

Predictors of distant metastasis in colorectal cancer: a multimodal statistical and machine-learning analysis

Wojciech Lewitowicz¹, Monika Kozłowska-Geller¹, Monika Wawszczak-Kasza^{1,2}, Agnieszka Plusa^{1,2}, Marcin Zaremba^{1,2}, Karol Romaszko¹, Piotr Lewitowicz^{1,2}

¹Collegium Medicum, Jan Kochanowski University, Kielce, Poland

²Meduniv Sp. z o.o., Kielce, Poland

Medical Studies 2025; 41 (4): 344–349

DOI: <https://doi.org/10.5114/ms.2025.158295>

Key words: colorectal cancer, metastasis, TP53, NRAS, logistic regression, random forest, LASSO, SVM, TRIPOD.

Abstract

Introduction: Distant metastasis (M1) is the principal determinant of outcome in colorectal cancer (CRC). Identifying robust molecular and clinical predictors at diagnosis may improve risk stratification.

Aim of the research: To develop and compare complementary statistical and machine-learning models for predicting metastatic status and to synthesize convergent predictors.

Material and methods: In a single-centre cohort ($N = 54$, $M1 = 26$, $M0 = 28$), we modelled a binary outcome – any type of CRC metastasis (M_{any}) using multivariable logistic regression (LR), Random Forest (RF), LASSO-penalized logistic regression, and Support Vector Machine (SVM). In a secondary analysis, we fitted a multinomial logistic regression with three categories Early_No_Mets (pT1-2M0), Advanced_No_Mets (pT3-4M0), and Metastatic (M1). Discrimination was summarized as AUC; classification metrics used native validation schemes (LR in-sample; RF out-of-bag [OOB]; SVM 10-fold cross-validation [CV]). Reporting follows TRIPOD guidance.

Results: LR identified *TP53* mutation as the strongest predictor (OR = 3.47; 95% CI: 0.92–14.5; $p = 0.073$; AUC = 0.713). RF achieved OOB error of 40.74% (accuracy = 59.3%); top features were *NRAS*, *TP53_pathway*, *TP53*, *WNT_pathway*, and *n_genes*. LASSO (10-fold CV) retained *NRAS*, *TP53*, and age (coefficients +1.15, +0.85, –0.02). SVM yielded AUC = 0.678, accuracy = 55.6%. The multinomial model revealed complete separation for *NRAS* in the Advanced_No_Mets group, precluding standard MLE inference.

Conclusions: *TP53* (gene/pathway) is a consistent risk signal across methods; *NRAS* carries high importance in ensemble/regularized models. Overall discrimination is modest, consistent with a small sample size; findings are hypothesis-generating and warrant validation.

Introduction

The incidence of early-onset colorectal cancer diagnosed in patients under the age of 50 years has been increasing around the world. The clinical and pathological features, genetic and epigenetic landscapes, and emerging data on the associated clinical risk factors are currently examined [1].

Distant metastasis at presentation is the dominant driver of prognosis in colorectal cancer (CRC). Globally, CRC remains a high-burden malignancy with substantial incidence and mortality; contemporary epidemiology underscores the clinical need for early risk stratification [2].

Targeted next-generation sequencing (NGS) enables high-depth, multi-gene profiling focused on clinically actionable hotspots at a fraction of the cost of whole-exome/genome sequencing, with improved sensitivity for low-frequency variants [3, 4]. In parallel, systematic screening and molecular work-up can support earlier detection and biologically informed

care pathways that are associated at a population level with stage shift and improved outcomes.

Artificial intelligence (AI) algorithms have made significant progress in the medical field. Their widespread use in diagnosing and treating various types of cancer, particularly colorectal cancer (CRC) is gaining substantial attention. CRC, the third most diagnosed malignancy in both men and women, remains a leading cause of cancer-related deaths worldwide [5]. In this study, we analysed a single-centre cohort profiled by a 50-gene hotspot panel and developed complementary statistical and machine-learning models to identify predictors of distant metastasis at diagnosis.

Aim of the research

To compare classical multivariable regression with ensemble and kernel methods and to synthesize convergent predictors of M1 status at diagnosis, using a unified dataset and reporting aligned with TRIPOD.

Material and methods

Cohort and outcome

We analysed $N = 54$ patients with histologically confirmed colorectal adenocarcinoma (NOS). The binary endpoint was M_any (0 = no distant metastasis; 1 = distant metastasis at diagnosis). Class counts: $M0 = 28$ (51.9%), $M1 = 26$ (48.1%).

Inclusion and exclusion criteria

Patients with confirmed colorectal adenocarcinoma (NOS) were included. Other histological subtypes were excluded due to their distinct molecular pathways. Moreover, all immunohistochemistry confirmed microsatellite unstable tumours were rejected. Additional exclusion criteria included prior radiotherapy or chemotherapy. Only patients with DNA of sufficient quality for next-generation sequencing (NGS) were eligible.

Age handling: Age was collected for all patients; primary analyses modelled age as a continuous predictor. Age strata (≤ 50 vs. > 50 years) were used only for descriptive summaries and did not drive model specification.

Genomic profiling (NGS workflow)

Tumour genomic DNA was extracted from FFPE using MagCore kit according to the manufacturer's protocol. Libraries were prepared with AmpliSeq Library PLUS for Illumina (Cancer Hotspot Panel v2), targeting 50 oncogenes/tumor suppressors (~2,800 hotspot variants; 207 amplicons). Indexed libraries were quantified (Quantus®/QuantiFluor®), pooled equimolarly, denatured and diluted per Illumina guidance, and spiked with 5% PhiX. Sequencing was performed on MiSeq Dx with MiSeq Reagent Micro Kit v2 (300-cycle). Standard demultiplexing and variant calling pipelines were applied; downstream variables included gene-level mutation flags, pathway indicators, and aggregate mutation counts used as predictors in the models.

Predictors

Clinical: age (years). Genes: *TP53*, *KRAS_any* (any type of mutation), *NRAS_any* (any type of mutation). Cancer pathway metrics: *n_mutations* (number of mutations), *n_genes* (number of genes), *n_genes_topN* (number of genes with any positive N), hotspot count, truncating_burden, pathway indicators: WNT_pathway, RAS, MAPK, pathway, PI3K pathway, *TP53* pathway, TGF- β pathway, and *n_pathways_affected*.

Models and validation

- Logistic regression (LR): prespecified predictors (age, *TP53*, *KRAS_any*, *n_pathways_affected*, *n_genes_topN*); discrimination summarized as in-sample AUC.

- Random Forest (RF): 500 trees (default *mtry*=4) using 17 predictors; performance via OOB error; importance by Mean Decrease in Accuracy (MDA) and Gini [6].
- LASSO logistic regression: L1-penalized model with 10-fold CV (λ_{1se}). Predictors standardized (Z-score) before penalization [7].
- Multinomial logistic regression: three categories—Early_No_Mets, Advanced_No_Mets, Metastatic; separation diagnostics performed [8, 9].
- SVM (RBF): 10-fold CV; tuning grid for C and γ ; importance estimated via caret permutation [10].

Statistical analysis and software

We summarize accuracy, sensitivity, specificity from each model's native validation (LR in-sample; RF OOB; SVM CV). Analyses were performed in R (random forest, glmnet, caret).

Performance metrics and statistical computations

For binary classification we additionally report: PPV (precision), NPV, F1-score, balanced accuracy ($(\text{sensitivity} + \text{specificity})/2$), Youden's J ($\text{sensitivity} + \text{specificity} - 1$), Matthews correlation coefficient (MCC), and Cohen's κ . [11–13] 95% CI for AUC in LR were computed using the Hanley–McNeil method with $n_{\{pos\}}=26$ and $n_{\{neg\}}=28$ [14]. For a fixed threshold (0.5) in LR we derived likelihood ratios ($LR+ = \text{sensitivity}/(1 - \text{specificity})$; $LR- = (1 - \text{sensitivity})/\text{specificity}$) and diagnostic odds ratio ($DOR = LR+/LR-$).

Results

Logistic regression (baseline specification)

Model specification is presented in Table 1. We fitted a multivariable logistic model with prespecified predictors (age [years], *TP53*, *KRAS_any*, *n_pathways_affected*, *n_genes_topN*). The dependent variable was *M_any* (0/1).

Point estimates. *TP53* exhibited the strongest association with metastatic presentation: OR = 3.47 (95% CI: 0.92–14.5; $p = 0.073$). Effect sizes for the remaining covariates were small and not statistically significant: age OR = 0.97 (0.92–1.01), *KRAS_any* OR = 0.81 (0.19–3.47), *n_pathways_affected* OR = 1.30 (0.49–3.53), *n_genes_topN* OR = 0.60 (0.23–1.45). The 95% CI for *TP53* was closest to excluding 1 (Figure 1), suggesting a borderline adverse association consistent with biological plausibility, albeit underpowered at $N = 54$.

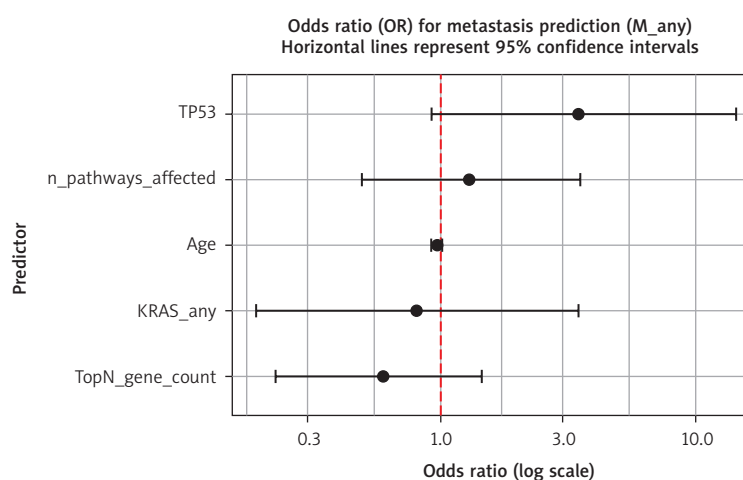
Discrimination. The model's AUC (in-sample) was 0.713 (95% CI: 0.574–0.852, Hanley–McNeil), indicating acceptable but not high discrimination.

Classification at 0.5 threshold. Accuracy 64.8% (35/54); sensitivity 65.4% (17/26); specificity 64.3% (18/28); PPV 0.630 (17/27); NPV 0.667 (18/27); F1 0.642;

Table 1. Multivariable logistic regression coefficients (baseline specification)

Predictor	β (SE)	OR	95% CI	P-value
<i>TP53</i>	1.244 (0.703)	3.47	0.92–14.50	0.073
Age (per year)	−0.030 (0.024)	0.97	0.92–1.01	0.193
<i>KRAS_any</i>	−0.211 (0.741)	0.81	0.19–3.47	0.767
<i>n_pathways_affected</i>	0.262 (0.504)	1.30	0.49–3.53	0.595
<i>n_genes_topN</i>	−0.511 (0.470)	0.60	0.23–1.45	0.266

Notes: β and SE are on the log-odds scale; SE was estimated from the reported 95% CI (normal approximation). Age effect is per 1-year increase.

**Figure 1.** Logistic regression – odds ratios with 95% CI (TP53, age, KRAS_any, n_pathways_affected, n_genes_topN)

balanced accuracy 0.648; Youden's J 0.297; MCC 0.296; Cohen's κ 0.296. Likelihood ratios: LR+ 1.83, LR− 0.54, DOR 3.40. These values indicate modest diagnostic separation; the LR pair suggests only small changes in post-test probability at this cut point.

Interpretation and caveats. With class sizes balanced ($M1 = 26$, $M0 = 28$), the borderline *TP53* signal and moderate AUC likely reflect limited sample size and wide CIs rather than absence of effect. No calibration assessment is presented; external validation is required before clinical translation.

Random Forest (OOB)

Performance. OOB error 40.74% corresponds to accuracy 59.3%. The OOB confusion matrix (TN = 18, FP = 10, FN = 12, TP = 14) yields sensitivity 0.538, specificity 0.643, PPV = 0.583, NPV = 0.600, F1 = 0.560, balanced accuracy 0.591, Youden's J = 0.181, MCC = 0.182, κ = 0.182 (Table 1). The modest gain over chance is consistent with small-N, multi-feature settings.

Variable importance. Rank-based importance (MDA/Gini) highlighted: 1) *NRAS_any*, 2) *TP53_pathway*, 3) *TP53*, 4) *WNT_pathway*, 5) *n_genes* (Figure 2). The prominence of *NRAS_any* – not included in the baseline LR – suggests non-linear effects and/or

interactions that are better captured by tree ensembles. Importance measures reflect predictive contribution and do not imply direction of effect.

Notes. OOB validation affords an internal performance estimate without a holdout set; however, OOB AUC was not computed here, and ranking can be affected by correlation among predictors.

LASSO-penalized logistic regression

Using 10-fold CV with the lambda.1se criterion, the final sparse model retained three predictors with non-zero standardized coefficients: *NRAS_any* ($\beta \approx +1.15$), *TP53* ($\beta \approx +0.85$), age ($\beta \approx -0.02$). Selection of *NRAS_any* alongside *TP53* corroborates the RF ranking and indicates that, under regularization, both markers contribute unique predictive signal. The small negative coefficient for age implies a weak inverse relationship after standardization, but the magnitude is modest and should not be overinterpreted.

Multinomial logistic regression

No predictor reached conventional significance across category comparisons (Advanced_No_Mets vs Early_No_Mets; Metastatic vs Early_No_Mets). For

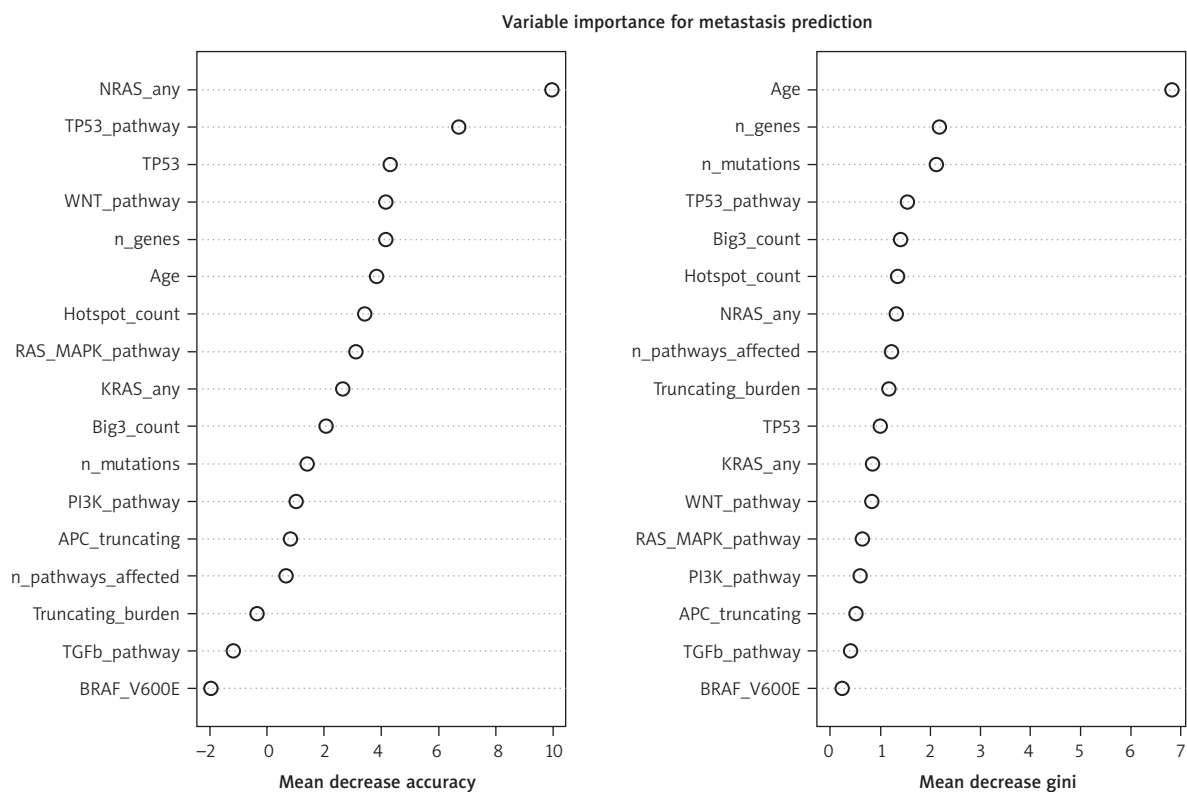


Figure 2. Random forest – feature importance (rank; 1 = highest): *NRAS_any*, *TP53_pathway*, *TP53*, *WNT_pathway*, *n_genes*

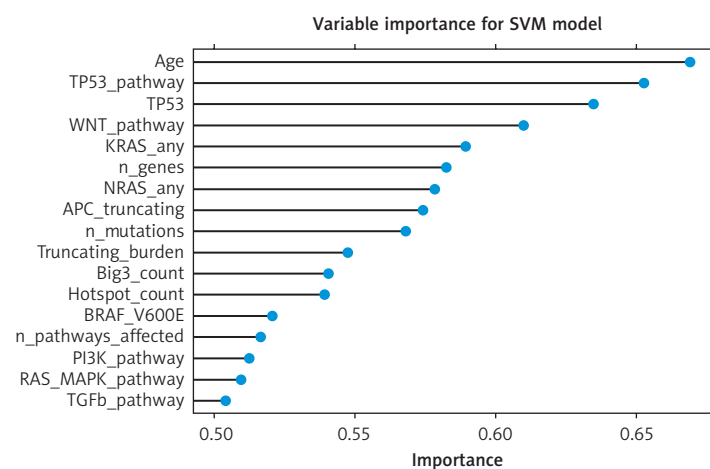


Figure 3. SVM (RBF) – Feature Importance (rank; 1 = highest): age, *TP53_pathway*, *TP53*, *WNT_pathway*, *KRAS_any*

NRAS_any in Advanced_No_Mets, the analysis exhibited complete separation (no events among cases), rendering standard MLE coefficients/*p*-values non-estimable. Penalized likelihood (e.g., Firth correction) or category consolidation would be appropriate remedies but were outside the present scope.

SVM (RBF; cross-validated)

Discrimination. AUC = 0.678 with accuracy = 55.6%, sensitivity = 50.0%, specificity = 61.7%, balanced accuracy = 0.559. The kernelized margin did not materially outperform LR, suggesting either limit-

ed non-linearity or insufficient sample size to capitalize on RBF flexibility. Variable importance prioritized age, followed by *TP53_pathway* and *TP53* (Figure 3).

Notes. Threshold-dependent metrics (PPV/NPV/F1) are not reported for CV due to the lack of a single pooled confusion matrix. Hyperparameters *C* and γ were tuned on a predefined grid.

Cross-model synthesis

TP53 (gene/pathway) emerges as a consistent risk signal across modelling families (borderline in LR; high importance in RF and SVM; retained by LASSO).

Table 2. Model performance metrics (rounded to 3 decimals)

Model	Validation	AUC (95% CI)	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	Balanced Acc	Youden J	MCC	Kappa	LR+	LR-	DOR
Logistic Regression	In-sample	0.713	0.65	0.654	0.643	0.630	0.667	0.642	0.648	0.297	0.296	0.296	1.831	0.538	3.400
Random Forest	OOB	–	0.593	0.538	0.643	0.583	0.600	0.560	0.591	0.181	0.182	0.182	1.508	0.718	2.100
SVM (RBF)*	10-fold CV	0.678 (–)	0.556	0.500	0.617	–	–	–	0.559	0.117	–	–	1.305	0.810	–

*NRAS*_any shows high importance in RF and is selected by LASSO, indicating added predictive value not captured in the baseline LR specification; its rank is lower in SVM.

Age is non-significant in LR but ranks highest in SVM and carries a small negative LASSO coefficient, suggesting a weak trend contingent on model class/regularization.

*KRAS*_any does not exhibit compelling independent contribution once other variables are considered.

Synopsis. A pragmatic two-feature signature (*TP53*, *NRAS*_any) recurs across methods, with age as an auxiliary factor. Overall discrimination is moderate, consistent with $N = 54$ and the breadth of predictors. These findings are hypothesis-generating and motivate harmonized cross-validation and external validation.

Performance summary

Performance summary was presented in Table 2.

Discussion

Our convergent findings suggest that *TP53* mutation status and the presence of *NRAS* mutation may support risk stratification for distant metastasis at the time of diagnosis. A pragmatic, testable framework is to treat *TP53*-mutant and/or *NRAS*-mutant tumours as high risk within a simple, multivariable score that also includes age. In practice after external validation and calibration – such a score could be used to: prioritize more intensive baseline staging and closer early postoperative surveillance for patients above a prespecified risk threshold; inform multidisciplinary decision-making in perioperative planning when occult dissemination is a concern; and finally select patients for clinical trials evaluating biologically directed treatment escalation/de-escalation strategies.

Prediction models are developed to help health care providers in estimating the clinical outcome (diagnostic models) or that a specific event will occur in the future (prognostic models). However, much evidence unveiled that the quality of reporting of prediction model studies is poor. Only with full and clear reporting of information on all aspects of a prediction model can risk of bias and potential usefulness of prediction models be adequately assessed. The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Initiative developed a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes [15–19].

We emphasize, however, that our models show moderate discrimination and were derived from $N = 54$; therefore, these markers should not be used in isolation to guide therapy. Independent validation, calibration assessment, and decision-curve analysis are required before clinical deployment.

Across complementary methods, *TP53* emerges as a robust marker associated with metastatic presentation, while *NRAS* any contributes strongly in non-linear/regularized settings. Discrimination (AUC: 0.68–0.71) is modest, which is expected for $N = 54$ with heterogeneous genomic inputs and balanced classes.

The multinomial model diagnosed separation, highlighting the importance of penalized likelihood in small samples. Method choices and validation schemes shape apparent performance; therefore, we emphasize consilience of signals over direct metric comparisons.

Limitations: Small, single-centre cohort; different validation paradigms across models (in-sample vs OOB vs CV); potential overfitting despite penalization; separation affecting multinomial estimates.

Conclusions

A minimal signature comprising *TP53* and *NRAS* any (with age as an auxiliary factor) recurs across independent modelling families. While predictive ability is moderate, these findings are hypothesis-generating and motivate external validation and calibration assessment.

Data Access Statement

De-identified data underlying this analysis are available from the corresponding author upon reasonable request.

Author contributions

Conceptualization: Wojciech Lewitowicz; data curation: Agnieszka Plusa, Marcin Zaremba, formal analysis: Monika Kozłowska-Geller; investigation: Monika Wawszczak-Kasza; methodology: Piotr Lewitowicz; project administration: Piotr Lewitowicz; resources: Monika Wawszczak-Kasza, Wojciech Lewitowicz; Karol Romaszko; supervision: Piotr Lewitowicz; writing – review & editing: Piotr Lewitowicz

Funding

No external funding.

Ethical approval

All procedures conformed to institutional/national research committee standards and the Declaration of Helsinki (1975) and its amendments.

Conflict of interest

The authors declare no conflict of interest.

References

- Patel SG, Karlitz JJ, Yen T, Lieu CH, Boland CR. The rising tide of early-onset colorectal cancer: a comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early detection. *Lancet Gastroenterol Hepatol*. 2022; 7(3): 262-274.
- Morgan E, Arnold M, Gini A, Lorenzoni V, Cabaasag CJ, La-versanne M, Vignat J, Ferlay J, Murphy N, Bray F. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimate from GLOBOCAN. *Gut*. 2023; 72(2): 338-344.
- Del Vecchio F, Mastroiaco V, Di Marco A, Compagnoni C, Capece D, Zazzeroni F, Capalbo C, Alesse E, Tessitore A. Next-generation sequencing applications in colorectal cancer. *J Transl Med*. 2017; 15: 246.
- Hussen BM, Abdullah ST, Salihi A, Khdr Sabir D, Sidiq KR, Rasul MF Hidayat HJ, Ghafouri-Fard S, Taheri M, Jamali E. Emerging roles of NGS in clinical oncology and personalized medicine. *Pathol Res Pract*. 2022; 230: 153760.
- Spaander MCW, Zauber AG, Syngal S, Blaser MJ, Sung JJ, You YN, Kuipers EJ. Young-onset colorectal cancer. *Nat Rev Dis Primers*. 2023; 9: 21.
- Mitsala A, Tsalikidis C, Pitiakoudis M, Simopoulos C, Tsaroucha AK. Artificial intelligence in colorectal cancer screening, diagnosis and treatment. A new era. *Curr Oncol*. 2021; 28(3): 1581-1607.
- Breiman L. Random forests. *Mach Learn*. 2001; 45: 5-32.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Statist Soc B*. 1996; 58: 267-288.
- Albert A, Anderson JA. On the existence of MLE in logistic regression. *Biometrika* 1984; 71: 1–10.
- Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993; 80: 27-38.
- Cortes C, Vapnik V. Support-vector networks. *Mach Learn*. 1995; 20: 273-297.
- Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975; 405: 442-451.
- Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960; 20: 37-46.
- Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950; 3: 32-35.
- Hanley JA, McNeil BJ. The meaning and use of the area under a ROC curve. *Radiology*. 1982; 143: 29-36.
- Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950; 3: 32-35.
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *BMC Med*. 2015; 13: 1.
- Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD Statement. *Ann Intern Med*. 2015; 162(1): 55-63.
- Chen K, Qu Y, Han Y, Li Y, Gao H, Zheng D. Performance of machine learning in diagnosing KRAS (Kirsten rat sarcoma) mutations in colorectal cancer: systematic review and meta-analysis. *J Med Internet Res*. 2025; 27: e73528.

Address for correspondence

Piotr Lewitowicz
Collegium Medicum
Jan Kochanowski University
Meduniv Sp.z.o.o
Kielce, Poland
E-mail: piotr.lewitowicz@ujk.edu.pl

Received: 4.10.2025

Accepted: 25.11.2025

Online publication: 30.12.2025