

DOI: [10.26354/bb.11.2.103.2026](https://doi.org/10.26354/bb.11.2.103.2026)

Krzysztof Iwanicki*

ORCID: 0000-0002-3322-2756

krzysztof.iwanicki@oirp.warszawa.pl

Machine Learning in the Validation of IRB Models: Regulatory Aspects of Its Use in the Light of the Positions of the EBA, ECB and KNF

ABSTRACT

This article aims to determine the scope of permissible use of machine learning in validating models used under the Internal Ratings Based (IRB) approach. After explaining the nature of IRB models and their validation process, it describes the regulatory aspects of using machine learning in this process, with particular attention to the development of challenger models. It also discusses the risks and limitations associated with using machine learning models, including the possibility that their predictive performance may deteriorate as a result of concept drift and data drift. The article also considers the significance of explainability in the validation process. It concludes that the applicable regulations permit machine learning to be used as one of the tools in the validation of IRB models. At the same time, the current state of computer science does not support the conclusion that machine learning alone can achieve all the objectives of validation. Consequently, machine learning cannot serve as a functional substitute for the validation process.

Keywords: IRB models, IRB model validation, machine learning, model risk, explainable artificial intelligence (XAI).

JEL Codes: G21, G28, C52, C45, C53

* Krzysztof Iwanicki – Warsaw Bar Association of Attorneys-at-Law.

IRB MODELS

Risk is inherent in banking. It stems from banks' role as financial intermediaries (Zawadzka, 2001, p. 52; Matuszyk, 2008, p. 13); banking involves, among other things, taking on and managing risk (Iwanicz-Drozdowska, 2024, p. 17). Risk is defined either as *"the possibility of failing to achieve an intended objective (or outcome), that is, a deviation below the desired state (a negative approach to risk), or any deviation, whether positive or negative, from the intended objective (a neutral approach to risk)"* (Iwanicz-Drozdowska and Chmielewski, 2024, p. 43), or as objectified uncertainty grounded in measurable indicators, which distinguishes it from uncertainty (Niedziółka, 2018, pp. 305–306). This article adopts the first definition.

Credit risk is one component of banking risk. The literature distinguishes narrow and broad conceptions of this category. In the narrow sense, it means *"the danger that a bank's debtor will fail to repay [...] the loan taken out, together with interest and fees"* (Matuszyk and Ptak-Chmielewska, 2024, p. 67); in the broad sense, it means *"the risk that a transaction counterparty will fail to meet its obligations to the bank, in relation to loans, guarantees, sureties, bankers' acceptances, letters of credit, interbank deposits, debt instruments and over-the-counter derivatives"* (Matuszyk and Ptak-Chmielewska, 2024, pp. 67–68). The following analysis adopts the first conception.

An essential element of managing individual credit risk is assessing creditworthiness, defined as *"the ability to fulfil all obligations under a credit agreement on time and in full"* (Wiatr, 2008, p. 77). This assessment is made, among other things, using models that simplify reality *"through the use of quantitative and expert methods"* (Komisja Nadzoru Finansowego, 2024, p. 13).

Credit risk, including the calculation of risk-weighted exposure amounts, may be measured either using the standardised approach, whose rules are prescribed by legislation and the supervisory authority, or using the Internal Ratings Based (IRB) approach. The latter is based on the assumption that *"a bank is able to assess the scale of the credit risk it takes on with sufficient accuracy. The essence of the IRB approach is the possibility of using internally estimated risk parameters"* (Schab, 2005, p. 93). The legal basis for using the IRB approach is Article 143(1) of Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012 (OJ L 176, 27 June 2013, pp. 1–337, as amended), hereinafter the *"CRR"*. Under this provision, the competent authority permits institutions to calculate risk-weighted exposure amounts using the IRB approach once specified conditions are met.

VALIDATION OF IRB MODELS

Use of the IRB approach entails an obligation to validate the models, that is, to *"assess the model's performance"* (Komisja Nadzoru Finansowego, 2015, p. 8).

The ECB states that validation is intended to *"form an opinion on whether the final rating system [...] meets the regulatory requirements"* (European Central Bank, 2023, p. 15) and to ensure an

effective challenge to the adequacy of model performance and the appropriateness of the rating system (European Central Bank, 2023, p. 16). In turn, the Polish Financial Supervision Authority (KNF) states that *“the primary objective of model validation [...] is to reduce model risk by verifying the correctness and assessing the performance and functioning of the model; identifying areas in which justified improvements may be made both directly to the model itself and to the process by which it is used; making relevant recommendations and assigning appropriate priorities; verifying the quality of implementation of recommendations; and promoting high standards of model management”* (Komisja Nadzoru Finansowego, 2015, p. 39).

The ECB indicates that the validation process uses, among other things, two types of sample. The first, *“performance metric samples”*, involve testing the model so that *“obligors or facilities are assigned to grades or pools and the empirical metrics (e.g. defaults) can be observed to assess the core model performance”* (European Central Bank, 2023, p. 66). The second, *“application test samples”*, involve assigning obligors or facilities to grades or pools when insufficient time has elapsed for the empirical realisations needed to assess the model to be observed (European Central Bank, 2023, p. 66). In practice, this sample entails testing the model on current exposures that are not yet due.

INTRODUCTION TO THE USE OF MACHINE LEARNING

The possibility of using machine learning in the validation of IRB models has appeared in EBA documents since at least 2021. That year, the EBA began consulting European financial-sector entities and issued a paper stating that *“the identified main areas of use are: [...] Model validation, where ML models can be used, for example, as model challengers or for benchmarking”* (European Banking Authority, 2021, p. 13). At that time, the use of machine learning more broadly in relation to IRB models was treated with caution (European Banking Authority, 2021, p. 10), and *“institutions have made the strategic decision to move away from using ML in regulatory areas”* (European Banking Authority, 2021, p. 11).

There is a distinction between basing an IRB model on machine learning and using machine learning to validate an IRB model that is not itself based on machine learning. The distinction is significant, although it is not always made *expressis verbis*. Before drawing a conclusion from a given source, it is therefore necessary to establish whether, and to what extent, that source addresses the former or the latter. For example, the ECB’s publication entitled *“Guide to Internal Models”* devotes an entire chapter to the use of machine learning in relation to internal models; however, that analysis concerns internal models based on machine learning (European Central Bank, 2025, pp. 15–26), which limits the relevance of that source to this article.

In its consultation paper, the EBA identified the principal areas in which machine learning might be used: (a) risk differentiation, (b) risk quantification and (c) model validation. It grouped the remaining areas under the heading *“other areas”*. In light of the distinction set out in the preceding paragraph, areas (a) and (b) concern the use of machine learning within the model

(an IRB model based on machine learning), whereas area (c) concerns its use in the validation process (the IRB model need not itself be based on machine learning).

The follow-up report did not identify areas substantially different from those listed at the outset. In relation to validation, it clarified that “*ML might be used [...] for internal validation purposes*” (European Banking Authority, 2023, p. 20), and that machine learning may be used to “*develop challenger models [...] and for benchmarking purposes in the initial as well as ongoing validation*” (European Banking Authority, 2023, p. 7), and to examine generalisation using out-of-sample and out-of-time data samples (European Banking Authority, 2023, p. 8).

PERMISSIBILITY OF USING MACHINE LEARNING

Legislation does not establish an exhaustive list of techniques that may be used in the validation process. Article 185(c) of the CRR refers broadly to “*other validation tools*”, indicating that the list is open-ended. The same approach is taken in Commission Delegated Regulation (EU) 2022/439 of 20 October 2021, supplementing Regulation (EU) No 575/2013 of the European Parliament and of the Council with regard to regulatory technical standards for the specification of the assessment methodology competent authorities are to follow when assessing the compliance of credit institutions and investment firms with the requirements to use the Internal Ratings Based Approach (OJ L 90, 18 March 2022, pp. 1–66), hereinafter “*Regulation 2022/439*”. Article 12(b), which addresses, among other things, the adequacy of validation methods, provides that “*validation methods [...] are appropriate to the nature, degree of complexity and range of application of the institution’s rating systems and the availability of data*”. This provision likewise does not establish a closed list.

This interpretation is consistent with the ECB’s position that “*all relevant validation techniques are expected to be used*” (European Central Bank, 2023, pp. 23–24). The ECB goes further, stating that institutions are expected to seek out state-of-the-art validation techniques (European Central Bank, 2023, p. 25).

USING MACHINE LEARNING TO DEVELOP CHALLENGER MODELS

IRB models, which are validated using, among other things, performance metric samples, are based on historical data. Even when the accuracy of those data is not in dispute, they may fail to reflect current conditions, which may have changed because of shifts in market circumstances (economic change, amendments to the regulatory environment, a lender’s revised credit policy, changes in borrower behaviour, and so forth). Challenger models can help identify this situation. Comparing the results of the IRB model under validation with those of a challenger model makes it possible to assess whether the IRB model adequately reflects the relationships present in the current portfolio. If the challenger model achieves greater predictive accuracy, this may indicate that the IRB model under validation no longer fits current market conditions.

In these circumstances, the challenger model does not replace the IRB model; it serves as a benchmark.

The KNF notes that “*machine-learning methods have higher (usually noticeably higher) discriminatory power than the standard logistic regression method*” (Komisja Nadzoru Finansowego, 2024, p. 70), which supports their usefulness, particularly since “*machine learning is more empirical [...] and accounts for higher-order interactions that are not specified in advance*” (Komisja Nadzoru Finansowego, 2026, p. 62). The ECB’s consultation found that machine learning was used primarily to develop models for calculating PD; it was used less frequently to calculate LGD, EAD, ELBE and CCF (European Central Bank, 2023, p. 6).

USING MACHINE LEARNING FOR DATA MANAGEMENT

A second area in which machine learning may be used is to support data-quality analysis. The preceding paragraph assumed that the statistical data were accurate; here, the assumption is that they contain errors characteristic of statistical databases.

According to Recommendation W, “*for each material model used, the bank should assess the model’s potential exposure to [...] data risk arising from [...] low-quality input data; deficiencies in processes for data acquisition, processing, aggregation and storage; an insufficient sample size; time series of inadequate length; an inappropriate frequency of data feeds to the model; and an insufficient range of available data to describe the course of the phenomenon*” (Komisja Nadzoru Finansowego, 2015, p. 24).

The KNF notes that “*the use of machine-learning techniques [...] enables effective use of data that are not always structured and come from different, sometimes unconventional, sources*” (Komisja Nadzoru Finansowego, 2026, p. 63). In the same document, the KNF also points to a related risk, noting that “*unbalanced datasets [...] In such situations, machine-learning models may not perform correctly*” (Komisja Nadzoru Finansowego, 2026, pp. 70–71).

THE RISK OF INSUFFICIENT EXPLAINABILITY

Under Article 175(4) of the CRR, “*where an institution uses statistical models as part of the rating process, it shall document the methods used in this respect. Such documentation: (a) contains a detailed description of the theory, assumptions, and mathematical and empirical bases*”. In some cases, the additional requirement in Article 3(1)(d) of Regulation 2022/439 must also be met. It provides that documentation of rating systems must enable “*third parties to examine and confirm the functioning of rating systems and, in particular, to examine and confirm that: [...] (i) the documentation of the design of the rating system is sufficiently detailed to enable a third party to understand the reasoning behind all aspects of the rating system*”. Against this background, the question arises whether validation carried out using machine

learning can satisfy the above requirement, referred to below for brevity as the “*requirement of sufficient detail*”.

It is not possible to determine whether machine learning “*enables an understanding of the reasoning behind all aspects of the rating system*” without drawing on computer science. According to Ch. Molnar, explainable models can be understood only with the aid of additional techniques, whereas understanding interpretable models requires only knowledge of their structure. From this perspective, interpretability may be understood as the extent to which a person can understand why a given system produced a given result; alternatively, it may be understood as the extent to which a person can predict the system’s result (Molnar, 2019, p. 15). Machine-learning systems can be classified by their degree of interpretability, from low to high. As M. Krzysztoń accurately observes, “*for systems that are more difficult to interpret, where an understanding of how they work is required, it is necessary to use additional tools from the field of XAI (explainable AI). A standard approach is post-hoc explainability, i.e. after the model has been trained*” (Krzysztoń, 2023, p. 52).

It is worth considering whether the legal requirement mentioned above, that the reasoning behind all aspects of the rating system be understood, should be equated with interpretability or explainability. It is also uncertain whether, and under what conditions, an explanation obtained with the aid of XAI tools would satisfy the requirement of sufficient detail. This issue remains unresolved in the documents analysed from the EBA, ECB and KNF. For example, the KNF states that “*a model is explainable if its internal workings can be understood by people or if it is possible to justify the model’s result*” (Komisja Nadzoru Finansowego, 2026, p. 72). However, the publication issued by the Office of the Polish Financial Supervision Authority (UKNF) does not expressly specify whether “*justifying the model’s result*” may consist in presenting an explanation generated by an XAI tool, or what requirements such an explanation would have to meet.¹

Following the EBA, it is also worth noting that, when developing machine-learning-based systems, “*unnecessary complexity [...] should be avoided if it is not justified by a significant improvement in predictive power*” (European Banking Authority, 2023, p. 21), so as not to impede the assessment of their explainability.

RISK OF REDUCED PREDICTIVE PERFORMANCE

The literature discusses extensively the risks of using machine learning that may lead to system degradation (Szostek et al., 2023, p. 25). However, some vulnerabilities to degradation do not appear to apply to systems used in the validation of IRB models. This is due, among other things, to the fact that these systems are unlikely to attract potential attackers: they play an

¹ Given the article’s length limits, it is not possible to resolve these issues here. The issues identified in this passage will be addressed in further research, the findings of which will be presented in separate publications.

indirect role and do not directly affect credit decisions. They therefore provide little opportunity for an attacker to monetise their activity. Attacks pose a separate set of risks; given the scope of this article, the following analysis is limited to risks arising from the properties of the models themselves.

Predictive performance may deteriorate without interference by third parties. J. Surma provides a useful analysis of this issue, drawing on T. Mitchell's definition: "*a system is able to learn from experience (E) with respect to the performance of tasks (T), as measured by a performance measure (P), if its performance at task T, as measured by P, improves with experience E. As a result of learning, the system is able to generate a correct response (output) for a given input object*" (Surma, 2023, p. 36). Surma identifies potential consequences of a system's faulty operation, including a decline in classifier quality through "*the generation of false-positive or false-negative errors. This type of attack results in a decline in classification accuracy, which implies reduced confidence in the system (confidence reduction) and, in an extreme case, abandonment of its use. This is due, among other things, to the fact that misclassifications generate real and potential costs (e.g. reputational loss) as a consequence of incorrect decisions or their absence*" (Surma, 2023, pp. 36–37).

A classifier's quality can deteriorate without human intervention. The literature highlights two causes in particular. First, "*data drift*" is a change in the distribution of input data that need not involve a change in the relationship between the input data and the target variable (Gama et al., 2014, p. 44:4). In other words, a model trained on historical data (such as the data used in a performance metric sample) will not perform equally well on a different dataset. Second, "*concept drift*" is a change in the relationship between the input variables and the target variable (Gama et al., 2014, p. 44:4). In the validation of IRB models, this may mean that an estimated parameter has a different predictive meaning under current conditions than it did in the past. For example, a high loan amount may historically have indicated a higher probability of default, whereas today it may reflect elevated inflation and changes in interest rates.

CONCLUSIONS

The analysis at the intersection of legal scholarship and computer science leads to the following conclusions about the scope of machine learning use in the validation of IRB models.

- 1) Analysis of the CRR and Regulation 2022/439, taking account of the positions of the EBA, ECB and KNF, indicates that machine learning may be used as one of the tools in the validation of IRB models.
- 2) The current state of computer science does not provide grounds for concluding that machine learning can, by itself, achieve all the objectives of IRB model validation. Machine-learning models are themselves subject to limitations that affect the reliability of their results.
- 3) Taken together, the normative and computer-science perspectives lead to the conclusion that machine learning cannot serve as a functional substitute for validation.

The analysis also brought to light further issues, allowing at least two research questions to be formulated:

- 1) Whether, and if so under what conditions, an explanation obtained with the aid of XAI tools will satisfy the requirement of sufficient detail referred to in Article 3(1)(d) of Regulation 2022/439?
- 2) Can “*justifying the model’s result*”, as used in the UKNF’s position, consist in presenting an explanation generated by an XAI tool?

Finally, it is worth emphasising the sound guidance in Recommendation W: “*when shaping the validation process, it should be remembered that there is no single universal approach to validation that will always fit the specific characteristics of a bank and the range of models it uses. The way in which tasks arising from the validation methodology (and, at times, tasks extending beyond its strictly specified scope) are carried out should therefore be adjusted flexibly, on a case-by-case basis, to the characteristics of the model*” (Komisja Nadzoru Finansowego, 2015, p. 40). Accordingly, the scope of machine learning use should also be determined on a case-by-case basis, taking account of the objectives of validation, the characteristics of the model, and the specific features of the bank.

REFERENCES

- European Banking Authority (2021). ‘Discussion Paper on Machine Learning for IRB Models’. Available at: https://www.eba.europa.eu/sites/default/files/document_library/Publications/Discussions/2022/Discussion%20on%20machine%20learning%20for%20IRB%20models/1023883/Discussion%20paper%20on%20machine%20learning%20for%20IRB%20models.pdf
- European Banking Authority (2023). ‘Machine Learning for IRB Models. Follow-up Report from the Consultation on the Discussion Paper on Machine Learning for IRB Models’. Available at: https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061483/Follow-up%20report%20on%20machine%20learning%20for%20IRB%20models.pdf
- European Central Bank (2023). ‘Supervisory Handbook on the Validation of IRB Rating Systems Under the Internal Ratings Based Approach’. Available at: https://www.eba.europa.eu/sites/default/files/document_library/Publications/Reports/2023/1061495/Supervisory%20handbook%20on%20the%20validation%20of%20IRB%20rating%20systems%20revised.pdf
- European Central Bank (2025). ‘ECB guide to internal models’. Available at: https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide202507.en.pdf
- Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). ‘A Survey on Concept Drift Adaptation’.
- Iwanicz-Drozdowska, M. (2024). ‘Wprowadzenie do zagadnień ryzyka bankowego’ [Introduction to Banking Risk]. [In:] M. Iwanicz-Drozdowska (ed.), ‘Zarządzanie ryzykiem bankowym’ [Bank Risk Management].

Iwanicz-Drozdowska, M. and Chmielewski, T. (2024). 'Kluczowe hasła' [Key Terms]. [In:] M. Iwanicz-Drozdowska (ed.), 'Zarządzanie ryzykiem' [Risk Management].

Komisja Nadzoru Finansowego (2015). 'Rekomendacja W dotycząca zarządzania ryzykiem modeli w bankach' [Recommendation W on Model Risk Management in Banks]. Available at: https://www.knf.gov.pl/knf/pl/komponenty/img/knf_137749_Rekomendacja_W_42219.pdf

Komisja Nadzoru Finansowego (2024). 'Zarządzanie ryzykiem modeli w zakresie działalności podmiotów sektora bankowego ze szczególnym uwzględnieniem modeli ryzyka kredytowego – podstawowe zagadnienia' [Model Risk Management in the Activities of Entities in the Banking Sector, with Particular Emphasis on Credit Risk Models: Key Issues]. Available at: https://www.knf.gov.pl/knf/pl/komponenty/img/publikacja_zarządzanie_ryzykiem_modeli_92496.pdf

Komisja Nadzoru Finansowego (2026). 'Rekomendacje dla podmiotów rynku finansowego dotyczące wpływu modeli Frontier AI na cyberbezpieczeństwo' [Recommendations for Financial Market Entities on the Impact of Frontier AI Models on Cybersecurity]. Available at: https://www.knf.gov.pl/knf/pl/komponenty/img/Rekomendacje_dla_podmiotow_rynku_finansowego_dot_wplywu_modeli_Frontier_AI_na_cyberbezpieczenstwo_99167.pdf

Krzysztoń, M. (2023). 'Weryfikacja wiarygodności systemów w erze uczenia maszynowego'. [In:] NASK *Cyberbezpieczeństwo w AI. AI w Cyberbezpieczeństwie*. Available at: <https://cyberpolicy.nask.pl/wp-content/uploads/2023/09/Cyberbezpieczenstwo-AI.-AI-w-cyberbezpieczenstwie.pdf>

Matuszyk, A. (2008). Credit scoring.

Matuszyk, A. and Ptak-Chmielewska, A. (2024). 'Ryzyko kredytowe' [Credit Risk]. [In:] M. Iwanicz-Drozdowska (ed.), 'Zarządzanie ryzykiem bankowym' [Bank Risk Management].

Molnar, Ch. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. Available at: https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf

Niedziółka, P. (2018). 'Pojęcie i rodzaje ryzyka bankowego' [The Concept and Types of Banking Risk]. [In:] M. Zalewska (ed.), 'Świat bankowości' [The World of Banking].

Commission Delegated Regulation (EU) 2022/439 of 20 October 2021 supplementing Regulation (EU) No 575/2013 of the European Parliament and of the Council with regard to regulatory technical standards for the specification of the assessment methodology competent authorities are to follow when assessing the compliance of credit institutions and investment firms with the requirements to use the Internal Ratings Based Approach, OJ L 90, 18 March 2022, pp. 1–66.

Regulation (EU) No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012, OJ L 176, 27 June 2013, pp. 1–337, as amended.

Schab, I. (2005). 'Ocena ryzyka kredytowego w ramach wewnętrznych systemów ratingowych – charakterystyka podejścia oraz podstawowych wymogów' [Credit Risk Assessment within Internal Rating Systems: A Description of the Approach and Basic Requirements], Bezpieczny Bank [Preprint], (1).

Surma, J. (2023). 'Wprowadzenie do ataków na systemy uczenia maszynowego' [Introduction to Attacks on Machine-Learning Systems]. [In:] 'Cyberbezpieczeństwo AI. AI w cyberbezpiec-

zeństwie' [AI Cybersecurity. AI in Cybersecurity]. Available at: <https://cyberpolicy.nask.pl/wp-content/uploads/2023/09/Cyberbezpieczenstwo-AI.-AI-w-cyberbezpieczenstwie.pdf>

Szostek, D., Kasprowski, P., Kozak, J., Kapczyński, A. and Prabucki, R. (2023). 'Wyzwania i zagrożenia z zakresu cyberbezpieczeństwa podczas projektowania lub wykorzystywania AI' [Cybersecurity Challenges and Threats in the Design or Use of AI]. [In:] CyberPolicy NASK (ed.), 'Cyberbezpieczeństwo AI. AI w cyberbezpieczeństwie' [AI Cybersecurity. AI in Cybersecurity]. Available at: <https://cyberpolicy.nask.pl/wp-content/uploads/2023/09/Cyberbezpieczenstwo-AI.-AI-w-cyberbezpieczenstwie.pdf>

Zawadzka, Z. (2001). 'Ryzyko bankowe' [Banking Risk], Bank [Preprint], (6).