Drug Safety (2022) 45:583–596
https://doi.org/10.1007/s40264-022-01159-2

ORIGINAL RESEARCH ARTICLE

Supervised Machine Learning-Based Decision Support for Signal Validation Classification

Muhammad Imran[1] · Aasia Bhatti[2] · David M. King[3] · Magnus Lerch[4] · Jürgen Dietrich[5] · Guy Doron[6] · Katrin Manlik[7]

Accepted: 13 February 2022

Abstract

Introduction Signal validation in pharmacovigilance is the process of evaluating data to decide whether evidence is sufficient to justify further assessment of a detected signal. During the signal validation process, safety experts in our organization are required to review signals of disproportionate reporting (SDRs) and classify them into one of six predefined categories.
Objective This experiment explored the extent to which predictive machine learning (ML) models can support the decision making of safety experts by accurately identifying the most appropriate predefined signal validation category.
Methods We extracted cumulative data for six medicinal products, consisting of historic SDR validations and Individual Case Safety Reports, from the company's safety database for training and testing of the ML model. We implemented a decision tree-based supervised multiclass classifier model termed Gradient Boosted Trees followed by a SHapley Additive exPlanations (SHAP) analysis to mitigate the "black box" effect of the ensemble model by identifying the key predicting features in the model. Following a retrospective analysis, a prospective experiment was conducted to test the model accuracy and user acceptance in a real-life setting.
Results The prediction accuracy of our ML model ranged from 83 to 86% over 3 months for the six medicinal products. The applicability of the model was confirmed by the company's safety experts. Additionally, the systematic predictions provided valuable information to the safety experts and assisted them in reviewing the SDRs efficiently and consistently.
Conclusions This experiment demonstrated that it is possible to train a multiclass classification model to accurately predict signal validation categories for SDRs. More importantly, the transparency of the predictions provided by the SHAP analysis led to high acceptance by the safety experts.

Key Points

This experiment demonstrated that signal validation in pharmacovigilance can be supported by a machine learning (ML)-based prevalidation step to improve process efficiency and consistency. Medical review by safety experts remains an essential part of the signal validation process, but this can be performed faster and more consistently when augmented by ML predictions.

Model explainability plays a major role in gaining trust and acceptance of ML outputs in pharmacovigilance. SHapley Additive exPlanations (SHAP) analysis was used to improve model explainability.

✉ Muhammad Imran
muhammad.imran5@bayer.com

Extended author information available on the last page of the article

1 Introduction

The goal of signal detection in pharmacovigilance is to detect the existence of new potentially causal associations, or new aspects of known associations, between medicinal products and events [1]. In the quantitative signal detection process, the use of disproportionality methods is a proven and widely used approach to identify signals from spontaneous adverse event reporting databases [2], which are termed signals of disproportionate reporting (SDRs). Filters are applied based on predetermined thresholds, trend flags, and further re-signaling criteria to greatly reduce the number of resulting SDRs. The remaining identified SDRs are reviewed and validated. Signal validation is the process of evaluating the data supporting the detected signals to verify whether evidence is sufficient to justify further analysis [3]. Safety experts evaluate relevant information and classify the validated signal into predefined categories. The signal validation process is complex and labor intensive and may show variability in its decisions because of the nature of this activity, which involves medical judgment that can vary between reviewers and over time.

There is reinforced interest and focus in research for the use of machine learning (ML) and artificial intelligence in a growing number of pharmacovigilance processes [4], including decision support and automation in the processing and reporting of Individual Case Safety Reports (ICSRs) [5–7], identification of adverse events or other medical concepts from spontaneous reports or social media supported by natural language processing [8–10], and adverse event prediction for personalized medicine [11]. Efforts are also increasing within pharmacovigilance research to support the signal detection process using ML approaches [12–14].

In this experiment, we explored the extent to which ML can support safety experts during the signal validation process. On the subject of decision support for signal prioritization, which is closely related to signal validation, we found previous work performed using a multiattribute decision analysis [15].

Our main objective was to test whether ML can reliably predict signal validation classifications and support the decisions of safety experts but not replace the medical review step. If successful, the efficiency and consistency of the currently manual signal validation process could be improved. In addition, we aimed to provide transparency around the ML outputs to achieve a high user acceptance for the ML-based approach.

2 Methods

2.1 Setup of the Experiment

Our experiment was guided by the following flow of activities.

1. We wanted to know whether an ML model could predict SDR validations and how accurate such predictions might be.
2. We used the data that SDRs are based on, i.e., ICSRs from the company's safety database, and the SDRs and their validations contained in the company's signal detection data mart.
3. In the first step (phase I), we used data retrospectively, transformed the data into features for ML, trained different models, tried some variations, compared the performances, and selected the most promising model.
4. In a second step (phase II), we applied the most promising model prospectively to new data, presented the predictions to safety experts, asked them whether the predictions and their presentation were helpful, and calculated the accuracy.
5. Finally, we reviewed what we learned and decided to share it in this publication.

2.2 Data Sources and Data Selection

2.2.1 Individual Case Safety Reports and Signals of Disproportionate Reporting (SDRs)

We used two data sources for our experiment: (1) the safety database containing ICSRs ("cases") and (2) the signal detection data mart containing SDRs and their validations. In the quantitative signal detection process, ICSRs, each containing one or multiple product–event combinations (PECs), are transferred from the safety database into the signal detection data mart and get aggregated by PEC, i.e., by counting the number of ICSRs for each PEC. The proportional reporting ratio (PRR) is run each month as a disproportionality method to identify which of the PECs meet the criteria and thresholds for an SDR. SDR criteria are defined for first-time SDRs with number of cases (N) ≥ 3 and PRR ≥ 2 and chi-squared (with Yates correction) ≥ 4 [16] and for re-signals in addition with a frequency increase ≥ 50% compared with the frequency at the latest prior validation [17].

The same two data sources were used in phase I and II of our experiment, with only the data selection criteria differing.

- Phase I (retrospective experiment conducted in September 2020)

For three medicinal products:

o Cumulative case data up to 31 August 2020.

p SDR data and their validations originating from monthly signal runs performed from August 2014 to September 2020, with a stratified split of 70% training and 30% test data.

q The phase I dataset contained 582,132 PEC records from ICSRs and 2105 SDRs and their validations from the signal detection data mart.

r Phase II (prospective experiment conducted in February, March, and April 2021)

For six medicinal products (including the three from phase I):

o Cumulative case data up to 31 January 2021, 28 February 2021, and 31 March 2021, respectively.

p SDR data and their validations originating from monthly signal runs performed from August 2014 to January 2021, plus SDRs for the subsequent month of the experiment—February, March, and April 2021—used for validation predictions. *Note:* SDR validations performed by safety experts for February and March 2021 were

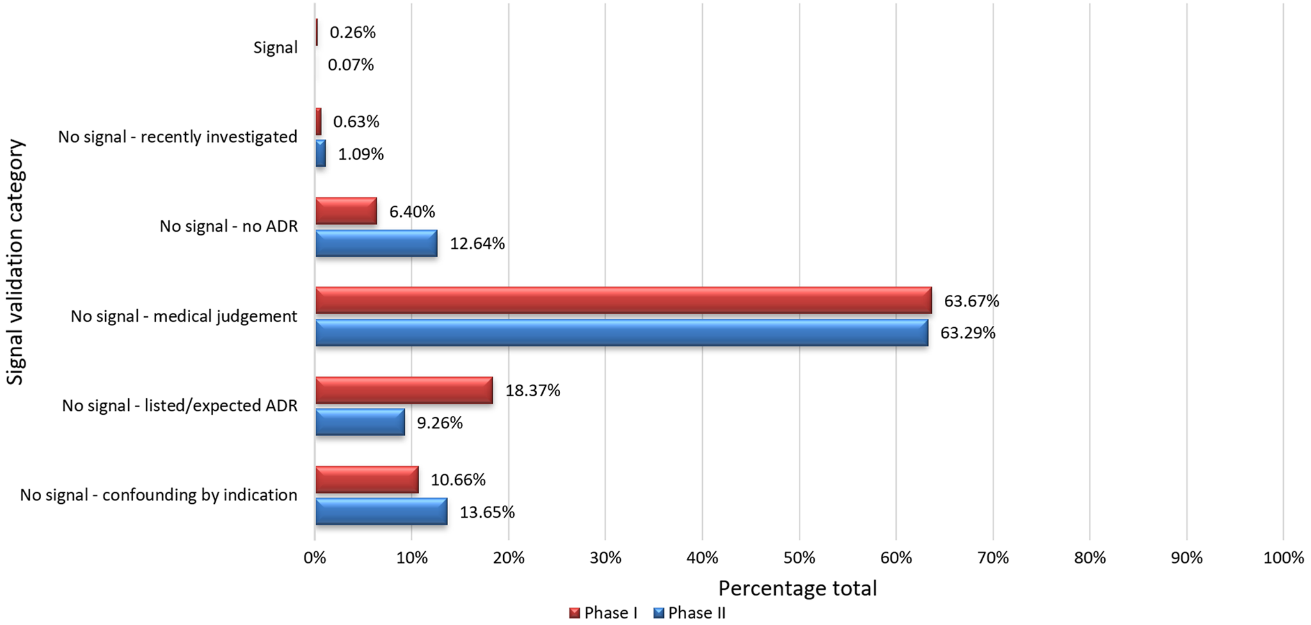

Fig. 1 Overall distribution of validated signals of disproportionate reporting over various categories in the historic signal validation data extracted for phase I and II of the experiment. *ADR* adverse drug reaction

included into the dataset for model retraining in March and April 2021, respectively.

q The latest phase II dataset contained 2.3 million PEC records and 6,606 SDRs and their validations.

The three products selected for phase I were two drugs and one biological product. They represent the late stage of the product life cycle and were chosen for the experiment because of their large dataset of historic ICSRs and SDRs. This helped to ensure the availability of a considerable amount of data for training the model. Phase II expanded the selection by an additional three drugs, which diversified the products across six different therapeutic areas and drug classes, from both the pharmaceuticals and the consumer health divisions.

2.2.2 SDR Validations

In our organization, SDRs are validated by safety experts as *signal* or *no signal* using one of five *no signal* classifications: *listed/expected adverse drug reaction* (ADR), *no ADR*, *recently investigated*, *medical judgment*, or *confounding by indication*. The five *no signal* classifications thus include the rationale for the *no signal* validation decision. These six predefined categories are specific to the authors' organization; other organizations may classify SDRs differently. The safety experts choose the signal validation category based on product knowledge and the evaluation of supporting information, including ICSR review.

Figure 1 shows the distribution of SDR validation classes observed in the data extracted for phase I and II. The vast majority of the SDRs were validated as *no signal*, with *medical judgment* being the most frequent category selected by the safety experts.

There is existing guidance as to which information shall be considered during signal validation, prioritization, and further assessment for decision making [3, 15, 18]. The guidance refers to previous awareness of the signal, strength of evidence about the causal relationship between the medicinal product and the event, and the clinical relevance of the ADR [3]. Regulatory guidance, as well as interviews with our company’s safety experts, helped to determine the selection of attributes for the data extraction and feature creation to inform the signal validation process. Both case data and SDR data were extracted on the level of medicinal product name and event Medical Dictionary for Regulatory Activities (MedDRA$^{®}$) preferred term (PT), as this is the data aggregation level used in the signal detection and validation process (Table 1).

2.3 Phase I: Set Up the Machine Learning Pipeline and Select a Promising Model

For the retrospective experiment (phase I), we considered ICSRs and historic SDRs and their validations from the past 6 years for three medicinal products. The data were used to train and test different ML models.

Table 1 Case data attributes extracted from the spontaneous reporting database, and signal of disproportionate reporting data attributes extracted from the signal detection data mart

Data source	Attribute level	Attributes	ICH E2B(R3) reference[a] [19]
Case data from safety database	Case attributes	Report type	C.1.3
		Country of incidence	E.i.9
		Case medically confirmed	E.i.8
	Patient attributes	Age group	D.2.3
		Gender	D.5
		Ethnicity	Not available
		Pregnancy	Not available
	Product attributes	Medicinal product name (suspect products)[b]	G.k.2.1.1b/ G.k.2.1.2b
		List of indications (1–3) as preferred terms	G.k.7.r.2b
	Event attributes	Event preferred term[b]	E.i.2.1b
		Event seriousness	E.i.3.2
		Event outcome	E.i.7
	PEC attributes	Time to onset of event	G.k.9.i.3.1
		Dechallenge	G.k.8 and E.i.7
		Rechallenge	G.k.9.i.4
		Event listedness	Not available
		Reporter causality	G.k.9.i.2.r.1 and r.3
		Company causality	G.k.9.i.2.r.1 and r.3
SDR data from signal detection data mart	SDR attributes	Medicinal product name (suspect product of interest)[b] Event preferred term[b] Flags: ●DME flag (as per company-specific DME list) ●listed flag (as per company core data sheet) ●trend flag (indicating an increased period frequency)	Not available
	Case counts	Case counts for this PEC: Each of them (a) cumulative, (b) for the current period,[c] and (c) for the prior period[c]: ●Total number of cases (all report types) ●Number of cases with report type spontaneous or literature ●Number of cases with report type study or published report from study ●Number of serious cases ●Number of fatal cases ●Case frequency[d]	Not available
	SDR validation attributes	SDR validation outcome[e]	Not available

DME designated medical event, *ICH* International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, *PEC* product–event combination, *SDR* signal of disproportionate reporting

[a] E2B(R3) data types and values were not explicitly retrieved and used as specified by the ICH. The E2B(R3) reference is only listed for the sake of clarity and attribute identification

[b] Attributes used for linking of the two datasets

[c] The signal detection periodicity is monthly. "Current period" refers to a 1-month look-back period into the previous month. "Prior period" refers to a look-back into the month when an SDR for the same PEC was validated the last time in the past

[d] Number of cases for this PEC divided by number of cases for the product

[e] Signal validation classification for the SDR done by safety expert in the past

2.3.1 Feature Engineering

The two feature sets used in our ML model were extracted from ICSR and SDR data. Most of the ICSR data were categorical in nature. They were converted into one-hot encoded representation [20] and then features were derived for each PEC by aggregating the ICSR data into a collection of features representing percentages and totals (see Table 2 for an example). These ICSR features were then combined with the SDR data, using the PEC as linking key. This approach of feature engineering provided unique data profiles of SDRs consisting of "percentage" and "total" ICSR features and the corresponding SDR validation annotations by safety experts.

Additional features were introduced at the SDR level, which counted how many times SDRs for the same PEC were assigned to which of the six possible signal validation

Table 2 Example of how features were engineered from the Individual Case Safety Report data for the Rechallenge attribute by creating two features (total and percent) for each available Rechallenge value (yes, no, unknown).

ICSR data

Case number	Product	Event	Rechallenge
1	3	2	Yes
2	3	2	Yes
3	3	2	No
4	3	2	No
5	3	2	Unknown

Resulting model features

Product	Event	Rechallenge Yes total	Rechallenge Yes percent	Rechallenge No total	Rechallenge No percent	Rechallenge Unknown total	Rechallenge Unknown percent
3	2	2	40%	2	40%	1	20%

ICSR Individual Case Safety Report

categories in the past. These count-based features were computed for all SDRs for which prior signal validations existed in the database and used as a *look-back mechanism* on past annotations of the safety experts while predicting the signal validation category. For the SDRs with no prior validations except the most recent one, these count-based features were filled with zeros.

2.3.2 Model Competition

In phase I, accuracy, weighted average F1 score (weighted by class frequency), and macro-average F1 score (arithmetic mean of class-wise F1 scores) [21] were used to decide upon the best-performing ML classification model and to compare models A and B (Sect. 2.3.3). All metrics were calculated using the Python Scikit-learn package [22].

To understand the behavior and performance of various types of ML models for our specific use case of signal validation classification, a classifier performance analysis was conducted where various ML models were trained and tested and the winner model was chosen based on the most stable and highest results in the model performance metrics. To ensure the stability of the results, a 3-fold cross validation and a feature ablation test were implemented. This ensured that the ML classifier neither overfit to a certain group of SDRs nor was dependent on only a certain subset of features. In this analysis, Random Forest, Linear Support Vector Classifier, Logistic Regression [23, 24] and eXtreme Gradient Boosting implementation of Gradient boosted trees ensemble model (XGBoost) [25] were compared. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) [26] was tested to address the class imbalance in the historic SDR data (Fig. 1).

The XGBoost model was the most stable and highest performing amongst all models tested for our use case with respect to performance scores. Therefore, we decided to use XGBoost for the classification task in the scope of our work.

Ultimately, the XGBoost model was trained with 100 boosting rounds using a learning rate of 0.1, a maximum tree depth of three layers, and L1 and L2 regularization terms equal to 0 and 1, respectively. The model was optimized using the multiclass classification error rate, which was calculated as the ratio of the number of wrongly classified SDRs to the total SDRs.

2.3.3 Model Variations: New vs. Recurring SDRs

To explore model variations, the data were split as shown in Fig. 2 and then fed into two instances of XGBoost, model A and model B. The first split was to separate out data for SDRs that had at least one preexisting validation from the SDRs that had no preexisting validation. This split allowed us to understand the behavior of the ML model in these two different groups of SDRs. The data for SDRs that had at least one preexisting validation were fed into model A, and the data for SDRs with no preexisting validation were fed into model B. The second split of the data was for the purpose of evaluating the ML models. The models were trained on 70% of the data and tested using the remaining 30% of the data by comparing the model predictions with the actual SDR validations completed by safety experts. This second split provided unbiased representative samples for model training and testing by first stratifying the data by SDR validation classes, randomly shuffling the data in each stratum, and then drawing training and test datasets.

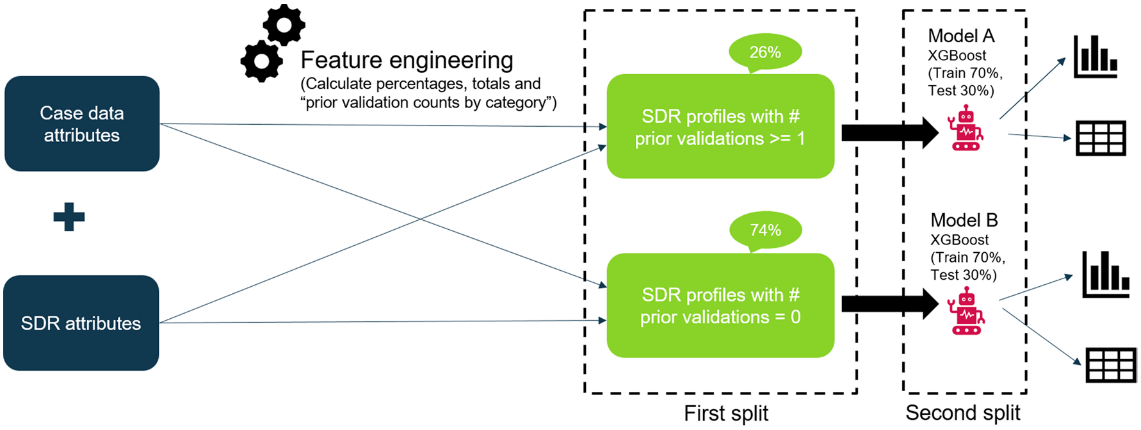
Feature engineering
(Calculate percentages, totals and "prior validation counts by category")
Case data attributes
SDR attributes
26%
SDR profiles with # prior validations >= 1
74%
SDR profiles with # prior validations = 0
Model A
XGBoost
(Train 70%, Test 30%)
Model B
XGBoost
(Train 70%, Test 30%)
First split
Second split

Fig. 2 Overall scheme of the data and model for phase I of the experiment showing the two splits in the data to evaluate the behavior of the model in each of the two groups of signal of disproportionate reporting (SDR) data

2.4 Phase II: Test the Model and Its Acceptance in a Real-Life Setting

Based on the promising results achieved in phase I of the experiment, the model was further tested in phase II in a 3-month prospective experiment. The experiment was expanded to include six products and ran in parallel with our organization's real-life monthly signal detection and validation process. It leveraged the same type of ML model as in phase I, i.e., XGBoost. However, in phase II, the model was trained on the entire phase II training dataset, and no separation into model A and B was performed because it was desired to have one single ML model that generalized to the complete dataset without separation of two groups of SDRs.

During this phase, each month, safety experts received the SDR validation predictions produced by the model for the respective month, performed their signal validation, and evaluated the usefulness of the model predictions. After each month, the model was retrained including the new SDR validations added by the safety experts based on their expertise. This scheme demonstrated a human feedback loop into the model to retrain it with the latest SDR validations.

For phase II of the experiment, accuracy was defined as an exact match percentage, i.e., percentage of matches of predicted classes to the classes assigned by safety experts. This accuracy was measured overall, as well as broken down by medicinal product and month, by novelty of SDR (first-time SDR vs. recurring SDR), and by signal validation class.

2.5 Model Explainability and Interpretability

To enable ML model interpretability, a SHapley Additive exPlanations (SHAP) analysis was implemented [27].

In phase I of the experiment, SHAP analysis was used to understand the "global" impact of input features on the overall model, which is further detailed in Sect. 3.

The SHAP framework also provides the capability to explore the "local" feature effects [28], which illustrates the impact of input features on individual predictions. This was used in phase II to present the three highest impact features for each model prediction to the safety experts.

The implications of model interpretability for use of ML in pharmacovigilance are further explained in Sect. 4.

3 Results

3.1 Phase I (Retrospective Experiment)

3.1.1 Model Performance

As described in Sect. 2.3.3, during phase I of the experiment, data were split and results computed for two types of SDRs in our data: 26% of the SDRs had at least one prior validation, and these data were used for model A (Fig. 3a); 74% of SDRs had no prior validation, and these data were used for model B (Fig. 3b).

In normalized confusion matrices (Fig. 3), better model performance is represented by higher numbers on the diagonal of the confusion matrix because entries on the diagonal represent correct classifications by the model. Off-diagonal entries show misclassifications. Fig. 3 shows that model A performed relatively better than model B overall despite the lower quantity of data for it. It can be seen in Fig. 3b that model B relatively misclassified more data from the *no signal—no adr*, *no signal—recently investigated*, and *no signal—listed/expected adr* categories into the false category of *no signal—medical judgment* because it had relatively

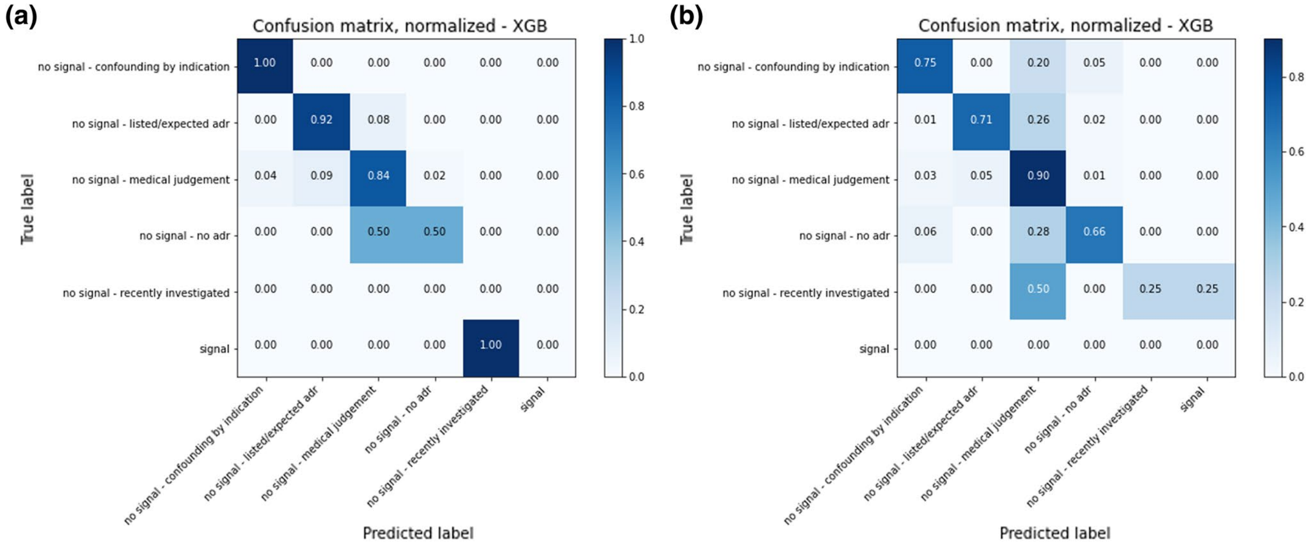

Fig. 3 Normalized confusion matrix for SDR validation classifications in phase I of the experiment. **a** Confusion matrix for model A: SDRs with at least one prior validation; 26% of SDRs belonged to this group. **b** Confusion matrix for model B: SDRs with no prior validation; 74% of SDRs belonged to this group. Values and color scale range from 0.00 (0% of true class) to 1.00 (100% of true class). Results are based on the 30% test datasets for model A and model B. *ADR* adverse drug reaction, *predicted label* signal validation prediction by ML model, *SDR* signal of disproportionate reporting, *true label* signal validation outcome determined by safety expert, *XGB* eXtreme Gradient Boosting model

less discriminatory power because of a lack of prior validations. However, Fig. 3a shows that model A performed relatively better overall by showing lower values in off-diagonal entries, suggesting better classifications produced by the model. These findings demonstrate that the presence of prior validation counts in the feature set contributed to more correct classifications by the model.

Table 3 shows the comparison of the performance in terms of the classification reports produced using model A and model B. It can be observed that model B achieved a better macro-average F1 score than model A (0.58 vs. 0.53, respectively). However, when comparing accuracy, model A performed slightly better than model B (0.84 vs. 0.83, respectively).

Furthermore, when comparing the class-wise F1 scores, the model performance for the classes *no signal—confounding by indication* and *no signal—listed/expected adr* benefited from prior validation count features. This finding was supported by observations in phase II of the experiment: when safety experts assigned a validation category of *no signal—confounding by indication* or *no signal—listed/expected adr* to an SDR, there was a high likelihood that their validation decision would stay the same for that SDR when it was re-signaled the next time by the signal detection system. Therefore, this knowledge of prior validation informing the future validation category led to a visible performance benefit of model A (see Table 3).

For the class *no signal—no adr*, model A showed a lower F1 score because of lower recall when compared with model B. The lower recall was because prior validations for *no signal—no adr* contained both *no signal—medical judgment* and *no signal—no adr*, and—in this scenario—model A did not benefit from the prior validations.

There were very few SDRs with validation class *signal* in the training data and only one SDR of such a class in the test set of model A. For model B, there were no SDRs with validation class *signal* in train or test data.

Another noticeable difference between models A and B is that there were no SDRs with signal categorization of *no signal—recently investigated* in the model A test data.

3.1.2 Model Explainability: "Global" Feature Impact

This section presents the results of the SHAP analysis. A notable difference in the SHAP-based overall feature importance can be observed for model A (Fig. 4a) versus model B (Fig. 4b).

Figure 4a indicates that the feature *N_PRIOR_no_signal—listed/expected adr*, which contains the count of how many times in the past a given SDR was categorized as *no signal—listed/expected adr*, had the highest overall impact on the ML model, and the purple color shows that it was the most informative for the same corresponding class *no signal—listed/expected adr*, based on which this feature was created. Also, we see the same for the following similar three highest impact features: *N_PRIOR_no_signal—medical judgment*, *N_PRIOR_no_signal – no adr*, and *N_PRIOR_no_signal—confounding by indication*. These were also highly informative about the respective classes, based on which these were calculated.

Table 3 Test set distribution and model performance metrics for model A and model B in phase I of the experiment

SDR validation class	Model A – 386 (26%) SDRs with one or more prior validations				Model B – 1519 (74%) SDRs without prior validation			
	Precision	Recall	F1 score	Test records (73)	Precision	Recall	F1 score	Test records (525)
No signal—confounding by indication	0.80	1.00	0.89	10.96%	0.76	0.75	0.75	10.48%
No signal—listed/expected adr	0.75	0.92	0.83	17.81%	0.79	0.71	0.75	18.48%
No signal—medical judgment	0.90	0.84	0.87	61.64%	0.87	0.90	0.88	64.19%
No signal—no adr	0.75	0.50	0.60	8.22%	0.68	0.66	0.67	6.10%
No signal—recently investigated	0.00	0.00	0.00	0.00%	1.00	0.25	0.40	0.76%
Signal	0.00	0.00	0.00	1.37%	0.00	0.00	0.00	0.00%
Accuracy			0.84				0.83	
Macro-average F1 score	0.53	0.54	0.53		0.68	0.54	0.58	
Weighted-average F1 score	0.84	0.84	0.83		0.83	0.83	0.83	

Results are based on the 30% test datasets for model A and model B

ADR adverse drug reaction, *SDR* signal of disproportionate reporting

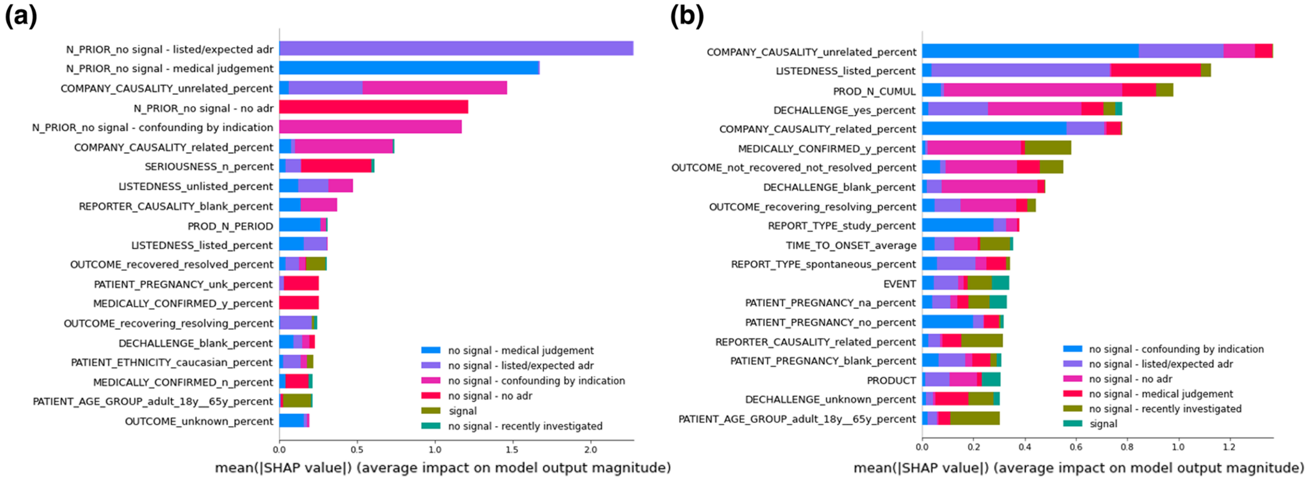

Fig. 4 Comparison of the overall feature importance for model A and model B in phase I of the experiment. **a** Plot for SDRs with one or more prior validations. **b** Plot for SDRs with no prior validations for the SDRs. The comparison between the two figures shows that the machine learning model benefits from the availability of prior validation features. When the model does not have prior validation information, it leverages features computed from case data. The length of the bars depicts the magnitude of the impact of various features on informing the machine learning model. The color within the bars explains the specific class or classes for which the feature contributed to informing the model. However, this plot does not indicate the direction of impact, i.e., whether the impact of the feature is positive or negative. The figure was produced using SHAP TreeExplainer package [28]. Results are based on the 30% test datasets for model A and model B. *ADR* adverse drug reaction, *SDR* signal of disproportionate reporting

Another interesting point to note here is that the feature *COMPANY_CAUSALITY_unrelated_percent*, which quantifies the percentage of unrelated event reports based on the company causality assessment for the PECs, was also informative to discriminate between the classes *no signal—listed/expected adr* and *no signal—confounding by indication*. This finding was discussed with safety experts, and they were in agreement that the company's causality assessment in the ICSR data also helps in deciding whether an SDR was *no signal* because of confounding by indication or because it was already listed and expected.

In model B, the feature *COMPANY_CAUSALITY_unrelated_percent* had the highest impact on the model performance and was the most informative feature about the *no signal—confounding by indication* class (Fig. 4b). Additionally, the feature named *LISTEDNESS_listed_percent*, which quantifies the percentage of listed events for the respective PEC, was the second most important feature for the model. The dominant purple color of this bar shows that it was the most informative feature for the *no signal—listed/expected adr* class in the data, which is expected because SDRs would be categorized into this class most likely if the majority of the underlying PECs are marked as *listed* in the respective ICSRs.

Importantly, the feature importance order and corresponding impacts on individual classes was different between the

Table 4 Example for a signal validation prediction for one SDR in month 2 of phase II of the experiment showing the information presented to safety experts

Product	Event	Signal validation prediction	Confidence score	Top three highest impact features	Probabilities for other signal validation classes
2	8	No signal—medical judgment	0.972	PROD_N_PERIOD, TREND_FLAG_new, OUTCOME_not_recovered_not_resolved_percent	No signal—confounding by indication: 0.01 No signal—listed/expected adr: 0.01 No signal—no adr: 0.005 Signal: 0.002 No signal—recently investigated: 0.001

These additional columns were embedded in a signal validation report containing all SDR information from the signal detection system with one line per SDR

OUTCOME_not_recovered_not_resolved_percent: percentage of cases where the event outcome was "not recovered/resolved" from all cases with this PEC, PEC product–event combination, *PROD_N_PERIOD*: number of new cases for the product in latest signal detection period, SDR signal of disproportionate reporting, *TREND_FLAG_new*: trend flag "new" indicating that this PEC was identified as SDR the first time

two models. It can also be seen that model A considered the prior validation count features as the most informative for discriminating between the classes. On the other hand, model B utilized almost all available features that were computed from ICSR data, in absence of prior validations.

3.2 Phase II (Prospective Experiment)

3.2.1 Presentation of Model Predictions: Confidence Scores and "Local" Feature Impact

Table 4 shows how the ML model predictions were presented to the safety experts in phase II of our experiment. To quantify the reliability of the model's predictions, probabilities for the predicted classes were calculated. The class with the highest probability was considered as the final prediction class and the corresponding probability was presented as the confidence score. To further assist with the interpretation of the results and to develop trust in the predictions of the model, all other class probabilities from the model were also presented to the safety experts in descending order. SHAP's "local" explanations capability was also used to display the three highest impact features for each prediction.

3.2.2 Model Performance

Overall, 133 SDRs were classified during the prospective phase II experiment for the six medicinal products. The accuracy in phase II was stable over the 3 months (83–86%; see Table 5) and confirmed the accuracy level found in phase I of the experiment. Accuracy for recurring SDRs (90.0%) was better than for SDRs that signaled for the first time (72.1%; see Table 6), which again confirmed the previous findings of phase I. During the 3 months, no SDRs were classified as *signal* or *no signal—recently investigated* by the safety experts for the six products in scope. The majority of SDRs again fell into the *no signal—medical judgment* category (94 of 133 SDRs; see Table 7). This corresponds with the distribution of classes in the historic signal validation data. The prediction accuracy for the class *no signal—medical judgment* was the highest (92.6%) of all classes.

3.2.3 User Acceptance

SHAP analysis was introduced during phase I of the experiment based on the safety experts' feedback. They wanted to understand how the model predicted the signal validations.

The SHAP analysis provided an explanation of the most important features that affected the decisions of the ML model. The safety experts accepted the decision-support tool because the SHAP information provided transparency of the model's decision rationale.

In addition to the most important features, the presentation of the model's confidence scores also contributed to higher user acceptance. We found that the SDR validations that matched between safety experts and the model had, in general, higher confidence scores in their predictions, whereas the "no matches" had widespread and generally lower confidence scores (data not shown).

4 Discussion

4.1 Model Performance: Strengths and Limitations of the Model

Our experiment to explore the predictive capabilities of ML for signal validation showed promising results. The results from the performance metrics (see Tables 3, 5, 6, 7) illustrate that an off-the-shelf XGBoost ML model can differentiate between the various classes of *no signal* SDRs by

Table 5 Accuracy of signal validation predictions by medicinal product over 3 subsequent months in phase II of the experiment

Product	Month 1				Month 2				Month 3				Total of SDRs	Accuracy
	Number of SDRs			Accuracy	Number of SDRs			Accuracy	Number of SDRs			Accuracy		
	Match[a]	No match	Total		Match	No match	Total		Match	No match	Total			
1	18	1	19	94.7%	10	2	12	83.3%	14	2	16	87.5%	47	89.4%
2	4	0	4	100.0%	2	1	3	66.7%	0	0	0	0.0%	7	85.7%
3	2	1	3	66.7%	7	1	8	87.5%	2	0	2	100.0%	13	84.6%
4	3	0	3	100.0%	3	0	3	100.0%	2	0	2	100.0%	8	100.0%
5	7	3	10	70.0%	9	3	12	75.0%	12	4	16	75.0%	38	73.7%
6	3	1	4	75.0%	9	1	10	90.0%	5	1	6	83.3%	20	85.0%
Total	37	6	43	86.0%	40	8	48	83.3%	35	7	42	83.3%	133	84.2%

SDR signal of disproportionate reporting

[a]*Match* Prediction by the machine learning model matched the signal validation outcome determined by safety expert

Table 6 Accuracy of signal validation predictions by novelty of signal of disproportionate reporting in phase II of the experiment

Novelty of SDR	Number of SDRs			Accuracy
	Match[a]	No match	Total	
New SDR[b]	31	12	43	72.1%
Recurring SDR[c]	81	9	90	90.0%
Total	112	21	133	84.2%

ML machine learning, *PEC* product–event combination, *SDR* signal of disproportionate reporting

[a]Prediction by the ML model matched the signal validation outcome determined by safety expert

[b]SDR for a specific PEC that was identified for the first time by the signal detection system

[c]SDR for a specific PEC that had already been identified one or more times but meets predefined re-signaling criteria

utilizing the company's ICSR and SDR data and without further data annotation.

The *no signal—medical judgment* category conceptually contains multiple subcategories from the decision criteria point of view, making it the majority class in the data and ML model. This resulted in better performance for this class, presumably since it had more examples for training of the model.

An important strength of our model is that it leverages the prior validation features that provide information about how many times historically the SDRs have been assigned to which signal validation categories. This provides a *look-back mechanism* to the model when making a prediction about a given SDR. For example, if a certain SDR has been categorized as *no signal—listed/expected adr* in the past, the model remembers this past validation of the SDR and provides consistency in the signal validation decision.

We hypothesized that an oversampling approach such as SMOTE should help the classifier improve its performance using more data points for learning. Surprisingly, the SMOTE implementation experiment did not help the model in improving its performance and even slightly decreased the accuracy and macro F1 score of the classifier, and thus was not utilized in the final model. One reason for this slight performance degradation of the XGBoost model could be that SMOTE generalized the model too much and thereby missed learning the nuances within the original data of the minority classes.

Further testing is needed to confirm the generalizability of the model since the experiment covered a limited number of products only. An extended diverse set of products from different phases of the product life cycle is recommended to be used for further testing of model generalization.

A limitation in our experiment was the low, single digit, number of SDRs classified as *signal* in the data that were used for model training and testing. Given such scarce

Table 7 Accuracy of signal validation predictions by signal of disproportionate reporting validation class in phase II of the experiment

SDR validation class	Number of SDRs			Accuracy
	Match[a]	No match	Total	
No signal—confounding by indication	9	2	11	81.8%
No signal—listed/expected adr	5	8	13	38.5%
No signal—medical judgment	87	7	94	92.6%
No signal—no adr	11	3	14	78.6%
No signal—recently investigated	0	1	1	0.0%
Signal	0	0	0	NA
Total	112	21	133	84.2%

NA not applicable, *SDR* signal of disproportionate reporting

[a] Prediction by the machine learning model matched the signal validation outcome determined by safety experts

"ground truth" to learn from, we expected a low performance of the model to correctly classify SDRs as *signal*. In fact, in phase I, the test data only contained one *signal*, and this was misclassified by the model. In phase II, the test data contained not even one *signal*, so the model performance for the *signal* class could not be calculated. Because the *signal* versus *no signal* classification is essential in the signal validation process, we plan to address this limitation in future enhancements (see Sect. 4.4).

4.2 Explainability

Ensemble tree models such as Random Forests and Gradient Boosted Trees are often go-to models as they can perform well in various domains [29, 30]. However, in addition to high accuracy, interpretability is also highly desirable. Especially in a domain such as pharmacovigilance, which is highly regulated and impacts on patient safety and public health, one needs to understand how an ML model uses input features to make predictions. Significant work can be found on explaining the overall impact of input features on ML models [31–33]. We used SHAP analysis [27] to enhance the explainability and transparency of the model. One successful example of using SHAP in healthcare is the application of the "Tree Explainer" from the SHAP framework for the explanation of predictions of hypoxemia [34].

The benefits of using SHAP analysis in this experiment were twofold. First, it provided an understanding of what features in the data are the most impactful features for the overall multiclass classifier model. Second, it provided a mechanism to build trust in the user community of the model by surfacing the features that impacted a particular prediction of the classification model.

The additional information presented to the safety experts in phase II of our experiment, together with the predicted class, comprised three important elements: the confidence score for each prediction, the three highest impact features for each prediction, and the probabilities for all other signal validation classes (see Table 4). The strength of providing this additional information together with the predictions is that it removed the "black box" character of the ML model by sharing the model's reasons for its decision making, which the safety experts could then review and consider. The SHAP analysis results presented to the safety experts significantly enhanced their understanding of the otherwise concealed decision criteria of the ML model and increased their confidence in the generated predictions.

The SHAP framework in the modeling process may reveal features that were considered less important by safety experts in the decision-making process. New insights can be brought to light by virtue of this data-driven approach. These informative features might positively influence and streamline the signal validation process.

4.3 User Acceptance

Overall, the modeling experiment was well accepted by the safety experts. One interesting piece of feedback from the safety experts was that the predictions made them think vigilantly about the assessment of SDRs when the model's categorization deviated from theirs.

The safety experts preferred using the model to support the validation process rather than letting it be totally autonomous. They also explained that additional product and disease knowledge is taken into consideration in their decision-making process, including mechanism of action of the drug and pharmacokinetic, toxicological, and epidemiological information that is not always included in the structured fields of ICSRs or in the SDR data.

Providing the safety experts with model predictions before they made their own assessment entailed the risk of biasing their judgment. However, we learned that the predictions worked like an independent second opinion that stimulated rather than biased the validation process.

In a concluding survey of the experiment, the involved safety experts confirmed the business value of the predictions provided by the model towards an increased efficiency, consistency, and quality of the signal validation process. In summary, the safety experts valued the predictions and would like to utilize them within their signal management application.

4.4 Outlook

As a next step, we plan to implement enhancements collected from the safety experts and project team. Examples of potential improvement ideas include engineering additional features from ICSR data and from reference data, such as MedDRA hierarchy levels or drug class information. An analysis of safety experts' comments (prospective and retrospective) when their signal validation category selection differed from the model's prediction will help to identify areas for future model improvements.

Furthermore, we aim to add more products and diversify them to include different phases of the product lifecycle to further test the model's ability to generalize. Prior to integrating the model into the signal management application, we will gather further experience by running the ML pipeline in parallel with the productive signal management process and creating a dashboard for the safety experts that presents the signal validation predictions from our model.

To successfully classify SDRs as *signal* versus *no signal*, we will extend our training data to include more products and consider augmenting our ICSR and signal validation data with external data. The resulting binary classifier model trained on this extended and augmented data could then be combined and used in a two-step approach as a sequence of two models. Specifically, the first model will be designed for supporting the *signal* versus *no signal* decision, and a second one will perform the classification for the different *no signal* justifications.

Finally, we believe that the knowledge gained in this experiment with the *quantitative* signal detection process could also be leveraged for a different use case: the signal validation of safety observations identified in the ongoing monitoring process of ICSRs, which is a periodic manual medical case review and as such a major component of the *qualitative* signal detection process [1]. Based on the promising results from this research, it may be worth further exploring whether this process could also be supported by ML.

5 Conclusions

This experiment demonstrated that signal validation in pharmacovigilance can be supported by an ML-based prevalidation step to improve process efficiency and consistency. We were able to train a multiclass classification model to predict signal validation classifications for SDRs, which showed promising results in terms of accuracy. Medical review by safety experts will always remain an essential part of the signal validation process, but it can be performed faster and in a more consistent way if it is augmented with ML predictions.

For safety experts, model explainability plays a major role in building trust in and acceptance of ML models. Using SHAP analysis helped to improve the model explainability.

As the training and test data only contained a limited amount of SDRs that were validated as *signals*, the data were not appropriate for training the supervised ML model to specifically distinguish between *signals* and *no signals* with considerable accuracy. Therefore, an area for further research is to combine this approach with a binary classifier supporting a *signal* versus *no signal* differentiation during the signal validation process.

Acknowledgements The authors thank Enrico Santus for his part in the ideation of the experiment and constructive comments. We also thank John Reinhard Pietzsch and Marion Legler for constructive comments. The MedDRA$^{®}$ trademark is registered by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).

Declarations

Funding This experiment was conducted during a project at Bayer AG that involved consultancy work carried out by Magnus Lerch. Magnus Lerch has received compensation for consultancy work from Bayer AG.

Conflict of interest Aasia Bhatti holds shares in Bayer AG. Muhammad Imran, Aasia Bhatti, David King, Magnus Lerch, Jürgen Dietrich, Guy Doron, and Katrin Manlik have no conflicts of interest that are directly relevant to the content of this experiment. Muhammad Imran, Jürgen Dietrich, Guy Doron, and Katrin Manlik are full-time employees of Bayer AG. Aasia Bhatti and David King are full-time employees of Bayer US LLC. The views expressed in this article are those of the authors and do not necessarily reflect the official policies or position of Bayer AG, Bayer US LLC or Lenolution GmbH.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The datasets generated during the experiment were retrieved from Bayer's safety database and contain private patient information that cannot be made publicly available.

Code availability The code used for this experiment is proprietary and therefore cannot be provided.

Author contributions Muhammad Imran, Aasia Bhatti, David King, Magnus Lerch, Jürgen Dietrich, and Guy Doron were involved in the conception and design of the experiment in phase I and phase II. Katrin Manlik was involved in the conception and design of the experiment in phase II. Muhammad Imran was responsible for the design and implementation of data processing, feature engineering pipeline, the ML models, the interpretation and presentation of the results, and the implementation of SHAP analysis. All authors contributed to the interpretation of the data analysis results and assisted with the concept and draft revisions of the manuscript. Guy Doron and Jürgen Dietrich assisted with literature search and review. All authors reviewed and approved the final manuscript and accept full responsibility for its overall content.

References

1. CIOMS. Practical aspects of signal detection in pharmacovigilance: Report of CIOMS Working Group VIII. Geneva. Geneva: CIOMS; 2010.
2. Candore G, Juhlin K, Manlik K, Thakrar B, Quarcoo N, Seabroke S, et al. Comparison of statistical signal detection methods within and across spontaneous reporting databases. Drug Saf. 2015;38(6):577–87. https://doi.org/10.1007/s40264-015-0289-5.
3. European Medicines Agency. Guideline on good pharmacovigilance practices (GVP) Module IX – Signal management (Rev 1). 2017. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-ix-signal-management-rev-1_en.pdf. Accessed 15 Aug 2021.
4. Bate A, Hobbiger SF. Artificial intelligence, real-world automation and the safety of medicines. Drug Saf. 2021;44(2):125–32. https://doi.org/10.1007/s40264-020-01001-7.
5. Abatemarco D, Perera S, Bao SH, Desai S, Assuncao B, Tetarenko N, et al. Training augmented intelligent capabilities for pharmacovigilance: applying deep-learning approaches to individual case safety report processing. Pharmaceut Med. 2018;32(6):391–401. https://doi.org/10.1007/s40290-018-0251-9.
6. Ghosh R, Kempf D, Pufko A, Barrios Martinez LF, Davis CM, Sethi S. Automation opportunities in pharmacovigilance: an industry survey. Pharmaceut Med. 2020;34(1):7–18. https://doi.org/10.1007/s40290-019-00320-0.
7. Schmider J, Kumar K, LaForest C, Swankoski B, Naim K, Caubel PM. Innovation in pharmacovigilance: use of artificial intelligence in adverse event case processing. Clin Pharmacol Ther. 2019;105(4):954–61. https://doi.org/10.1002/cpt.1255.
8. Du J, Xiang Y, Sankaranarayanapillai M, Zhang M, Wang J, Si Y, et al. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning. J Am Med Inform Assoc. 2021;28(7):1393–400. https://doi.org/10.1093/jamia/ocab014.
9. van Stekelenborg J, Ellenius J, Maskell S, Bergvall T, Caster O, Dasgupta N, et al. Recommendations for the use of social media in pharmacovigilance: lessons from IMI WEB-RADR. Drug Saf. 2019;42(12):1393–407. https://doi.org/10.1007/s40264-019-00858-7.
10. Comfort S, Perera S, Hudson Z, Dorrell D, Meireis S, Nagarajan M, et al. Sorting through the safety data haystack: using machine learning to identify individual case safety reports in social-digital media. Drug Saf. 2018;41(6):579–90. https://doi.org/10.1007/s40264-018-0641-7.
11. Lee CY, Chen YP. Machine learning on adverse drug reactions for pharmacovigilance. Drug Discov Today. 2019;24(7):1332–43. https://doi.org/10.1016/j.drudis.2019.03.003.
12. Bate A, Lindquist M, Edwards IR, Olsson S, Orre R, Lansner A, et al. A Bayesian neural network method for adverse drug reaction signal generation. Eur J Clin Pharmacol. 1998;54(4):315–21. https://doi.org/10.1007/s002280050466.
13. Bae JH, Baek YH, Lee JE, Song I, Lee JH, Shin JY. Machine learning for detection of safety signals from spontaneous reporting system data: example of nivolumab and docetaxel. Front Pharmacol. 2020;11: 602365. https://doi.org/10.3389/fphar.2020.602365.
14. Ibrahim H, Abdo A, El Kerdawy AM, Eldin AS. Signal detection in pharmacovigilance: a review of informatics-driven approaches for the discovery of drug-drug interaction signals in different data sources. Artif Intell Life Sci. 2021;1:100005. https://doi.org/10.1016/j.ailsci.2021.100005.
15. Levitan B, Yee CL, Russo L, Bayney R, Thomas AP, Klincewicz SL. A model for decision support in signal triage. Drug Saf. 2008;31(9):727–35. https://doi.org/10.2165/00002018-200831090-00001.
16. Evans SJ, Waller PC, Davis S. Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. Pharmacoepidemiol Drug Saf. 2001;10(6):483–6. https://doi.org/10.1002/pds.677.
17. Lerch M, Nowicki P, Manlik K, Wirsching G. Statistical signal detection as a routine pharmacovigilance practice: effects of periodicity and resignalling criteria on quality and workload. Drug Saf. 2015;38(12):1219–31. https://doi.org/10.1007/s40264-015-0345-1.
18. Pacurariu A, van Haren A, Berggren AL, Grundmark B, Zondag D, Harder H, et al. SCOPE Work Package 5 Signal Management. Best Practice Guide, Annex 2. 2016. https://www.ema.europa.eu/documents/other/scope-training-signal-management-best-practice-guide_en.pdf. Accessed 08 Aug 2021.
19. European Medicines Agency. ICH guideline E2B (R3) on electronic transmission of individual case safety reports (ICSRs)—data elements and message specification—implementation guide. 2013. https://www.ema.europa.eu/en/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human-use_en-4.pdf. Accessed 16 Dec 2021.
20. Poenaru-Grigorescu CJ, Ghic G. Analyzing the dummy variable in econometric models highlighting the binary choice regression models. Qual Access Success. 2016;17(S3):182–7.
21. Opitz J, Burst S. Macro F1 and Macro F1. 2021. https://arxiv.org/abs/1911.03347. Accessed 12 Aug 2021.
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12(85):2825–30.

23. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer; 2013.
24. Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.
25. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. 2016. https://doi.org/10.1145/2939672.2939785
26. Chawla NV, Bowyer KW, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953.
27. Lundberg SM, Lee S-I. A Unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., (eds) Advances in neural information processing systems 30: Curran Associates, Inc.; 2017. pp. 4765-74
28. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9.
29. Freeman EA, Moisen GG, Coulston JW, Wilson BT. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Can J For Res. 2016;46(3):323–39. https://doi.org/10.1139/cjfr-2014-0562.
30. Krauss C, Do XA, Huck N. Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur J Oper Res. 2017;259(2):689–702. https://doi.org/10.1016/j.ejor.2016.10.031.
31. Molnar C. Interpretable machine learning—a guide for making black box models explainable. 2021. https://christophm.github.io/interpretable-ml-book/. Accessed 04 Aug 2021.
32. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. Electronics. 2019;8(8):832. https://doi.org/10.3390/electronics8080832.
33. Tjoa E, Guan C. A Survey on explainable artificial intelligence (XAI): towards medical XAI. IEEE Trans Neural Netw Learn Syst. 2020. https://doi.org/10.1109/TNNLS.2020.3027314.
34. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749–60. https://doi.org/10.1038/s41551-018-0304-0.

Authors and Affiliations

Muhammad Imran[1] · Aasia Bhatti[2] · David M. King[3] · Magnus Lerch[4] · Jürgen Dietrich[5] · Guy Doron[6] · Katrin Manlik[7]

[1] Bayer AG, Digital Transformation and Information Technology Pharma, Decision Science and Advanced Analytics for Medical Affairs, Pharmacovigilance and Regulatory Affairs, Müllerstr. 178, 13353 Berlin, Germany

[2] Bayer US LLC, Pharmaceuticals, Pharmacovigilance, Benefit-Risk Management TA Radiology, Whippany, NJ, USA

[3] Bayer US LLC, Digital Transformation and Information Technology Pharma, Adverse Event Management, Morristown, NJ, USA

[4] Lenolution GmbH, Berlin, Germany

[5] Bayer AG, Pharmaceuticals, Pharmacovigilance, Innovation and Digitalization, Berlin, Germany

[6] Bayer AG, Pharmaceuticals, Pharmacovigilance, R&D, Data Sciences, Berlin, Germany

[7] Bayer AG, Pharmaceuticals, Pharmacovigilance, Data Science and Insight Generation, Berlin, Germany