Reliably Filter Drug-Induced Liver Injury Literature With Natural Language Processing and Conformal Prediction

Xianghao Zhan, Fanjin Wang, and Olivier Gevaert

***Abstract*—** **Drug-induced liver injury describes the adverse effects of drugs that damage the liver. Life-threatening results were also reported in severe cases. Therefore, liver toxicity is an important assessment for new drug candidates. These reports are documented in research papers that contain preliminary *in vitro* and *in vivo* experiments. Conventionally, data extraction from publications relies on resource-demanding manual labeling, which restricts the efficiency of the information extraction. The development of natural language processing techniques enables the automatic processing of biomedical texts. Herein, based on around 28,000 papers (titles and abstracts) provided by the Critical Assessment of Massive Data Analysis challenge, this study benchmarked model performances on filtering liver-damage-related literature. Among five text embedding techniques, the model using term frequency-inverse document frequency (TF-IDF) and logistic regression outperformed others with an accuracy of 0.957 on the validation set. Furthermore, an ensemble model with similar overall performances was developed with a logistic regression model on the predicted probability given by separate models with different vectorization techniques. The ensemble model achieved a high accuracy of 0.954 and an F1 score of 0.955 in the hold-out validation data in the challenge. Moreover, important words in positive/negative predictions were identified *via* model interpretation. The prediction reliability was quantified with conformal prediction, which provides users with a control over the prediction uncertainty. Overall, the ensemble model and TF-IDF model reached satisfactory classification results, which can be used by researchers to rapidly filter literature that describes events related to liver injury induced by medications.**

***Index Terms*—Drug-induced liver injury, natural language processing, ensemble learning, sentence embedding, conformal prediction.**

Manuscript received 4 February 2022; revised 9 May 2022; accepted 18 July 2022. Date of publication 25 July 2022; date of current version 5 October 2022. *(Xianghao Zhan and Fanjin Wang contributed equally to this work.) (Corresponding author: Olivier Gevaert.)*

Xianghao Zhan is with the Department of Bioengineering, Stanford University, Stanford, CA 94305 USA (e-mail: xzhan96@stanford.edu).

Olivier Gevaert is with the Department of Biomedical Data Science, Stanford University, Stanford, CA 94305 USA (e-mail: olivier.gevaert@stanford.edu).

Fanjin Wang is with the Department of Pharmaceutics, UCL School of Pharmacy, University College London, WC1N 1AX London, U.K. (e-mail: fanjin.wang.20@ucl.ac.uk).

Digital Object Identifier 10.1109/JBHI.2022.3193365

I. Introduction

DURG-INDUCED liver injury (DILI) is defined as the unexpected adverse reaction of the liver to drugs. DILI is a common and critical cause of liver injury because liver plays a key role in drug metabolism [1]. Liver toxicity caused by drugs can be classified into two types: intrinsic and idiosyncratic. Intrinsic liver toxicity of drugs is more predictable and is directly related to the dosage of a specific drug. The damage to the liver occurs within a short time window, typically several hours after administration of the drugs. In comparison, idiosyncratic liver toxicity is more patient-specific and has a longer onset of occurrence. For drugs with high lipophilicity, idiosyncratic liver damage could be triggered even below the recommended daily dosage [2]. The severity of DILI can be different among different patients considering the interaction of genetic and environmental factors [3]. Although most patients can recover from DILI, DILI cases may lead to acute liver failure [4]. For example, the intrinsic liver toxicity of paracetamol, often caused by overdosing, is reported to account for 73.7% of acute liver injury and acute liver failure in Scotland from 1992 to 2014 [5]. Additionally, approximately 75% of the idiosyncratic drug reactions result in liver transplantation or death [1]. Therefore, DILI has become one of the most common reasons that reject the promising novel drug candidates and is strictly evaluated during the drug development process [3].

The complex mechanism of DILI and the severity of the DILI consequences call for a better monitoring of DILI events [3]. However, the majority of DILI reports are from clinical practices or experimental studies in the free text of publications. Conventionally, scientific publications need to be manually checked and processed by researchers and pharmacists. However, thousands of new articles are published in journals on a daily basis, let alone millions of previous publications on the PubMed archive, making it almost impossible for manual inspection. Recently, the rapid development of natural language processing (NLP) technology has enabled data mining applications based on free text. To give some examples, long short-term memory (LSTM) structure in recurrent neural networks (RNN) allows the understanding of long-term dependencies in texts [6]. Bidirectional encoder representation from transformers (BERT) has been developed as a pre-trained language model for understanding text information [7]. Generally, to use these learning algorithms, words are converted into vectors using word vectorization (embedding) techniques like bag-of-words (BOW), term frequency-inverse document frequency (TF-IDF), and Word2Vec. BOW is the most straightforward word embedding method. It counts the time that a word appears in the document. However, this leads to a very sparse feature matrix since only a few words in the whole vocabulary (the collection of words) will appear in the document. As its name suggests, TF-IDF uses the term frequency and the inverse document frequency to represent a document. It assigns more importance to less-frequently occurring words which might contain more meaningful information in a document. On top of these methods, Word2Vec embedding uses a pre-trained neural network to vectorize words to fixed-length vectors. These techniques made it possible to process scientific publications automatically. For example, Wang *et al.* constructed an NLP model to extract clinical information to support clinical decisions [8]. Zhan *et al.* used TF-IDF as the word embedding and logistic regression to extract ICD-10 codes of common cardiovascular disease from electronic health records [9]. Thus, it is promising to utilize NLP techniques to expedite the labelling process of the publications with DILI results and enable researchers to fast filter the literature.

Besides the success of individual classification models developed on different NLP text vectorization techniques, another machine learning strategy, ensemble learning has gained success over past decades [10]. Ensemble learning relies on a set of models and develops a model (known as the meta learner) to combine the different individual models (known as separate learners) under certain rules to improve the model generalizability [10]. With the typical ensemble learning strategies such as bagging, boosting, stacking and decision fusion, ensemble learning makes use of the diversity across different individual models to reduce the variance of the model, which can improve the model's performance on the unseen data.

In the current study, with the data from the National Institute of Health (NIH) LiverTox database [11], the current study developed a model to filter the DILI literature from irrelevant literature based on the title and abstract of publications with multiple text vectorization algorithms in NLP. This study also leveraged the ensemble learning strategy by building a logistic regression meta learner on the top of the predicted probability of different separate learners (also known as the decision fusion strategy) and compared the performance of the ensemble learning model with that of individual models. The model showed high classification performance and interpretable results. Finally, we quantified the prediction reliability with the conformal prediction framework.

II. Methods

The dataset comprises approximately 14,000 DILI-related papers ('positive samples') and approximately 14,000 papers irrelevant to DILI ('negative samples'). For the contest released by the Annual International Conference on Critical Assessment of Massive Data Analysis (CAMDA 2021), only 50% of the positive samples (7,177) and negative samples (7,026) was released while the remaining samples were held out for model assessment. For the hold-out test data, which was referred to as the hold-out test dataset 1, there are 14,211 samples with labels masked to test the model performance on unseen data. For each sample, there are the publication titles and/or the abstracts. The challenge released an additional hold-out test data set with 2,000 abstracts, which was referred to as the hold-out test dataset 2.

A. Data Pre-Processing

The published data (excluding the hold-out datasets) was partitioned into 80% training and 20% internal validation data. Furthermore, pre-processing was carried out on the free text by lowercasing, removing punctuation, numeric, special characters, multiple white spaces, stop words, and finally tokenizing the text with Gensim library on Python 3.7 [12]. Stemming was also performed by changing the terms into their word stems (e.g. "hepatotoxicity" to "hepatotox") to reduce the number of distinct terms for sake of avoiding model overfitting. Here, stemming was regarded as a model hyperparameter tuned based on the performance of the five-fold cross-validation on the training set.

B. Text Vectorization

To extract features from the free-text literature, several different text vectorization algorithms were used to transfer the text into numerical features (i.e. word/sentence vectors): Bag-of-words (BOW), term frequency-inverse document frequency (TF-IDF), word2vec (W2V), and sent2vec (S2V).

Both BOW [13] and TF-IDF [14] are based on word counts. They are among the simple word vectorization algorithms which are widely used to classify text. As a basic word vectorization approach, BOW enumerates the number of each term's occurrences in a piece of text and uses the number of occurrences of each term as the feature. Since the BOW relies on the counts of all terms, the dimensionality of the features from the extracted text equals the number of all different terms in the training corpus. Based on the basic features extracted based on BOW, TF-IDF further regularized the features by calculating the ratio of term frequency (TF), which denotes the number of term occurrences, and inverse document frequency (IDF), which denotes the number of text samples that contain this term. As a result, the value of a feature for a sample increases as the number of occurrences in the text increases but decreases as the total number of texts that include the term increases. With the regularization, the TF-IDF algorithms emphasize the rare terms over the entire training corpus. After applying the BOW and TF-IDF word vectorization algorithms, the feature dimensions were 30,753 and 42,452 with/without stemming.

Word2vec (W2V) [15] is a neural-network-based vectorization algorithm. Without directly relying on the number of occurrences of distinct terms, W2V tries to create an embedding matrix E of the term embeddings for all the terms that occurred in the training corpus. Then, W2V optimizes the embedding matrix E and finally when applying the embedding matrix for downstream tasks, W2V maps the terms to their associated embedding vectors in the embedding matrix. The embedding matrix is the goal of optimization in the W2V training process. The randomly initialized embedding matrix is learned with shallow neural networks based on simple prediction tasks, such as the continuous bag-of-words (CBOW) and the continuous skip-gram. In this study, two pre-trained biomedical W2V models were used: one was trained on a corpus from Wikipedia, PubMed, and PMC (W2V1) [16], and the other was trained on a corpus from PubMed and MIMIC-III (W2V2) [17]. Both models include 16,545,452 terms with an embedding dimension of 200. After converting each term in a text into a 200-dimension embedding, an average of all the term embeddings was taken as the embedding for a text.

S2V is another unsupervised sentence embedding algorithm that allows researchers to compose sentence embeddings using word vectors along with n-gram embeddings [18]. It simultaneously trains the composition and the embedding vectors. S2V can be regarded as an extension of the word contexts from CBOW to a larger sentence context, while the sentence words are specifically optimized via an unsupervised objective function. In the current study, a biomedical S2V model trained on PubMed and MIMIC-III corpora with a dimensionality of 700 for a text was examined [19].

C. Classification Model Development and Assessment

To develop the DILI literature classification model, two protocols were adopted: 1) non-ensemble learning, where classifiers were based on 80% of the training data and the four different vectorization algorithms; 2) ensemble learning, where the 80% training data was further partitioned into 60% data for training separate learners and 20% data for training and hyperparameter tuning for the meta-learner. The separate learners output the predicted probabilities on the samples for meta-learner training while the meta-learner aggregates the predicted probabilities in a logistic regression model. The rationale for developing ensemble learning models is because the different embedding algorithms contain much diversity as they adopted different frameworks to compute the word vectors and document vectors and therefore the potential of better classification performance was tested with the ensemble learning model [10]. The diversity of the separate learners may boost the performance of document classification via the fusion of predicted probabilities given by different models. It is worth noting that, in the ensemble learning protocol, the BOW model was discarded because it generally performs worse when compared with TF-IDF while being similar to TF-IDF as word-count-based algorithms [20]. Additionally, three different weights were added to the positive/negative classes for the TF-IDF/W2V/S2V models in the ensemble learning, to add divergence and focus more on the positive cases, because in a real-world application setting, considering the broad range of research fields archived in PubMed, the positive case prevalence is likely to be much lower than that in the training data.

In the development of separate learners and in the non-ensemble learning protocol, logistic regression (LR) and random forest (RF) models where benchmarked as the classification algorithms. These two algorithms were applied because of their interpretability of important features in the decision-making process. In the meta-learner training, LR was used to reduce the variance and avoid overfitting caused by more flexible classifiers such as RF. The hyperparameters including the strength of L2 penalty, different class weights for LR, the number of estimators, the number of maximum splits for RF, were fine-tuned via five-fold grid search cross-validation on the training data, with classification accuracy as the optimization goal. Similarly, the hyperparameters for the meta-learner were optimized on the 20% separated training set with a five-fold cross-validation.

To evaluate the model performance, on the 20% internal validation data partitioned from the released dataset (for which the labels are known), the classification accuracy, the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and the F1 score were calculated. AUROC is the area under the curve in which the x-axis denotes the false positive rate (FPR) and the y-axis denotes the true positive rate (TPR), while AUPRC is the area under the curve in which the x-axis denotes the recall and the y-axis denotes the precision. AUPRC and AUROC are both commonly used metrics in bioinformatics and cheminformatics studies [21], [22]. The reasons why both AUROC and AUPRC were considered in evaluating our models are: firstly, AUROC has been a widely used metric in evaluating binary classifiers without reliance on the decision threshold set on predicted class probability; secondly, AUPRC was also used because it is more sensitive to the prevalence and can better reflect model performance in an imbalanced data set [23]. Therefore, AUPRC and F1-score were particularly emphasized as in real-world applications, the prevalence of the DILI-positive cases can be much lower and the AUPRC and F1 score can be more comprehensive in evaluating model performance with different weights on positive/negative cases. For the hold-out validation dataset 1 and dataset 2, the accuracy, F1 score, precision, and recall were calculated to validate the model performance on unseen data after the model predictions are submitted.

D. Model Robustness Test With Bootstrapping

To test the reproducibility and robustness of the model performance and the statistical significance in the model comparison, 100 bootstrapping experiments were performed on the 80% training data for the BOW/TF-IDF/W2V/S2V model and on the 60% training data for the ensemble learning model and the variation of the model performance metrics (AUROC, AUPRC, accuracy and F1 score) was tested. To compare the computational cost of each method, the vectorization time, modeling time (for the classification algorithm) and the time to load the pre-trained models were also recorded. Additionally, to test whether the text-vectorization-based approaches perform better than the recently developed transformer-based large pre-trained language model, we tested the performance of the BioBERT (Biomedical Bidirectional Encoder Representations from Transformers) developed by Lee *et al.* [24]. The latest version (BioBERT-Base v1.2) trained with one million PubMed documents was used to extract the text embeddings from the second last layer output and then the prediction of DILI was done with the logistic regression classifier on the text embeddings. The mean values and the 95% confidence interval of the metrics were reported in the result section. Furthermore, paired t-tests were used to test the statistical significance among the metrics given by different models.

E. Prediction Reliability Analysis Using Conformal Prediction

To quantify the uncertainty of the predictions, the conformal prediction framework was implemented. Conformal prediction, developed by Vladimir Vovk, assumes the data is generated from an independent and identical distribution and calculates the credibility and confidence level in the model inference from a statistics aspect [25]. Simplified descriptions of conformal prediction enough for applications can be found in the previous articles [26], [27]. The conformal prediction framework relies on the statistical significance on the nonconformity measurement: a metric to quantify how well a new prediction conform to the existing training data. Briefly, the nonconformity measurement α_i was firstly calculated with respect to each prediction label within the training set by an previously proposed nonconformity calculation function (details shown in [27]):

$$\alpha_i = 0.5 - \frac{\hat{p}(y_i|x_i) - \mathrm{max}\hat{\mathrm{p}}_{\mathrm{y!=y_i}}(\mathrm{y_i}|\mathrm{x_i})}{2} \tag{1}$$

where y_i and x_i is the label and features, respectively. In our case of binary classification with LR as the classifier, the equation can be simplified as:

$$\alpha_i = \begin{cases} \hat{p}(y_i|x_i) & y_i = 0 \\ 1 - \hat{p}(y_i|x_i) & y_i = 1 \end{cases} \tag{2}$$

where the $\hat{p}(y_i|x_i)$ is prediction probability given by LR. At the inference stage, the α^* for the new sample x^* with regards to negative/positive label is also calculated with (2). Then the P-value of the prediction is calculated by:

$$p^{*,y} = \frac{|\{i = 1, \ldots, n | \alpha_i^{*,y} \leq \alpha_i^n\}|}{n} \tag{3}$$

where $p^{*,y}$ is the P-value of the assumed label y (0/1) for the new sample x^*. The credibility of prediction is defined as the larger P-value [26]. In this study, to quantify the prediction uncertainty, the credibility of each predictions made for the validation samples were analyzed over the 100 times of bootstrapping. The mean, median, 95% confidence interval, and quartiles of these credibility values for correct and incorrect predictions were reported for the TF-IDF and the ensemble model.

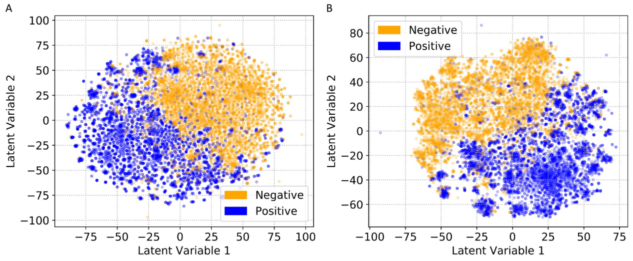
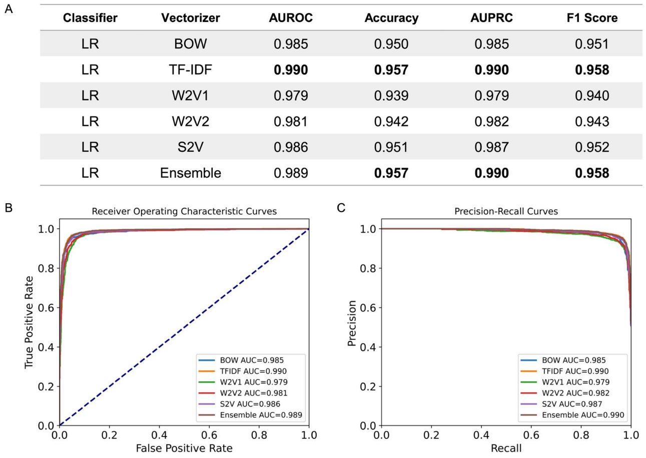


Fig. 1. The t-SNE visualization of the text vectors of the training data. (A) TF-IDF text vector visualization; (B) S2V text vector visualization.



Classifier	Vectorizer	AUROC	Accuracy	AUPRC	F1 Score
LR	BOW	0.985	0.950	0.985	0.951
LR	TF-IDF	**0.990**	**0.957**	**0.990**	**0.958**
LR	W2V1	0.979	0.939	0.979	0.940
LR	W2V2	0.981	0.942	0.982	0.943
LR	S2V	0.986	0.951	0.987	0.952
LR	Ensemble	0.989	**0.957**	**0.990**	**0.958**

Fig. 2. The classification performance of the models on the internal validation data. (A) The table of classification performance metrics of different vectorization models. The receiver operating characteristic curves (B), and the precision-recall curves (C) of the different models.

III. Results

Before experiments, the data were firstly visualized to give a better understanding of the data distribution. First, text vectors given by TF-IDF and S2V were visualized as examples, with the unsupervised non-linear dimensionality reduction method: t-distributed stochastic neighbour embedding (t-SNE), which has been shown effective in visualizing high-dimensional data [20], [28]. The results show that the positive samples and the negative samples cluster separately, indicating the potential feasibility of classifying the DILI-positive samples (Fig. 1). It should be noted that the t-SNE visualization is completely unsupervised and the clusters shown are labelled with the ground truth labels from the dataset.

A. Model Performance on Validation Data

With text vectors, classifiers based on each of the four vectorization algorithms were built. After performing hyperparameter tuning in the five-fold cross validation on the training data, the best strength of L2 penalty were 10 for the BOW model, 0.1 for the TF-IDF, W2V1 and W2V2 models, and 1 for the S2V model. Word stemming was used for BOW and TF-IDF models but not in the W2V1, W2V2 and S2V models. The performance on the validation data is shown in Fig. 1. The results show that besides the ensemble learning model, TF-IDF outperformed the other models with the highest AUROC (0.990), accuracy (0.957), AUPRC (0.990), and F1-score (0.958) (Fig. 2). The RF models did not outperform the LR models and were therefore not shown and used in our ensemble models.

Additionally, after plotting the confusion matrices, it can be seen that among the separate learners, the TF-IDF model has the fewest false-positive cases while the S2V model has the fewest false-negative cases (Fig. 3). Venn plots of the false predictions were shown in Fig. 3(A). According to the results, although most false-positive and false-negative cases overlap across different word-vectorization models, there is divergence

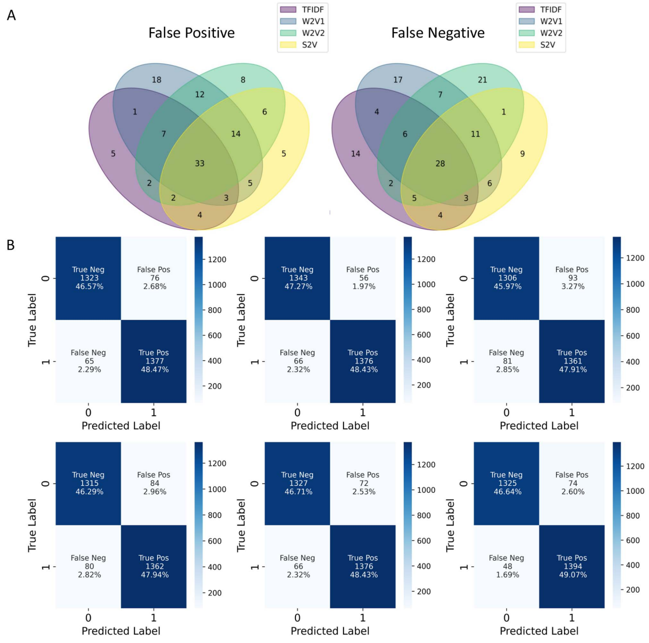


Fig. 3. The confusion matrices and the numbers of false-positive cases and false-negative cases of different models on the validation data. (A) The Venn plot of the false-positive cases and false-negative cases given by four separate learners. (B) The confusion matrices of the six different algorithms, top, from left to right: BOW, TF-IDF, W2V1, bottom: from left to right, W2V2, S2V, and ensemble learning models.

among different models, which motivates the ensemble learning protocol which aggregates the diverse knowledge of diverse learners for potentially better model performance. In addition, Cohen's Kappa values were calculated for quantitative analysis of model differences before developing ensemble models [29], from which a similar conclusion can be interpreted as the Venn plot.

The classifiers based on TF-IDF, two W2V models, and S2V were then assembled. The results show that it has reached the same highest AUPRC, accuracy, and F1-score as the TF-IDF model, with a similar AUROC (Fig. 2). According to the confusion matrices (Fig. 3), the ensemble model shows the fewest false-negative cases and a decent number of false-positive cases. Based on the hyperparameter tuning processes, the hyperparameters of the final ensemble model have the strength of L2 penalty at 10 and the class weight at 1:1 (tuned based on five-fold cross-validation on the 20% data used to train the meta learner). The hyperparameters of the 12 separate LR models leveraged in the ensemble models (tuned based on five-fold cross-validation on the 60% data used to train separate learner): strength of L2 penalty: 0.1 for the nine TF-IDF/W2V1/W2V2 models, 1 for three S2V models.

B. Model Interpretation

On the internal validation data, the best-performing models of TF-IDF and ensemble learning were interpreted. For the TF-IDF model, the training data was bootstrapped 2000 times, extracted the coefficients of LR, and averaged the coefficients which correspond to each term. Then, the mean coefficients were ranked and the top 10 most important words for the positive prediction and negative prediction were selected and shown respectively. The results show the important words in the stemmed version: the important words for the positive prediction such as "safety", "hepatotoxic/hepatotoxicity", and "liver" show clear meaning related to DILI, illustrating model interpretability (Fig. 4(A)). Additionally, the contribution of different vectorization models in the ensemble learning was visualized (Fig. 4(B)). The results show that all coefficients are positive indicating the positive contributions of different text vectorization models. Based on these results, it is shown that among these models, TF-IDF models and S2V models perform the best.

C. Model Robustness Test With Bootstrapping

According to the protocol introduced in Section 2.D, 100 times of bootstrapping resampling were done on the training set and the model performance fluctuations were tested, with the results shown in Fig. 5. The statistics and the boxplots show that the ensemble learning model outperformed the other models on all four classification model performance metrics with statistical significance ($p \leq 0.001$, paired t-tests), while the TF-IDF model generally ranked the second outperforming all models other than the ensemble learning model ($p \leq 0.001$). However, since the computation of the ensemble learning model requires the TF-IDF, W2V1, W2V2 and S2V models, the TF-IDF models significantly outperformed the ensemble learning model in the computational cost. Additionally, according to the results shown in Fig. 5, the BioBERT performance was generally inferior to the BOW, TF-IDF, S2V and ensemble learning models.

D. Prediction Reliability Analysis Using Conformal Prediction

The credibility values, calculated according to the conformal prediction framework, were shown in Fig. 6(A). Wrong predictions (false negative (FN) and false positive (FP) predictions) have relatively low credibility around 0.5 and correct predictions have a much higher mean credibility around 0.74. A clear trend was shown in Fig. 6(B) that a better prediction accuracy went along with a higher prediction credibility. A few example texts were given in Table II for wrong and correct predictions with relatively high (>0.99) and low credibilities (<0.55).

E. Model Performance on Hold-Out Data and Additional Hold-Out Test Data

After testing the models on the internal validation data, two best models: the TF-IDF model and the ensemble model were chosen and tested on the hold-out test data. The performance on the hold-out test data is shown in Table III and it should be noted that the labels on the hold-out datasets are completely blinded and only the performance metrics can be given upon the submission of predictions. On the hold-out test dataset 1, both models reach the same accuracy of 0.954, while the ensemble model reaches a higher precision of 0.960, and the TF-IDF model reaches a higher recall of 0.961. On the additional hold-out test data (hold-out test dataset 2) with abstract only, the TF-IDF

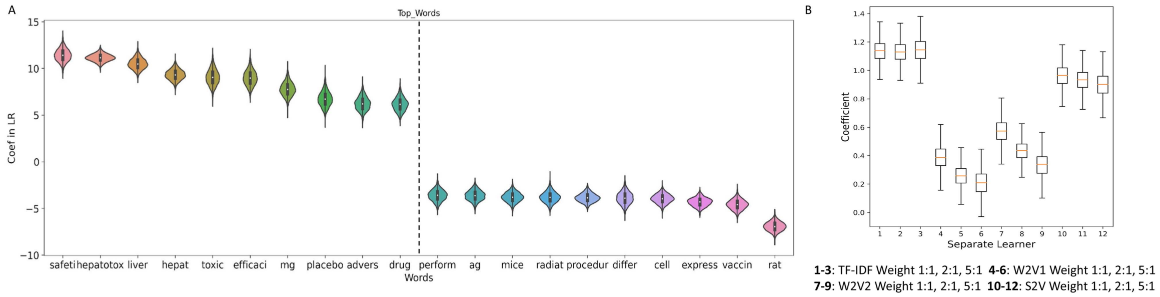
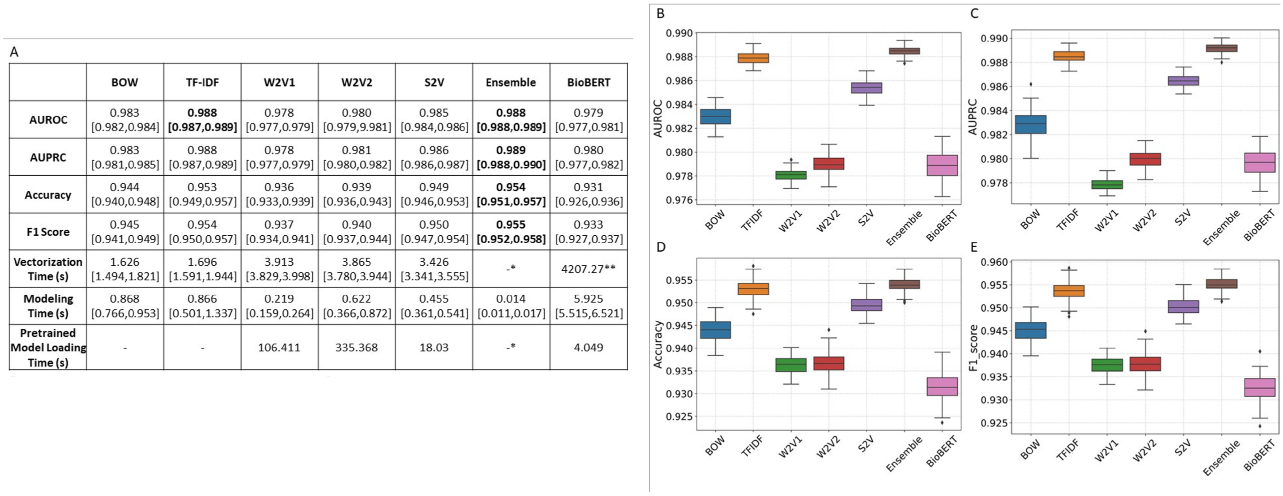


Fig. 4. Interpretation of the important words for classification based on TF-IDF and the contribution of separate learners in the ensemble learning model. (A) The top 10 most important words for positive predictions and negative predictions with the distribution of logistic regression coefficients in 2,000 bootstrapping experiments. (B) The normalized logistic regression coefficients of separate learners in 2,000 bootstrapping experiments.


A

	BOW	TF-IDF	W2V1	W2V2	S2V	Ensemble	BioBERT
AUROC	0.983 [0.982,0.984]	**0.988** **[0.987,0.989]**	0.978 [0.977,0.979]	0.980 [0.979,9.981]	0.985 [0.984,0.986]	**0.988** **[0.988,0.989]**	0.979 [0.977,0.981]
AUPRC	0.983 [0.981,0.985]	0.988 [0.987,0.989]	0.978 [0.977,0.979]	0.981 [0.980,0.982]	0.986 [0.986,0.987]	**0.989** **[0.988,0.990]**	0.980 [0.977,0.982]
Accuracy	0.944 [0.940,0.948]	0.953 [0.949,0.957]	0.936 [0.933,0.939]	0.939 [0.936,0.943]	0.949 [0.946,0.953]	**0.954** **[0.951,0.957]**	0.931 [0.926,0.936]
F1 Score	0.945 [0.941,0.949]	0.954 [0.950,0.957]	0.937 [0.934,0.941]	0.940 [0.937,0.944]	0.950 [0.947,0.954]	**0.955** **[0.952,0.958]**	0.933 [0.927,0.937]
Vectorization Time (s)	1.626 [1.494,1.821]	1.696 [1.591,1.944]	3.913 [3.829,3.998]	3.865 [3.780,3.944]	3.426 [3.341,3.555]	-*	4207.27**
Modeling Time (s)	0.868 [0.766,0.953]	0.866 [0.501,1.337]	0.219 [0.159,0.264]	0.622 [0.366,0.872]	0.455 [0.361,0.541]	0.014 [0.011,0.017]	5.925 [5.515,6.521]
Pretrained Model Loading Time (s)	-	-	106.411	335.368	18.03	-*	4.049

Fig. 5. Model robustness test and classification performance variation in 100 bootstrapping experiments. (A) The variation of the metrics of model performance. Here, mean values and 95% confidence interval values were reported. The best performing model according to each classification performance metric was marked in bold. (*: The time of ensemble learning approach depends on the time of separate learners including TF-IDF, W2V1, W2V2 and S2V. **: Due to the lengthy computational time of text vectorization with BioBERT, the time was tested once). The variation of model (B) AUROC, (C) AUPRC, (D) accuracy, and (E) F1 score from different models. Machine specs: Google Cloud Platform, 16 N1 vCPUs, 104 GB RAM, no GPU acceleration.

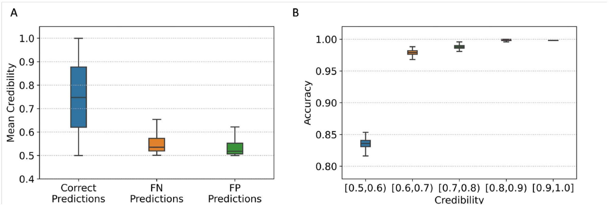


Fig. 6. Ensemble model prediction reliability analysis from 100 bootstrapping experiments. (A) The mean credibility of correct, FN, and FP predictions. (B) The accuracy distribution of predictions made with different credibility.

TABLE I
Pair-Wise Cohen's Kappa Values for Diversity Analysis

	BOW	TF-IDF	W2V1	W2V2	S2V
BOW	1.000	0.918	0.887	0.887	0.912
TF-IDF	0.918	1.000	0.913	0.913	0.930
W2V1	0.887	0.913	1.000	0.927	0.924
W2V2	0.887	0.913	0.927	1.000	0.927
S2V	0.912	0.930	0.924	0.927	1.000

model outperformed the ensemble model with higher accuracy of 0.927, a higher F1 score of 0.930 and a higher precision of 0.886, while the ensemble showed a higher recall of 0.988. Generally, on the two datasets of the hold-out data released by the challenge, both models show high performance in classifying the literature with high accuracy and F1 score, which may enable researchers to accurately filter the DILI-negative literature for further analysis.

IV. Discussion

In this study, to develop a machine-learning-based model to automatically filter drug-induced liver injury (DILI) related publications out of the irrelevant publications, four different natural language processing (NLP) text vectorization methods were used to vectorize scientific publications. Then, logistic regression models and random forest models were built based on the text vectors and ensemble learning, to predict whether the publications are related to DILI or not. The TF-IDF model and the ensemble learning model with logistic regression classifiers

TABLE II
SAMPLES FROM THE ENSEMBLE MODEL WITH HIGH-CREDIBILITY CORRECT PREDICTIONS AND LOW CREDIBILITY WRONG PREDICTIONS

Predicted Label	True Label	Credibility	Title	PMID
1	1	0.996	Two cases of severe clinical and histologic hepatotoxicity associated with troglitazone	9652997
0	0	0.999	Multivalent HA DNA vaccination protects against highly pathogenic H5N1 avian influenza infection in chickens and mice	19293944
1	0	0.509	Identifying hepatitis C virus genotype 2/3 patients who can receive a 16-week abbreviated course of peginterferon alfa-2a (40KD) plus ribavirin	20196118
0	1	0.542	Mitochondrial carbonic anhydrase VA deficiency resulting from CA5A alterations presents with hyperammonemia in early childhood	24530203

TABLE III
THE PERFORMANCE OF THE TF-IDF MODEL AND THE ENSEMBLE MODEL ON THE HOLD-OUT DATA (1) AND ADDITIONAL HOLD-OUT DATA (2)

Model	Accuracy	F1 Score	Precision	Recall
TF-IDF-1	0.954	0.954	0.947	0.961
Ensemble-1	0.954	0.955	0.960	0.950
TF-IDF-2	0.927	0.930	0.886	0.979
Ensemble-2	0.900	0.908	0.840	0.988

reached the highest classification performance in terms of accuracy, AUPRC, and F1 score on the internal validation set. Both models show high classification performance on the hold-out data. Additionally, the TF-IDF model is also interpretable with the important words for making positive predictions showing meanings clearly related to DILI, and important words for making negative predictions not directly related to DILI. As DILI, which may cause acute liver failure and even death, has become the major killer of prospective new drug candidates and most DILI research results are in the free-text format in scientific publications, the models would enable researchers to fast filter the DILI literature without time-consuming manual work.

The reason for using an ensemble model is that the different text vectorizations are believed to capture different details in the sentence. The concept is validated by the Venn plot in Fig. 3(A) which showed overlapping false negatives and false positives of predictions given by models built on various text vectorizations. On top of these overlapping wrong predictions, different text vectorizations also owned their distinctive false positives and false negatives. By using a logistic regression meta-learner in the ensemble learning process to leverage the predicted probability from different separate learners (logistic regression models with different vectorization techniques), this ensemble model could potentially compensate for limitations and boost the strength of different text vectorizations in this task. And this is evidenced by its better prediction results on the hold-out data (Table I). More importantly, on the internal validation data, the ensemble learning strategy yields a lower false-negative rate (reduced to 1.69%, while the individual word-vectorization model can only reach 2.29%). The fewer false negative predictions better help prevent missing DILI-information for researchers, which can be critical for the drug candidates.

Recently, there has been research pointing out the peeking problem of ensemble learning models where the model performance may be overestimated through multiple submissions of the predicted results in machine learning challenges [30]. However, in this study, the protocol for developing the ensemble learning models and non-ensemble learning models prevents the peeking problem from happening. Because the hyperparameter tuning of the ensemble learning model was done during five-fold cross-validation on the 20% data used for the training of the meta learner while the hyperparameters for the non-ensemble learning models were tuned based on the five-fold cross-validation on the 80% training data, the model was completely developed before performance evaluation was performed and therefore the performance metrics on the internal validation set, hold-out validation sets were not used to feedback the model training and hyperparameter tuning process. Additionally, to prevent any model fine-tuning in the model performance evaluation process, the final results on the holdout validation sets were all based on one submission of the ensemble learning model and one submission of the TF-IDF model.

When compared with the models developed by other candidates in the CAMDA challenge, the models developed in this study outperformed the other models reported in the CAMDA challenge presentation at the 29th Conference on Intelligent Systems for Molecular Biology and the 20th European Conference on Computational Biology on the same hold-out validation dataset [31]. For example, Katritsis *et al.* developed a text-mining approach for the identification of DILI-related literature named dialogi with linear classifier, and reported an accuracy of 94.1%, a recall of 94.9% and a precision of 93.3% on the external validation set; Liu *et al.* developed an AI-based DILI literature classifier named DILIc and reported an accuracy of 94.14% on the same validation set. However, the best model performance in this study on the same external validation set are: accuracy: 95.4% (TF-IDF model and ensemble learning model), precision: 96.0% (ensemble learning model) and 94.7% (TF-IDF model), recall: 96.1% (TF-IDF model) and 95.0% (ensemble learning model), which outperformed the other candidates in the contest on the same validation dataset.

Conformal prediction enables the model to output not only the predicted labels but also the reliability of the predictions [25], [32]. This can effectively provide users with a failure protection tool by referring to the prediction credibility. In actual applications, the predictions made with high credibility should be fairly accurate and thus trustworthy. For example, according to the results, predictions made by the classifier with credibility higher than 0.60 can be generally accepted by the users. For lower credibility predictions, researchers are recommended to double check the article manually to avoid missing any FN.

Examples of low credibility predictions were shown in Table II. For the FP case, the title contains “hepatitis C virus”. The word “hepatitis” may account for the reason why the model wrongly classified it as a DILI-positive paper. For the FN sample listed, we cannot see any liver-related description except for the term "hyperammonemia". Although acquired hyperammonemia is known to be caused by acute liver failure [33], the title itself seems to be not discussing the drug-induced liver failure but more related to a gene deficiency. The credibility of this prediction is low: 0.542, which indicates that the uncertainty of this prediction is high. In this case, the researcher can read into the article to dig out what has been reported.

Next, a comparison between the conventional text-embedding-based approaches and a transformer-based large pre-trained language model is also worthy of discussion. In this study, the results show that compared with the pre-trained biomedical BERT [24], the simple word-vectorization-based approaches such as the BOW and TF-IDF have statistically significantly better AUROC, AUPRC, accuracy and F1 score. The BioBERT's performance was also not better than the pre-trained word-embedding and sentence-embedding models. The results correspond well with many previous research papers showing that transformer-based models did not show significant improvement over conventional models. For example, Wieting *et al.* [34] showed that complex neural-network-based methods are outperformed by simpler methods with basic linear regression, while Arora *et al.* [35] further validated this by showing that the unsupervised sentence embedding can be a formidable baseline for complicated models. Furthermore, the computational cost of the large pre-trained language models is high: the vectorization of the training and validation corpus took over 4,000 seconds in this study, which can lead the model efficiency to significantly deteriorate with no gain in the better classification performance. Additionally, transformer-based and recurrent-neural-network-based models are not easily interpretable when compared with the conventional word-vector-based method with logistic regression. As a result, the conventional NLP approaches are recommended for the task of filtering DILI literature for their satisfactory performances and low computational cost.

As for the application of this study, firstly, the models could be applied to filter DILI literature for drug discovery researchers from the large corpus of publications and monitor the latest research on DILI, as assistant systems for information retrieval. The scenario of this application is that the researchers can collect the titles and abstracts of the newly published biomedical papers and then apply the DILI-filter models developed in this study. As there are tens of thousands of new publications everyday, the models can enable the researchers to automatically fast filter out the papers unrelated to DILI and this can improve the efficiency of paper reading for the researchers. Albeit concerns about ensemble learning were raised for a longer runtime and being more computationally demanding, the application scenarios of our model are less time- and resource-critical. The time consumed (Fig. 5) for vectorization and model inference for our ensemble model is still negligible when compared with human performances. And thus model inference can be done everyday to filter out related articles. Secondly, these literature-filtering models can lay a foundation for future quantitative structure-property relationship (QSPR) modeling in drug discovery [36] and drug development [37] because the systems can expedite the DILI-labeling process for different drug candidates effectively. Furthermore, this paradigm of literature-filtering system development can be expanded to other fields of biomedical research well.

Although models developed in this study show good classification performance, there are limitations that can be further addressed. For example, the training corpus is relatively small for the model training when compared with the majority of natural language processing applications. Here, the pre-trained biomedical word/sentence embedding models such as S2V and W2V were leveraged to address the issue of limited training data. In the future, a larger training corpus can be made with the addition of multiple additional biomedical text corpora (which may be referred to as unlabelled data for semi-supervised learning or data augmentation) to further improve the model performances on real-world applications. The supplementary data are not necessarily directly related to the DILI topic, but this unbalanced training set better represents the actual application scenario where the majority of the publications on PubMed are irrelevant to DILI.

V. Conclusion

Several models were developed to filter DILI-related literature based on four text vectorization techniques (bag-of-words, TF-IDF, two biomedical word2vec models, biomedical sent2vec model) and ensemble learning. The model with TF-IDF and LR outperformed others with an AUROC of 0.990, an accuracy of 0.957, and an AUPRC of 0.990. An ensemble learning model with overall similar performance but the fewest false-negative cases was developed based on the prediction probability from 12 individual word-vectorization models, which shows the highest accuracy (0.954) and F1 score on the hold-out data (0.955). Both models performed well on the two hold-out test data with each model taking a lead in different classification metrics (the ensemble learning model accuracy: 0.954, F1-score: 0.955, precision: 0.960, recall: 0.950). Moreover, conformal prediction was implemented to obtain the reliability of filtering results, serving as a evaluation tool for researchers to further avoid FN predictions. The development of both TF-IDF and ensemble models enables the users to apply these two models for applications and the ensemble-learning-based model enables a more efficient literature filter with fewer false negatives for researchers who focus on the DILI in the field of drug discovery.

VI. Code Availability

The code and models can be found at: https://github.com/xzhan96-stf/dili-filter

Acknowledgment

The authors thank the support from Department of Bioengineering and Department of Biomedical Data Science, Stanford University.

References

[1] W. M. Lee, "Drug-induced hepatotoxicity," *New England J. Med.*, vol. 349, no. 5, pp. 474–485, Jul. 2003. [Online]. Available: https://doi.org/10.1056/NEJMra021844

[2] M. Chen, J. Borlak, and W. Tong, “High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury,” *Hepatology*, Baltimore, MD, USA, vol. 58, no. 1, pp. 388–396, Jul. 2013.

[3] R. J. Andrade et al., “Drug-induced liver injury,” *Nature Rev. Dis. Primers*, vol. 5, no. 1, pp. 1–22, Aug. 2019. [Online]. Available: https://www.nature.com/articles/s41572-019-0105-0

[4] N. Kaplowitz, “Drug-induced liver injury,” *Clin. Infect. Dis.*, vol. 38, no. Supplement_2, pp. S44–S48, Mar. 2004. [Online]. Available: https://doi.org/10.1086/381446

[5] M. C. Donnelly, J. S. Davidson, K. Martin, A. Baird, P. C. Hayes, and K. J. Simpson, “Acute liver failure in Scotland: Changes in aetiology and outcomes over time (the Scottish look-back study),” *Alimentary Pharmacol. Therapeutics*, vol. 45, no. 6, pp. 833–843, Mar. 2017.

[6] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] Y. Wang, M. Rastegar-Mojarad, R. Komandur-Elayavilli, S. Liu, and H. Liu, “An ensemble model of clinical information extraction and information retrieval for clinical decision support,” in *Proc. TREC Conf. Nat. Inst. Standards Technol.*, 2016, pp. 1–10.

[9] X. Zhan, M. Humbert-Droz, P. Mukherjee, and O. Gevaert, “Structuring clinical text with AI: Old versus new natural language processing techniques evaluated on eight common cardiovascular diseases,” *Patterns*, vol. 2, no. 7, Jul. 2021, Art. no. 100289. [Online]. Available: https://www.cell.com/patterns/abstract/S2666-3899(21)001227

[10] M. Ganaie et al., “Ensemble deep learning: A review,” 2021, *arXiv:2104.02395*.

[11] J. H. Hoofnagle, J. Serrano, J. E. Knoben, and V. J. Navarro, “LiverTox,” *Hepatology*, Baltimore, Md., vol. 57, no. 3, pp. 873–874, Mar. 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5044298/

[12] R. Řehůřek and P. Sojka, “Software framework for topic modelling with large corpora,” in *Proc. LREC Workshop New Challenges NLP Frameworks*, Valletta, Malta, ELRA, 2010, pp. 45–50.

[13] Z. S. Harris, “Distributional Structure,” *WORD*, vol. 10, no. 2–3, pp. 146–162, Aug. 1954. [Online]. Available: https://doi.org/10.1080/00437956.1954.11659520

[14] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *J. Documentation*, vol. 28, no. 1, pp. 11–21, Jan. 1972. [Online]. Available: https://doi.org/10.1108/eb026526

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2013, vol. 26, pp. 1–9.

[16] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, “Distributional semantics resources for biomedical text processing,” in *Proc. LBM*, 2013, pp. 39–44.

[17] Y. Zhang, Q. Chen, Z. Yang, H. Lin, and Z. Lu, “BioWordVec, improving biomedical word embeddings with subword information and MeSH,” *Sci. Data*, vol. 6, no. 1, pp. 1–9, May 2019. [Online]. Available: https://www.nature.com/articles/s41597-019-0055-0

[18] M. Pagliardini, P. Gupta, and M. Jaggi, “Unsupervised learning of sentence embeddings using compositional n-gram features,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., Volume 1 (Long Papers)*, 2018, pp. 528–540. [Online]. Available: http://arxiv.org/abs/1703.02507

[19] Q. Chen, Y. Peng, and Z. Lu, “BioSentVec: Creating sentence embeddings for biomedical texts,” in *Proc. IEEE Int. Conf. Healthcare Informat.*, 2019, pp. 1–5.

[20] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[21] N.-Q.-K. Le and B. P. Nguyen, “Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2189–2197, Nov./Dec. 2021.

[22] T. N. K. Hung et al., “An AI-based prediction model for drug-drug interactions in osteoporosis and paget’s diseases from SMILES,” *Mol. Informat.*, Jan. 2022, Art. no. e2100264.

[23] J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” in *Proc. 23rd Int. Conf. Mach. Learn.*, New York, NY, USA, Association for Computing Machinery, Jun. 2006, pp. 233–240. [Online]. Available:https://doi.org/10.1145/1143844.1143874

[24] J. Lee et al., “BioBERT: A pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[25] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learn. in a Random World*. New York, NY, USA: Springer, 2005. [Online]. Available: http://link.springer.com/10.1007/b106715

[26] X. Zhan, Z. Wang, M. Yang, Z. Luo, Y. Wang, and G. Li, “An electronic nose-based assistive diagnostic prototype for lung cancer detection with conformal prediction,” *Measurement*, vol. 158, Jul. 2020, Art. no. 107588. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0263224120301251

[27] L. Liu et al., “CPSC: Conformal prediction with shrunken centroids for efficient prediction reliability quantification and data augmentation, a case in alternative herbal medicine classification with electronic nose,” *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 4001211.

[28] L. Liu et al., “Boost AI power: Data augmentation strategies with unlabelled data and conformal prediction, a case in alternative herbal medicine discrimination with electronic nose,” *IEEE Sensors J. 21.20*, pp. 22995–23005, 2021.

[29] T. Toprak, B. Belenlioglu, B. Aydin, C. Guzelis, and M. A. Selver, “Conditional weighted ensemble of transferred models for camera based onboard pedestrian detection in railway driver support systems,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5041–5054, May 2020.

[30] A. E. Kavur, L. I. Kuncheva, and M. A. Selver, “Basic ensembles of vanilla-style deep learning models improve liver segmentation from ct images,” 2020, *arXiv:2001.09647*.

[31] “Camda poster presentation,” [Online]. Available: https://www.iscb.org/cms_addon/conferences/ismbeccb2021/posters.php?track=CAMDA&session=E#search, accessed: 2021–07-31.

[32] X. Zhan et al., “Online conformal prediction for classifying different types of herbal medicines with electronic nose,” in *Proc. IET Doctoral Forum Biomed. Eng., Healthcare, Robot. Artif. Intell.*, 2018, pp. 1–8.

[33] S. Matoori and J.-C. Leroux, “Recent advances in the treatment of hyperammonemia,” *Adv. Drug Del. Rev.*, vol. 90, pp. 55–68, Aug. 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169409X15000654

[34] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, “Towards universal paraphrastic sentence embeddings,” 2015, *arXiv:1511.08198*.

[35] S. Arora, Y. Liang, and T. Ma, “A simple but tough-to-beat baseline for sentence embeddings,” in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–16.

[36] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, “Machine learning in chemoinformatics and drug discovery,” *Drug Discov. Today*, vol. 23, no. 8, pp. 1538–1546, Aug. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359644617304695

[37] M. Elbadawi, S. Gaisford, and A. W. Basit, “Advanced machine-learning techniques in drug discovery,” *Drug Discov. Today*, vol. 26, no. 3, pp. 769–777, Dec. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1359644620305213