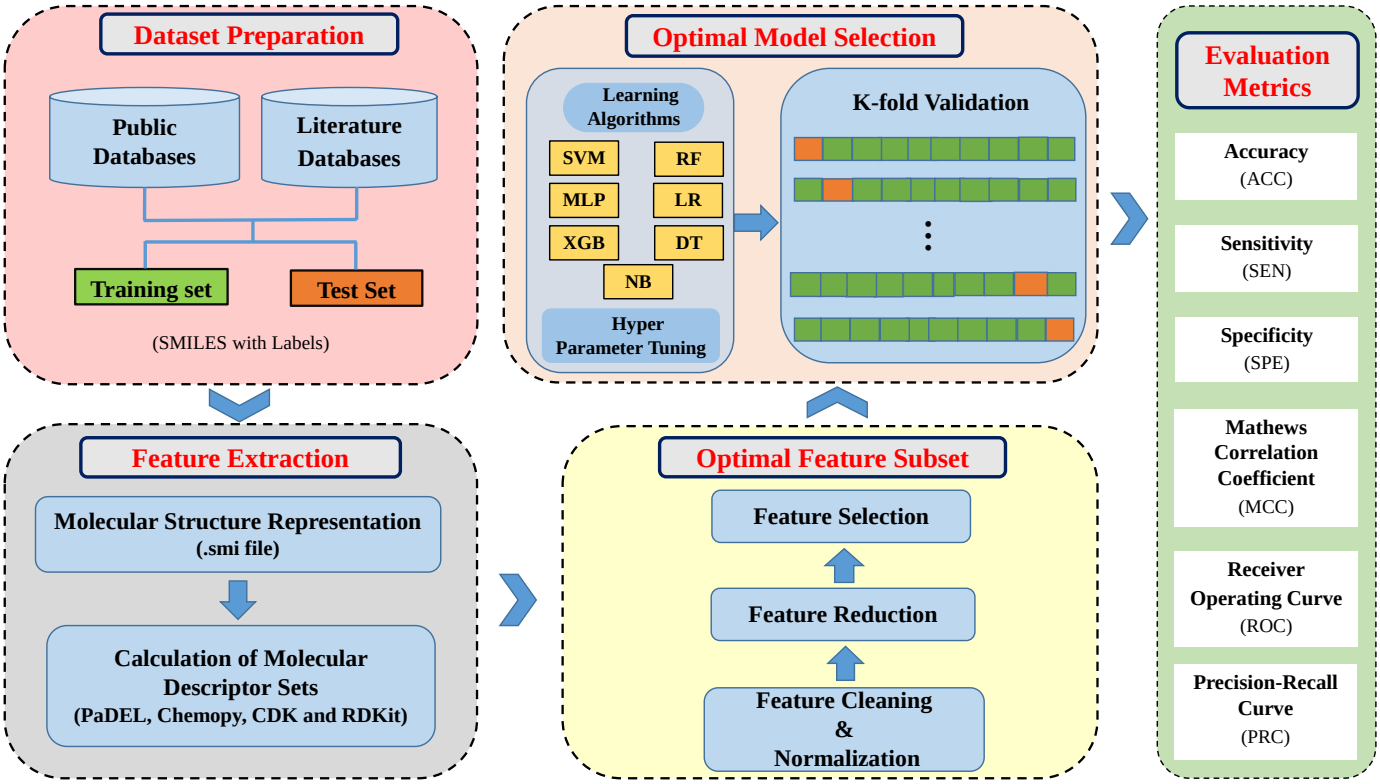

Figure 1. Illustration of the overall workflow.

2. Materials and Methods

2.1. Datasets

We obtained the training dataset compounds to develop a drug-induced liver toxicity prediction model from previously published work [33]. He et al. constructed a training dataset by integrating most of the data from publicly available datasets, i.e., DILIrank [34], LiverTox [35], and LTKB [36], and also performed an extensive study from the PubMed database and various scientific publications [37,38] to retrieve new hepatotoxic and hepaprotective compounds. Furthermore, a crucial data filtering procedure was performed to make a large scale and chemically diverse training set of 1254 compounds. The Simplified Molecular Input Line Entry System (SMILES) for each compound was acquired from the PubChem compound database [39]. We excluded a compound from the previous study because it may create an outlier as most of the compounds have SMILES sequence lengths of less than 150. Finally, our training set has 1253 compounds, consisting of 636 hepatotoxic and 617 non-hepatotoxic compounds. We collected drug compounds for the test dataset from the literature [33,40]. After eliminating duplicate and structurally similar compounds, we randomly selected 208 drug compounds, consisting of 94 hepatotoxic and 114 non-hepatotoxic compounds.The training and test datasets used in this study can be found in the supplementary file (Tables S1 and S2).

2.2. Molecular Descriptors

Molecular descriptors are commonly utilized to quantitatively represent molecular characteristics for drug compounds [41]. We can compute numerous descriptors from the SMILES string format through various open source packages [42–45]. In this study, we calculated four sets of descriptors (CDK, Chemopy, PaDEL and RDKit) using an integrated publicly available web-based platform ChemDes [46]. We used individual descriptor sets and their combinations as shown in Figure 2. The individual descriptor set count is shown in Table 1. In total, 2648 descriptors were computed. In combined descriptor sets, the

redundant descriptors which were calculated by more than one software were eliminated in the data preprocessing phase as discussed in upcoming section.

Table 1. Details of the descriptor sets.

Descriptor Set	Descriptors Count
PaDEL 1&2D	1544
Chemopy 1&2D	633
CDK	275
RDKit	196
Total	2648

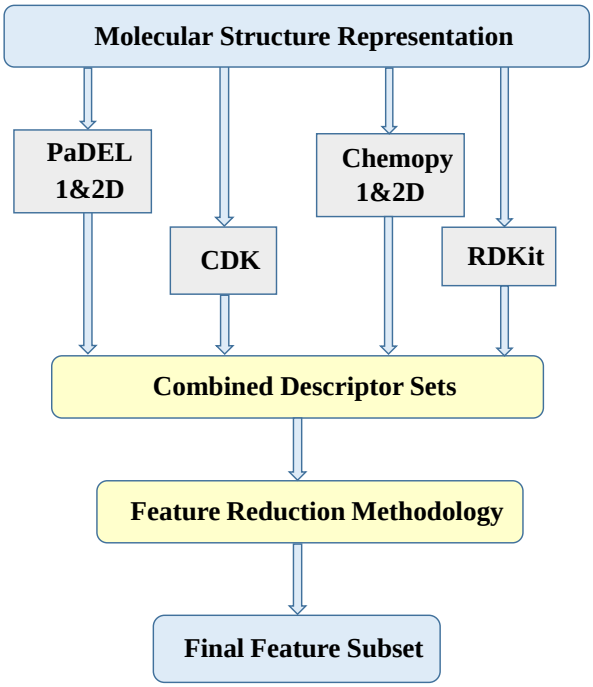

Figure 2. Combining individual descriptor sets.

2.3. Data Preprocessing and Feature Selection

Data preprocessing is an essential step in machine learning modeling as it improves the quality of the data and impacts the learning capability of the model. The descriptor preprocessing, reduction and selection methodology is shown in Figure 3.

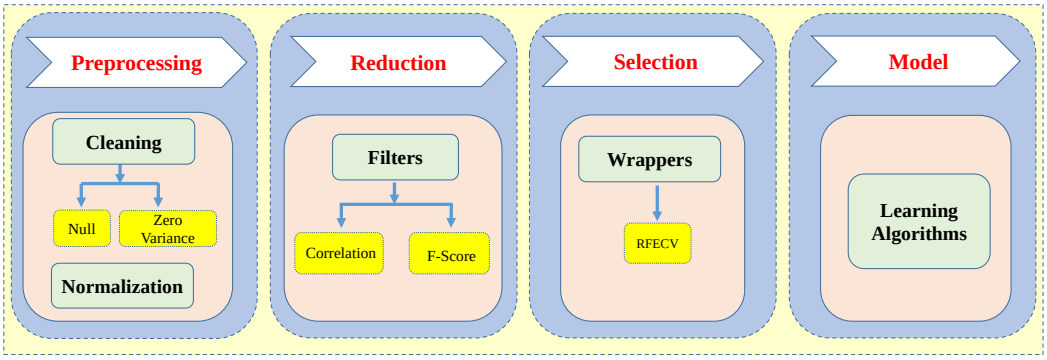

Figure 3. Illustration of the phases to develop final descriptors subset.

The main purpose of data cleaning is to identify and remove the noisy data by dropping the missing and identical value features. A variance threshold algorithm was applied to remove the zero variant features, i.e., the features with the same value in all drug compounds. The selection of most important subset of features is the challenging optimization step in machine learning-based model development. Feature selection techniques reduce the computational cost and complexity of the model. There are several feature selection techniques to select the best molecular descriptor subset for training the model [47]. We utilized the feature selection algorithms implemented by open source machine learning library Scikit-learn [48] in Python. Filter-based selection methods are faster and generally used in the case of the high dimensional features. In filtering, the selection of features i performed without considering the predictive model. The filter-based linear correlation method was used to eliminate the redundant and irrelevant features by using the Pearson correlation coefficient [49]. Molecular descriptors having a mutual correlation of more than 0.9 have been reduced by dropping one of the highly correlated features. The F-score algorithm was implemented to rank all the features according to the feature importance score. The feature importance score was calculated based on the correlation value of each feature with the target label and not considering the mutual information among the features [50].

In addition to filter methods, wrapper methods were proposed to search for the best performing subset of features by iterative training of a supervised learning estimator. Though wrapper-based selection methods are computationally expensive, it avoids over-fitting and improves the learning accuracy of the predictive model. We applied the Recursive Feature Elimination and Cross Validation (RFECV) technique to select the high ranked features by training the SVM linear classifier while recursively eliminating the low importance features [51–54]. The optimal feature subset of 155 molecular descriptors was selected after eliminating 5% of less important molecular descriptors in each iteration using 10-fold CV method. The final optimized subset selected from the training set and external test set was used for model development, internal validation and external validation, respectively.

2.4. Model Building and Optimization

Machine learning models can be used to predict hepatotoxicity given the molecular descriptor of a compound as input. We mainly focused on the following machine learning algorithms to develop binary classification models, among several methods that have been applied in QSAR modeling: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Logistic Regression (LR), Random Forest (RF), XG Boosting (XGB), K-Nearest Neighbors (KNN), Naive Bayes (NB) and Decision Tree (DT) classifier [27,32,33,53,55,56]. These robust algorithms are highly efficient and can accommodate numerous features. We implemented these machine learning algorithms by using the widely used library Scikit-learn in Python [48]. The machine learning algorithms' performances can be effectively improved by tuning their parameter values. The hyper-parameters of the models were optimized by the grid search method with cross-validation over a parameter grid. We trained the optimized algorithms with the best selected molecular descriptors and known hepatotoxicity labels. In this study, an SVM-based binary classifier was mainly used for performance comparison.

Support Vector Machine

SVM is a powerful supervised learning method and widely used for solving classification problems. The algorithm performs the classification by identifying the optimal hyper-plane using several kernel functions that discriminate between the positive and negative class molecules in a high dimensional space. In this study, we used the most popular radial basis function (RBF) as the kernel function which showed better performance than the others (linear, sigmoid and polynomial). In addition, we optimized the penalty parameter C and the kernel coefficient Gamma of the RBF kernel through the grid search method with cross-validation. The regularization parameter C controls the trade-off

between the smooth decision boundaries and correct classification [57]. The higher values of kernel width parameter Gamma denotes an exact fit as per the training dataset and causes an over-fitting problem. The optimal values of C and Gamma used in this study were 100 and 0.01, respectively.

2.5. Model Training and Validation

In this study, the reliability and quality of the proposed model was evaluated by performing external validation in addition to 10-fold cross-validation (CV). In the CV method, the training dataset was randomly divided into 10 subsets. The optimized model was trained with nine subsets and the remaining one subset as a internal validation set. The training and validation procedure was repeated ten times with different training subsets and internal validation sets, respectively. Finally, the performance of the binary classifier was calculated by averaging the results of the 10 corresponding internal validation sets.

2.6. Performance Evaluation Metrics

To assess the predictive ability of the proposed model, we employed several statistical metrics, including accuracy (*ACC*), the overall prediction accuracy; sensitivity (*SEN*), the prediction accuracy of hepatotoxic compounds; specificity (*SPE*), the prediction accuracy of non-hepatotoxic compounds; Matthew's correlation coefficient (*MCC*); and *F*1-*Score*, which are mathematically defined as follows:

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

$$SPE = \frac{TN}{TN + FP} \quad (2)$$

$$SEN = \frac{TP}{TP + FN} \quad (3)$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

$$F1\text{-}Score = \frac{2 * TP}{(2 * TP) + FN + FP} \quad (5)$$

$$RndACC = \frac{(TP + FN) * (TP + FP) + (TN + FP) * (TN + FN)}{N^2} \quad (6)$$

$$\Delta ACC = 100 * (ACC - RndACC)(\%) \quad (7)$$

where true positive (*TP*) denotes the number of hepatotoxic molecules that are predicted correctly, true negative (*TN*) indicates the number of non-hepatotoxic molecules that are predicted correctly, false positive (*FP*) is the count of non-hepatotoxic compounds that are incorrectly predicted as hepatotoxic compounds, false negative (*FN*) is the count of hepatotoxic compounds that are incorrectly predicted as non-hepatotoxic compounds. The MCC is used to measure the balanced classification performance, the coefficient value 1 indicates perfect classification and -1 represents perfect misclassification [58]. The statistical parameter F1-Score is calculated for estimating the quality of binary classification models using an imbalanced dataset. The random accuracy (*RndACC*) and its difference with real accuracy (ΔACC in %) can be estimated to rank the predictive quality of the QSAR models [59]. The receiver operating curves (ROCs) and the precision recall curves (PRCs) were plotted to summarize the binary classification performance. Additionally, we calculated the area under the ROC curve (AUC-ROC) and the area under the PRC curve (AUC-PRC) for classifier comparisons.

3. Results and Discussion

3.1. Data Analysis

To estimate the chemical diversity of the dataset used in this study, we calculated the Tanimotto similarity index [60] based on Morgan Fingerprint with radius 2. The majority of the compounds in the training and test sets had similarity indices in the range below 0.30 and the mean value was only 0.0921. These results suggest that the chemical structures used in our dataset were diverse. We plotted the heat map corresponding to the Tanimotto similarity index of molecules from the entire dataset (Figure 4).

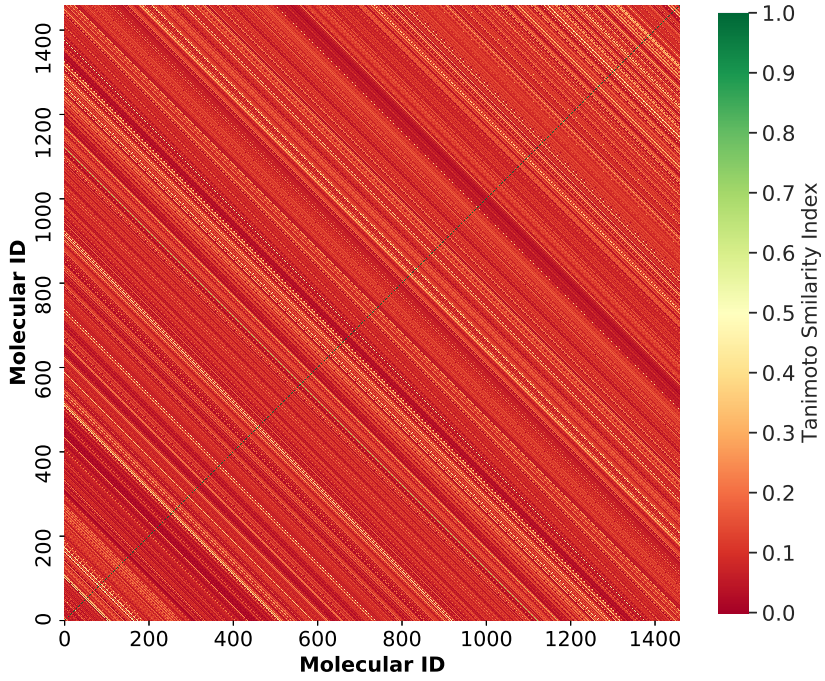

Figure 4. Heat map showing the molecular similarity of the molecules used in the entire dataset plotted by Tanimoto similarity index calculated using Morgan Fingerprints. The *x*-axis and *y*-axis represent the number of molecules used in the whole dataset.

t-distributed stochastic neighbor embedding (t-SNE) is a non-linear technique for dimensionality reduction and it is used to create graphical representation of the chemical space covered by the set of molecules [61]. It is recommended to reduce the number of dimensions before performing the t-SNE algorithm, which will speed up the computation and suppress some noise. In this study, principal component analysis (PCA) was performed on 2048 bit Morgan fingerprints to obtain 100 principal components, which represent 56.76% of the overall variance in the data. Figure 5 represents the chemical space visualization of the entire dataset using t-SNE algorithm. Furthermore, we explored the chemical space of the whole dataset using molecular weight and AlogP (octanol/water partition coefficient) as demonstrated in Figure 6. The molecular weight values varied from 74 to 843.88 and AlogP values ranged from −7.88 to 15.61. The scatter diagram distributions (Figures 5 and 6) illustrate that the hepatotoxic and non-hepatotoxic compounds shared the same chemical space.

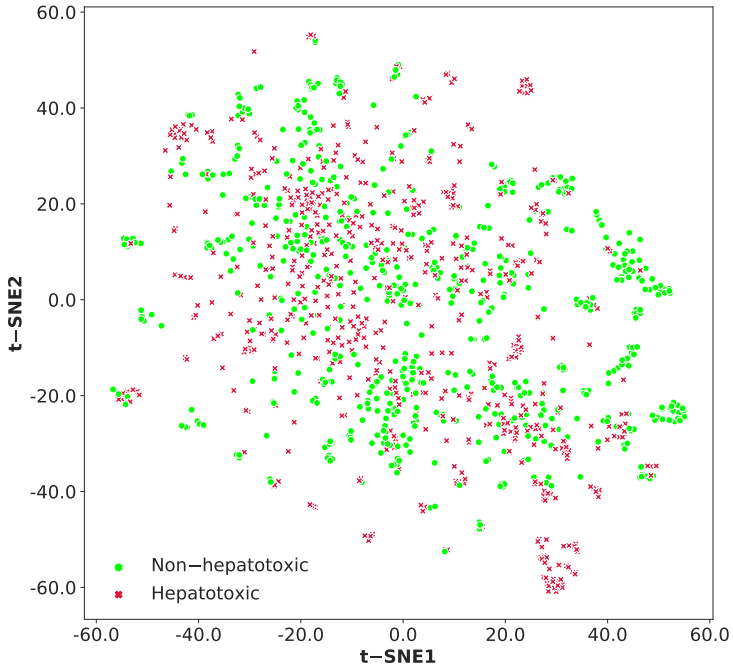

Figure 5. Visualizing the chemical space coverage using t-distributed stochastic neighbor embedding (t-SNE) on PCA output with 100 principal components (accounts for 56.76% of the overall variance) of the entire dataset (training and test datasets). Red x markers represent the hepatotoxic compounds and the green circle markers represent the non-hepatotoxic compounds.

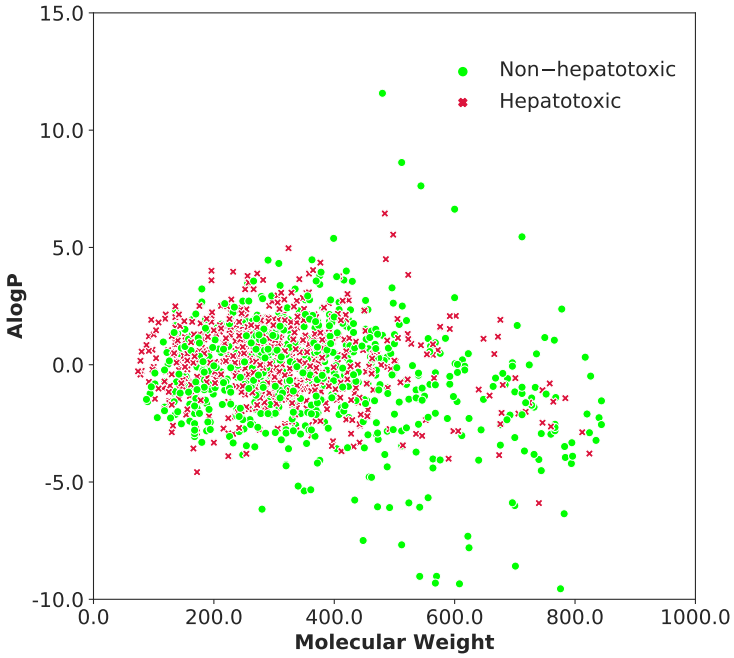

Figure 6. Chemical space defined by molecular weight and AlogP of the entire dataset (training and test datasets). Red x markers represent the hepatotoxic compounds and the green circle markers represent the non-hepatotoxic compounds.

3.2. Performance of Models Using Cross-Validation

Various machine learning methods were used to build prediction models based on molecular descriptor subsets. Initially, we worked on individual descriptor sets and evaluated their performances with few machine learning models. Then, we made different combinations of molecular descriptor subsets and found a good combination of the descrip-

tor set that produced the best prediction results for the validation data. The best prediction model was built based on the selected combination of the individual descriptor sets.

3.2.1. Experiments with Individual Descriptor Sets

We employed the experimental workflow approach on the four individual descriptor sets such as CDK, Chemopy, PaDEL and RDKit. At first, empty valued and zero variant features were removed from the original molecular descriptor set. After applying the feature importance score-based ranking and RFECV techniques, we selected an optimal subset of molecular descriptors from each set for model training and validation. The SVM-based classification model was used to compare the 10-fold cross-validation performance of the individual descriptor sets.

Table 2 shows the selected number of descriptors from each set and their prediction performance results. It is evident that the PaDEL descriptor set performs better in terms of accuracy (ACC), Sensitivity (SEN) and Mathews Correlation Coefficient (MCC) than the other individual descriptor sets.

Table 2. The best performance of individual descriptor sets.

Descriptor Set	Selected No. of Descriptors	ACC	SPE	SEN	MCC
CDK	77	0.771	0.772	0.773	0.545
Chemopy	104	0.766	0.763	0.773	0.536
PaDEL	91	0.781	0.749	0.815	0.565
RDKit	86	0.752	0.729	0.777	0.507

3.2.2. Experiments with Combined Descriptor Sets

We examined various combinations of the individual descriptor sets to improve the model performance. Here, the similar experimental workflow, i.e., the preprocessing and feature selection methods, were also applied for the combined feature sets to obtain the low dimensional optimal descriptor subset. The PaDEL descriptor set was present in all the combination groups as it computes a large number of descriptors and showed better prediction performance compared to the other individual descriptor sets. The optimal number of descriptors and their SVM classifier prediction outcomes for different possible combinations are shown in Table 3.

Table 3. Performance details of best combination descriptor sets.

Best Combination Descriptor Sets	Optimal No. of Descriptors	ACC	SPE	SEN	MCC
PaDEL-RDKit	132	0.796	0.784	0.809	0.593
PaDEL-RDKit-CDK	162	0.804	0.796	0.813	0.609
PaDEL-RDKit-CDK-Chemopy	155	0.811	0.783	0.840	0.623

From Table 3, it can be seen that every combined descriptor set showed improved prediction ACC and MCC compared to the best performing PaDEL descriptor set with 91 optimal descriptors. The number of descriptors for each group has been obtained as the result of optimizing the feature selection algorithms mentioned in the methods section. The combined group of three descriptor sets with 162 optimal features gave improved MCC over PaDEL-RDKit combination with 132 features. The combination of all the individual descriptor sets selected less than 6% of features after feature reduction and selection steps from the total number of 2648 original features. This combo showed an improved prediction in 10-fold cross-validation compared to all other combinations with respect to evaluation metrics ACC, SEN and MCC.

The details of 155 best selected descriptor subsets from combination of all the descriptor sets are given in the supplementary file (Table S3). Most of the selected features were from the PaDEL descriptor set and belong to auto-correlation, E-state and topological

descriptors (Table 4). From Chemopy 1&2D, most of the descriptors are E-state and MOE (Molecular Operating Environment) descriptors. For CDK, topological and Kappa descriptors represented the major part. At last, all the selected RDKit descriptors are constitutional descriptors.

Table 4. Details of the selected optimal descriptor subset.

Descriptor Set	Descriptor Type	Total-Type	Total-Set	% of Selection
PaDEL 1&2D	Autocorrelation Descriptors	46	83	54
	E-state Descriptors	13		
	Topological Descriptors	11		
	Constitutional Descriptors	11		
	Others	2		
Chemopy 1&2D	MOE-type descriptors	11	37	24
	E-state Descriptors	11		
	Autocorrelation Descriptors	10		
	Others	5		
CDK	Topological Descriptors	17	28	18
	Kappa Descriptors	5		
	Autocorrelation Descriptors	3		
	Others	5		
RDKit	Constitutional descriptors	7	7	4

The Shapeley Additive Explanations (SHAP) technique is adopted to understand the most important descriptors and their contribution to the model prediction [62,63]. The SHAP technique is based on the game theory approach and was developed using Python. Figure 7 shows the summary plot of top 20 descriptors used for training the proposed SVM model. The SHAP summary plot indicates the relationship between the descriptor value and its impact on the model prediction. In the violin plot, the red color indicates the higher feature values and the blue color indicates the lower feature values. The descriptors are ordered based on their importance. The E-state descriptor "minHsOH" is the primary feature and it causes either a large positive or large negative in the model outcome and "maxHBint5" is the next most important descriptor.

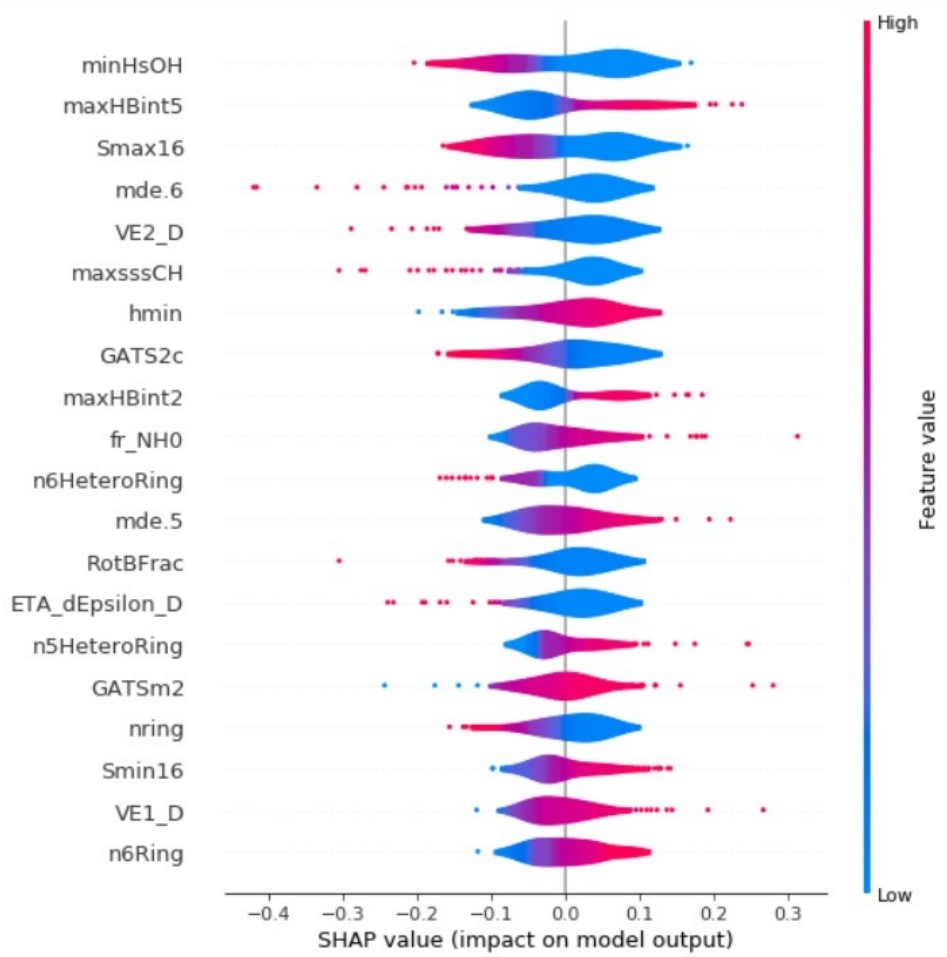

Figure 7. SHAP summary plot displays the distribution of top 20 important descriptors used for training the proposed model for hepatotoxicity prediction.

3.2.3. Comparison of SVM with Other Classifiers

We evaluated the performance of the SVM classifier built on the final selected optimal descriptors subset with other conventional supervised learning methods such as Multi-Layer Perceptron (MLP), Logistic Regression (LR), Random Forest (RF), XG Boosting (XGB), K-Nearest Neighbor (KNN), Naive Bayes (NB) and Decision Tree (DT) classifier. The performance evaluation metrics comparison of all classifiers using a 10-fold cross-validation is shown in Figure 8. The comparison results confirm that the SVM-based binary classification model is performing better for drug-induced liver toxicity prediction. Particularly, the accuracy of SVM is 14.2% more than DT, 13% more than NB, 6% more than KNN, 4.8% more than XGB, 4.3% more than RF, 3.3% more than LR, 2% more than MLP. The MCC value of the SVM-based model is above 0.6. In addition, the F1-score value of the proposed SVM-based model is above 0.8, which shows the quality of the binary classification model using imbalanced data.

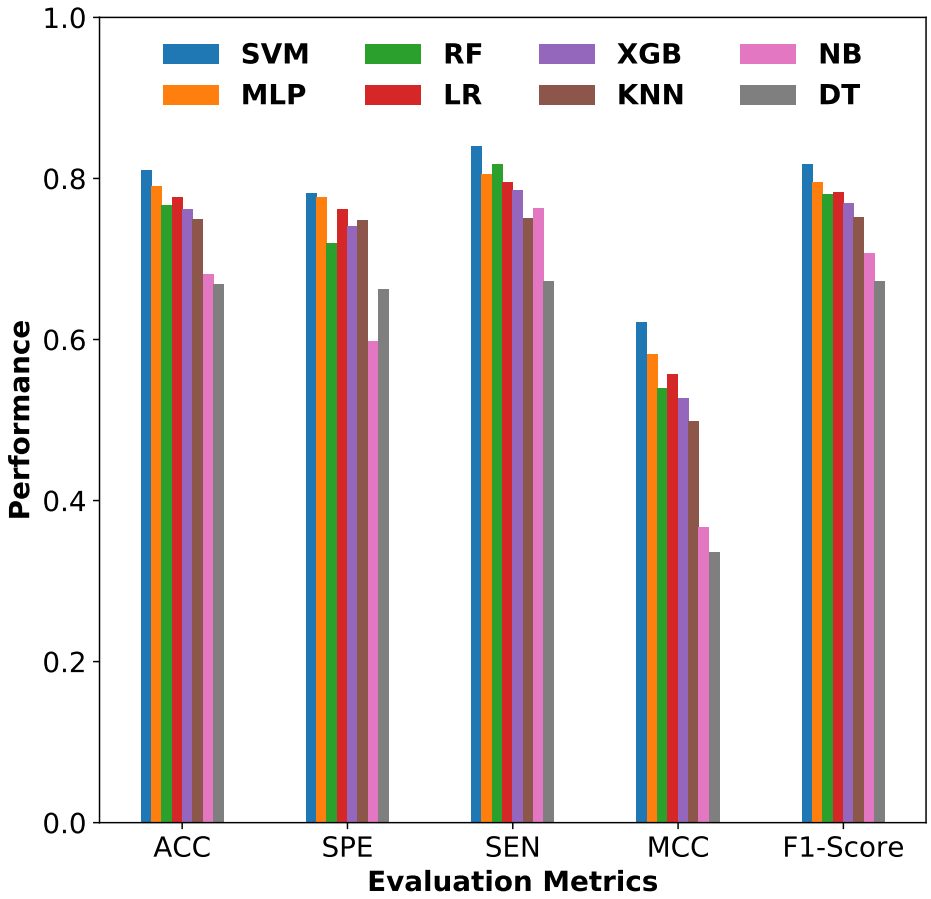

Figure 8. Performance comparison of various classifiers using 10-fold cross-validation.

The most probable random accuracy (RndACC) is calculated from confusion matrix values for all the models used for performance comparison and all the models reported in this study have a maximum random accuracy of value 0.5. The difference (ΔACC in %) between the real model accuracy (ACC) and the random accuracy was also calculated and used for ranking the models. The descending order ranking of the models based on this values is SVM, MLP, LR, RF, XGB, KNN, NB, DT classifier. The proposed SVM-based model has the highest accuracy difference value among all the models used for comparison. The confusion matrix values (TN, FP, FN & TP) and other evaluation parameter values (ACC, SPE, SEN, MCC, F1-Score, RndACC& ΔACC) of the 10-fold cross-validation for each model used in the comparison is provided in the supplementary file (Table S4).

The comparison of receiver operating characteristic (ROC) curves and precision recall curves (PRCs) with area under the curve (AUC) values for all the models are shown in Figures 9 and 10. As depicted from the results, SVM with the Radial Basis Function (RBF) kernel achieved AUC-ROC of 0.811 and AUC-PRC of 0.860, which are relatively better than all other methods.

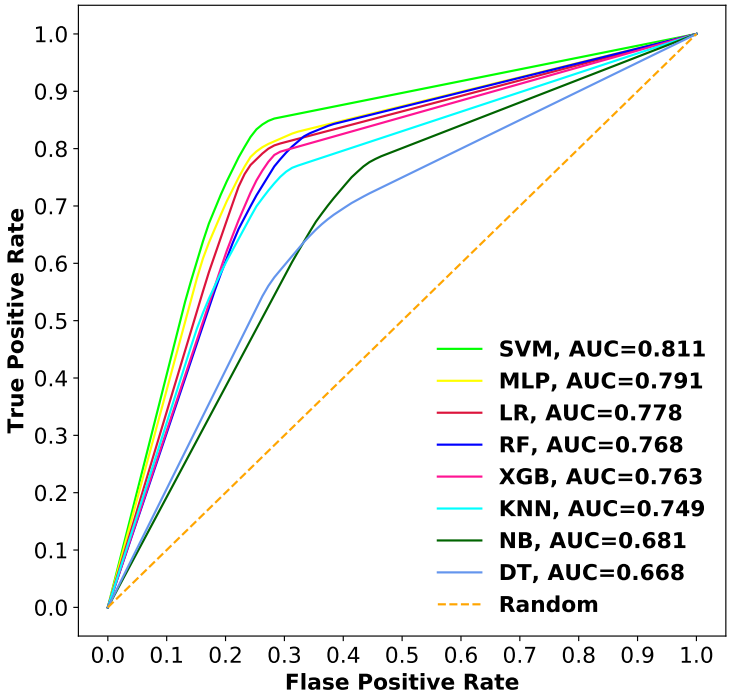

Figure 9. ROC comparison of different classifiers with corresponding AUC values using 10-fold cross-validation.

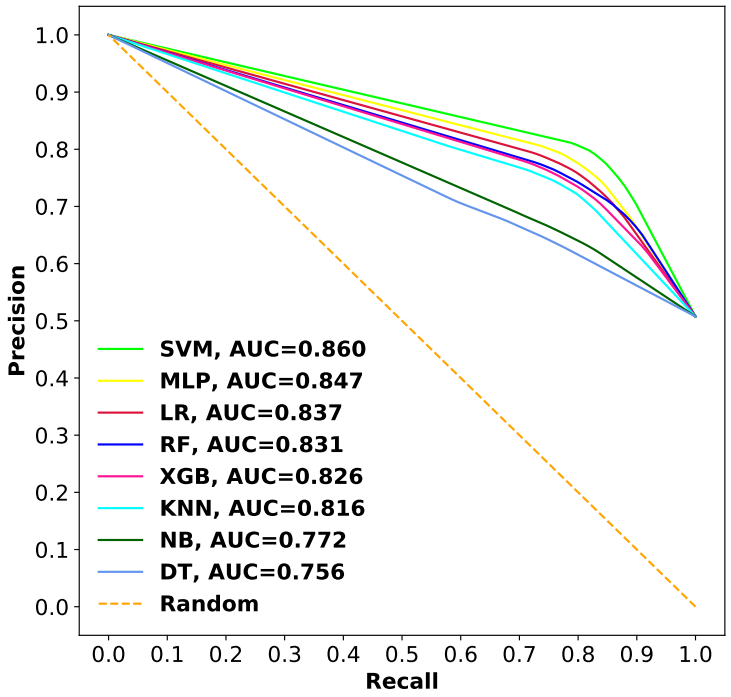

Figure 10. PRC comparison of different classifiers with corresponding AUC values using 10-fold cross-validation.

3.3. Performance Comparison with Previous Work

Various QSAR models have been published for drug-induced liver toxicity prediction by using machine learning algorithms [27–29,31,32,37]. We only selected computational models that were cross-validated [27,28,32,33] for comparing with our proposed model. Although the proposed prediction model showed good performance in internal validation, it is required to perform the external validation to determine the robustness of the SVM-

based model. In particular, the external validation dataset used in this study did not have any identical or high structural similarity compounds with the training dataset.

Table 5 shows that the results of cross-validation and external validation set comparison between proposed model with the previously published models. The accuracy of the proposed model is better than the other methods considered for comparison. The recently published ensemble model [33] yielded an ACC of 0.783, SPE of 0.748, and SEN of 0.818 with the 10-fold cross-validation. These results were improved with our proposed model, ACC by 2.8%, SPE by 3.5%, SEN by 2.2% and with a good MCC value of 0.623 (MCC was not given in the prior work). The SVM-based proposed model also achieved good performance compared to the ensemble model in external validation. Thus, the proposed SVM-based model is a promising hepatotoxicity predictor compared to the ensemble model.

Table 5. Proposed model performance compared with the previously published literature.

Model Name	No. of Compounds	Test Method	ACC	SPE	SEN
Proposed Model	1253	10-fold CV	0.811	0.783	0.840
	208	External Validation	0.756	0.708	0.807
Ensemble Model [33]	1254	10-fold CV	0.783	0.748	0.818
	204	External Validation	0.730	0.658	0.773
Ensemble-Top5 [32]	1241	5-fold CV	0.711	0.603	0.799
SVM [27]	978	5-fold CV	0.797	0.585	0.948
	88	External Validation	0.750	0.379	0.932
RF [28]	996	10-fold CV	0.65	0.62	0.68
	966	External Validation	0.58	0.38	0.75

4. Conclusions

Drug-induced liver toxicity estimation is one of the significant safety related challenges in the pharmaceutical industry. In this study, we focused on the prediction of liver toxicity based on computational models using a large and diverse dataset of 1253 unique compounds. We used a total of 2648 molecular descriptors calculated from four different descriptor sets as modelling features. Initially, null values and highly correlated features were dropped from the high dimensional feature space, and then feature selection techniques were applied to select the optimal subset of molecular descriptors for effective model training. Eight different supervised learning models were constructed and optimized with the best selected final features and their cross-validation prediction performance was analyzed. The SVM-based binary classification models utilizing less than 6% of the original features achieved improved performance compared to the other machine learning models. Moreover, the proposed model demonstrated better performance than the previous study in 10-fold cross-validation and external validation. It was observed from the comparison that the extended molecular descriptor feature space could improve the prediction performance. Meanwhile, the selection of discriminating model features is also a challenging task to obtain good prediction results. In the future, with great understanding of drug-induced liver toxicity mechanisms, we intend to investigate deep learning architectures using improved dataset considering biological data along with the chemical structure for improving the hepatotoxicity prediction. Additionally, a large-scale dataset with standard DILI definition and dose-level information will aid to build an efficient models for DILI assessment.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms22158073/s1.

Author Contributions: Conceptualization, K.J., H.T. and K.T.C.; methodology, K.J.; software, K.J. and H.T.; validation, K.J., H.T. and K.T.C.; investigation, K.J., H.T. and K.T.C.; writing—original draft

preparation, K.J.; writing—review and editing, K.J., H.T. and K.T.C.; supervision, H.T. and K.T.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C2005612) and in part by the Brain Research Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No.NRF-2017M3C7A1044816).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were generated in this study.

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

References

1. Almazroo, O.A.; Miah, M.K.; Venkataramanan, R. Drug metabolism in the liver. *Clin. Liver Dis.* **2017**, *21*, 1–20. [CrossRef]
2. Real, M.; Barnhill, M.S.; Higley, C.; Rosenberg, J.; Lewis, J.H. Drug-induced liver injury: Highlights of the recent literature. *Drug Saf.* **2019**, *42*, 365–387. [CrossRef]
3. Albrecht, W.; Kappenberg, F.; Brecklinghaus, T.; Stoeber, R.; Marchan, R.; Zhang, M.; Ebbert, K.; Kirschner, H.; Grinberg, M.; Leist, M.; et al. Prediction of human drug-induced liver injury (DILI) in relation to oral doses and blood concentrations. *Arch. Toxicol.* **2019**, *93*, 1609–1637. [CrossRef]
4. Lee, W.M. Drug-induced hepatotoxicity. *N. Engl. J. Med.* **2003**, *349*, 474–485. [CrossRef] [PubMed]
5. Onakpoya, I.J.; Heneghan, C.J.; Aronson, J.K. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: A systematic review of the world literature. *BMC Med.* **2016**, *14*, 10. [CrossRef]
6. Njoku, D.B. Drug-induced hepatotoxicity: Metabolic, genetic and immunological basis. *Int. J. Mol. Sci.* **2014**, *15*, 6990–7003. [CrossRef] [PubMed]
7. Garcia-Cortes, M.; Robles-Diaz, M.; Stephens, C.; Ortega-Alonso, A.; Lucena, M.I.; Andrade, R.J. Drug induced liver injury: An update. *Arch. Toxicol.* **2020**, *94*, 3381–3407. [CrossRef] [PubMed]
8. Assis, D.N.; Navarro, V.J. Human drug hepatotoxicity: A contemporary clinical perspective. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 463–473. [CrossRef]
9. Chen, M.; Vijay, V.; Shi, Q.; Liu, Z.; Fang, H.; Tong, W. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov. Today* **2011**, *16*, 697–703. [CrossRef]
10. Issa, A.M.; Phillips, K.A.; Van Bebber, S.; Nidamarthy, H.G.; Lasser, K.E.; Haas, J.S.; Alldredge, B.K.; Wachter, R.M.; Bates, D.W. Drug withdrawals in the United States: A systematic review of the evidence and analysis of trends. *Curr. Drug Saf.* **2007**, *2*, 177–185. [CrossRef]
11. Suh, J.I. Drug-induced liver injury. *Yeungnam Univ. J. Med.* **2020**, *37*, 2. [CrossRef]
12. Olson, H.; Betton, G.; Robinson, D.; Thomas, K.; Monro, A.; Kolaja, G.; Lilly, P.; Sanders, J.; Sipes, G.; Bracken, W.; et al. Concordance of the toxicity of pharmaceuticals in humans and in animals. *Regul. Toxicol. Pharmacol.* **2000**, *32*, 56–67. [CrossRef]
13. Sertkaya, A.; Wong, H.H.; Jessup, A.; Beleche, T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin. Trials* **2016**, *13*, 117–126. [CrossRef] [PubMed]
14. Sistare, F.D.; Mattes, W.B.; LeCluyse, E.L. The promise of new technologies to reduce, refine, or replace animal use while reducing risks of drug induced liver injury in pharmaceutical development. *ILAR J.* **2017**, *57*, 186–211. [CrossRef] [PubMed]
15. Tai, W.; He, L.; Zhang, X.; Pu, J.; Voronin, D.; Jiang, S.; Zhou, Y.; Du, L. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: Implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell. Mol. Immunol.* **2020**, *17*, 613–620. [CrossRef] [PubMed]
16. Segall, M.D.; Barber, C. Addressing toxicity risk when designing and selecting compounds in early drug discovery. *Drug Discov. Today* **2014**, *19*, 688–693. [CrossRef]
17. Przybylak, K.R.; Cronin, M.T. In silico models for drug-induced liver injury–current status. *Expert Opin. Drug Metab. Toxicol.* **2012**, *8*, 201–217. [CrossRef]
18. Saini, N.; Bakshi, S.; Sharma, S. In-silico approach for drug induced liver injury prediction: Recent advances. *Toxicol. Lett.* **2018**, *295*, 288–295. [CrossRef]
19. Thakkar, S.; Chen, M.; Fang, H.; Liu, Z.; Roberts, R.; Tong, W. The Liver Toxicity Knowledge Base (LKTB) and drug-induced liver injury (DILI) classification for assessment of human liver injury. *Expert Rev. Gastroenterol. Hepatol.* **2018**, *12*, 31–38. [CrossRef]
20. Chierici, M.; Francescatto, M.; Bussola, N.; Jurman, G.; Furlanello, C. Predictability of drug-induced liver injury by machine learning. *Biol. Direct* **2020**, *15*, 1–10. [CrossRef]
21. Xu, Y.; Dai, Z.; Chen, F.; Gao, S.; Pei, J.; Lai, L. Deep learning for drug-induced liver injury. *J. Chem. Inf. Modeling* **2015**, *55*, 2085–2093. [CrossRef] [PubMed]
22. Kuna, L.; Bozic, I.; Kizivat, T.; Bojanic, K.; Mrso, M.; Kralj, E.; Smolic, R.; Wu, G.Y.; Smolic, M. Models of drug induced liver injury (DILI)–current issues and future perspectives. *Curr. Drug Metab.* **2018**, *19*, 830–838. [CrossRef]

23. Marchant, C.A.; Fisk, L.; Note, R.R.; Patel, M.L.; Suárez, D. An expert system approach to the assessment of hepatotoxic potential. *Chem. Biodivers.* **2009**, *6*, 2107–2114. [CrossRef]
24. Greene, N.; Fisk, L.; Naven, R.T.; Note, R.R.; Patel, M.L.; Pelletier, D.J. Developing structure- activity relationships for the prediction of hepatotoxicity. *Chem. Res. Toxicol.* **2010**, *23*, 1215–1222. [CrossRef] [PubMed]
25. Pizzo, F.; Lombardo, A.; Manganaro, A.; Benfenati, E. A new structure-activity relationship (SAR) model for predicting drug-induced liver injury, based on statistical and expert-based structural alerts. *Front. Pharmacol.* **2016**, *7*, 442. [CrossRef] [PubMed]
26. Ekins, S.; Williams, A.J.; Xu, J.J. A predictive ligand-based Bayesian model for human drug-induced liver injury. *Drug Metab. Dispos.* **2010**, *38*, 2302–2308. [CrossRef]
27. Zhang, H.; Ding, L.; Zou, Y.; Hu, S.Q.; Huang, H.G.; Kong, W.B.; Zhang, J. Predicting drug-induced liver injury in human with Naïve Bayes classifier approach. *J. Comput. Aided Mol. Des.* **2016**, *30*, 889–898. [CrossRef]
28. Kotsampasakou, E.; Montanari, F.; Ecker, G.F. Predicting drug-induced liver injury: The importance of data curation. *Toxicology* **2017**, *389*, 139–145. [CrossRef]
29. Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P.W.; Tang, Y. In silico prediction of drug induced liver toxicity using substructure pattern recognition method. *Mol. Inform.* **2016**, *35*, 136–144. [CrossRef] [PubMed]
30. Kim, E.; Nam, H. Prediction models for drug-induced hepatotoxicity by using weighted molecular fingerprints. *BMC Bioinform.* **2017**, *18*, 227. [CrossRef] [PubMed]
31. Mulliner, D.; Schmidt, F.; Stolte, M.; Spirkl, H.P.; Czich, A.; Amberg, A. Computational models for human and animal hepatotoxicity with a global application scope. *Chem. Res. Toxicol.* **2016**, *29*, 757–767. [CrossRef]
32. Ai, H.; Chen, W.; Zhang, L.; Huang, L.; Yin, Z.; Hu, H.; Zhao, Q.; Zhao, J.; Liu, H. Predicting drug-induced liver injury using ensemble learning methods and molecular fingerprints. *Toxicol. Sci.* **2018**, *165*, 100–107. [CrossRef]
33. He, S.; Ye, T.; Wang, R.; Zhang, C.; Zhang, X.; Sun, G.; Sun, X. An in silico model for predicting drug-induced hepatotoxicity. *Int. J. Mol. Sci.* **2019**, *20*, 1897. [CrossRef] [PubMed]
34. Chen, M.; Suzuki, A.; Thakkar, S.; Yu, K.; Hu, C.; Tong, W. DILIrank: The largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov. Today* **2016**, *21*, 648–653. [CrossRef]
35. Hoofnagle, J.H.; Serrano, J.; Knoben, J.E.; Navarro, V.J. *LiverTox: A website on Drug-Induced Liver Injury*; Wiley Online Library: Hoboken, NJ, USA, 2013.
36. Chen, M.; Zhang, J.; Wang, Y.; Liu, Z.; Kelly, R.; Zhou, G.; Fang, H.; Borlak, J.; Tong, W. The liver toxicity knowledge base: A systems approach to a complex end point. *Clin. Pharmacol. Ther.* **2013**, *93*, 409–412. [CrossRef] [PubMed]
37. Liew, C.Y.; Lim, Y.C.; Yap, C.W. Mixed learning algorithms and features ensemble in hepatotoxicity prediction. *J. Comput. Aided Mol. Des.* **2011**, *25*, 855–871. [CrossRef]
38. Huang, S.H.; Tung, C.W.; Fülöp, F.; Li, J.H. Developing a QSAR model for hepatotoxicity screening of the active compounds in traditional Chinese medicines. *Food Chem. Toxicol.* **2015**, *78*, 71–77. [CrossRef]
39. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.; Thiessen, P.; Yu, B.; et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 2019,*47*, *D1102–D1109*. [CrossRef]
40. Ivanov, S.; Semin, M.; Lagunin, A.; Filimonov, D.; Poroikov, V. In Silico Identification of Proteins Associated with Drug-Induced Liver Injury Based on the Prediction of Drug-Target Interactions. *Mol. Inform.* **2017**, *36*, 1600142. [CrossRef]
41. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 41.
42. Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E.L. Recent developments of the chemistry development kit (CDK)-an open-source java library for chemo-and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120. [CrossRef] [PubMed]
43. Cao, D.S.; Xu, Q.S.; Hu, Q.N.; Liang, Y.Z. ChemoPy: Freely available python package for computational biology and chemoinformatics. *Bioinformatics* **2013**, *29*, 1092–1094. [CrossRef] [PubMed]
44. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474. [CrossRef]
45. Landrum, G. Rdkit documentation. *Release* **2013**, *1*, 4.
46. Dong, J.; Cao, D.S.; Miao, H.Y.; Liu, S.; Deng, B.C.; Yun, Y.H.; Wang, N.N.; Lu, A.P.; Zeng, W.B.; Chen, A.F. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminform.* **2015**, *7*, 1–10. [CrossRef] [PubMed]
47. Saeys, Y.; Inza, I.; Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517. [CrossRef]
48. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
49. Freedman, D.A. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*; Cambridge University Press: Cambridge, UK, 2010.
50. Güneş, S.; Polat, K.; Yosunkaya, Ş. Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome. *Expert Syst. Appl.* **2010**, *37*, 998–1004. [CrossRef]
51. Darst, B.F.; Malecki, K.C.; Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* **2018**, *19*, 1–6. [CrossRef]

52. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
53. Abdelbaky, I.; Tayara, H.; Chong, K.T. Prediction of kinase inhibitors binding modes with machine learning and reduced descriptor sets. *Sci. Rep.* **2021**, *11*, 1–13. [CrossRef]
54. Khanal, J.; Lim, D.Y.; Tayara, H.; Chong, K.T. i6ma-stack: A stacking ensemble-based computational prediction of dna n6-methyladenine (6ma) sites in the rosaceae genome. *Genomics* **2021**, *113*, 582–592. [CrossRef] [PubMed]
55. Minerali, E.; Foil, D.H.; Zorn, K.M.; Lane, T.R.; Ekins, S. Comparing Machine Learning Algorithms for Predicting Drug-Induced Liver Injury (DILI). *Mol. Pharm.* **2020**, *17*, 2628–2637. [CrossRef]
56. Li, X.; Chen, Y.; Song, X.; Zhang, Y.; Li, H.; Zhao, Y. The development and application of in silico models for drug induced liver injury. *RSC Adv.* **2018**, *8*, 8101–8111. [CrossRef]
57. Ben-Hur, A.; Weston, J. A user's guide to support vector machines. In *Data Mining Techniques for the Life Sciences*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 223–239.
58. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 1–13. [CrossRef] [PubMed]
59. Lučić, B.; Batista, J.; Bojović, V.; Lovrić, M.; Kržić, A.S.; Bešlo, D.; Nadramija, D.; Vikić-Topić, D. Estimation of random accuracy and its use in validation of predictive quality of classification models within predictive challenges. *Croat. Chem. Acta* **2019**, *92*, 379–391. [CrossRef]
60. Butina, D. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750. [CrossRef]
61. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
62. Zhang, Z.; Beck, M.W.; Winkler, D.A.; Huang, B.; Sibanda, W.; Goyal, H. Opening the black box of neural networks: Methods for interpreting neural network models in clinical applications. *Ann. Transl. Med.* **2018**, *6*, 216. [CrossRef]
63. Alam, W.; Tayara, H.; Chong, K.T. XG-ac4C: Identification of N4-acetylcytidine (ac4C) in mRNA using eXtreme gradient boosting with electron-ion interaction pseudopotentials. *Sci. Rep.* **2020**, *10*, 1–10.