BMJ Health & Care Informatics

Machine learning prediction of germline *BRCA1/2* pathogenic variants in patients with ovarian cancer

Giovanni Innella ,[1,2] Giulia Erini,[2] Antonio De Leo,[3,4] Lea Godino,[1] Luca Caramanna,[2] Simona Ferrari,[1] Sara Miccoli,[1] Anna Myriam Perrone,[2,5] Claudio Zamagni,[6] Pierandrea De Iaco,[2,5] Daniela Turchetti,[1,2] Paola Rucci[7]

To cite: Innella G, Erini G, De Leo A, *et al*. Machine learning prediction of germline *BRCA1/2* pathogenic variants in patients with ovarian cancer. *BMJ Health Care Inform* 2025;**32**:e101751. doi:10.1136/bmjhci-2025-101751

Received 06 August 2025
Accepted 14 December 2025

[1]Medical Genetics Unit, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy
[2]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy
[3]Department of Experimental, Diagnostics and Specialty Medicine, University of Bologna, Bologna, Italy
[4]IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy
[5]Division of Oncologic Gynecology, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy
[6]Breast and Gynecological Medical Oncology, IRCCS Azienda Ospedaliero-Universitaria di Bologna, Bologna, Italy
[7]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy

Correspondence to
Dr Daniela Turchetti;
daniela.turchetti@unibo.it

ABSTRACT

Objectives To assess the performance of machine learning (ML) algorithms to predict the presence of germline *BRCA1/2* pathogenic variants in ovarian cancer (OC) patients based on clinical–pathological features.

Methods Clinical–pathological features of 648 patients with OC tested for *BRCA1/2* were analysed using three supervised ML algorithms: random forest, boosting and support vector machine.

Results In the 'test' sample, boosting proved to be the most effective algorithm (accuracy: 84.5%; precision: 80.0%; recall: 3.1%; area under the curve (AUC): 78.8%), followed by support vector machine (accuracy: 81.4%; precision: 72.7%; recall: 27.6%; AUC: 62.3%) and random forest (accuracy: 74.4%; precision: 55.6%; recall: 14.7%; AUC: 71.3%). In the 'validation' sample, accuracy was 79.8% for boosting, 81.7% for support vector machine, 80.8% for random forest.

In the most effective algorithm (boosting), family history of OC showed the highest relative influence (52.9), followed by histotype (19.5), personal history of breast cancer (BC) (17.1), age at diagnosis (8.4) and family history of BC (2.2), while Federation of Gynecology and Obstetrics stage had no influence.

Discussion We identified the predictive algorithm that best estimates the a priori likelihood of being a carrier of germline *BRCA1/2* pathogenic variants in patients with OC. These findings support a role for ML approaches in predicting *BRCA1/2* status in patients with OC, but accuracy and precision are still suboptimal for clinical use, suggesting the need for additional research.

Conclusions Results support the selection of relevant clinical features for predictive purposes, which could have significant implications for the clinical management of patients with OC.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Machine learning (ML) models have been used to predict development and clinical outcomes of ovarian cancer (OC), but never the presence of germline *BRCA1/2* pathogenic variants.

WHAT THIS STUDY ADDS

⇒ Here, we assessed the performance of three ML algorithms in predicting *BRCA* status in patients with OC and boosting demonstrated the best performance, followed by support vector machine and random forest.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Although accuracy and precision of our method did not exceed 90%, raising the need for additional refinements, our findings are encouraging and pave the way for future research.

INTRODUCTION

Machine learning (ML) is gaining an increasingly important role in ovarian cancer (OC) research and has a considerable potential for clinical translation; in particular, ML models have demonstrated a potential to facilitate earlier and more precise interventions throughout the care process. However, their extensive implementation in real-world contexts still necessitates additional prospective validation.[1 2]

Regarding early detection, ML models have shown greater accuracy in combining biomarkers, clinical variables and imaging data, providing support for diagnosis and risk-based triage.[1 2] ML is also being used to refine tumour subtype classification by analysing genomic and epigenetic profiles, helping to identify biologically distinct subgroups with different therapeutic responses.[3] Regarding prognosis and survival prediction, ML algorithms have been employed to predict progression-free survival and recurrence risk by combining clinical, pathological and molecular data.[4] These models could support treatment planning and follow-up strategies.

In the context of personalised treatment, ML has been used to predict responses to chemotherapy and targeted therapies (based on multiomics and clinical data).[5 6] Germline pathogenic variants of the *BRCA1/2* genes are strongly associated with the risk of developing OC and with response to treatment.

We previously developed a logistic regression model to predict the presence of germline *BRCA1/2* pathogenic variants in patients with OC based on their clinical–pathological characteristics,[7] with the aim of providing clinicians involved in patients with OC care with a weighted risk score of *BRCA1/2* variants based on available clinical–pathological parameters. The proposed risk score yet requires clinical validation and further improvements.

Since ML may represent a valuable approach for *BRCA1/2* prediction, as suggested by Liu and Wu,[8] in the present study we applied ML algorithms to the set of independent variables used in Innella *et al*[7] to assess their performance in predicting *BRCA1/2* variants and to validate the relevance of individual variants considered in the previous risk score.

MATERIALS AND METHODS

Patient sample

The study population was previously described.[7] Briefly, it includes all the patients with OC tested for *BRCA1/2* from 2012 to 2022 at the Bologna hub of the Emilia-Romagna Network; for each patient, information on the following clinical–pathological features was collected: age at diagnosis, Federation of Gynecology and Obstetrics (FIGO) stage, tumour grade, tumour histotype, presence/absence of a family history of breast cancer (BC) and OC in first-degree/second-degree relatives (considered together) and a personal history of BC.[9]

Statistical analysis

Categorical variables were compared between carriers and non-carriers of germline *BRCA1/2* pathogenic variants using the χ^2 test, and mean age was compared between groups using the t-test.

ML classification algorithms were then used to predict the probability of being a *BRCA1/2* germline pathogenic variant carrier as a function of the clinical–pathological features listed above, randomly partitioning the cohort into training (65%), validation (15%) and testing (20%) samples. The performance of the models was compared using the following evaluation metrics: accuracy (% of correctly classified cases), precision (positive predictive value), recall (true positive rate) and area under the curve. All these indices range from 0 to 1, with higher values denoting better performance.

The relative influence score was used to denote the predictive ability of the variables.

The ML algorithms used were:

Random forest: this method of classification creates a set of decision trees that consists of a large number of individual trees, which operate as an ensemble. Each individual tree in the random forest returns a class prediction and the class with the most votes becomes the model's prediction.

Boosting classification: this algorithm sequentially adds features to a decision tree ensemble, each one correcting its predecessor. However, instead of changing the weights for every incorrectly classified observation at every iteration, the boosting method tries to fit the new feature to the residual errors made by the previous feature.

Support vector machine classification: this supervised learning algorithm maps training examples to points in space so as to maximise the width of the gap between the two categories (*BRCA1/2*, no *BRCA1/2*). New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

The random forest, boosting classification and support vector machine classification algorithms partition the study population into training, testing and validation samples to generate robust results. Statistical analyses were conducted using SPSS, V.28.0 and JASP, V.0.18.

RESULTS

Patient characteristics

The study population included 648 patients with complete data for the features of interest out of the 1009 in the database.[7] Although selection bias cannot be ruled out, the pattern of missing data and the absence of significant differences between patients with complete and incomplete data mitigate this issue. In fact, the missing data were sparse (23.9% of cases had missing data on just one variable and 5.6% had missing data on two variables), and the variable with the most missing values was stage, which was excluded from the predictive score based on evidence from our logistic regression model.[7] 216 patients (21.4%) were carriers of a *BRCA1/2* germline pathogenic variant. The characteristics of the study population are shown in table 1. All variables that differed significantly between the two groups were included in the ML classification algorithms.

ML algorithms

For all the three ML algorithms, data were partitioned into training/validation/testing samples (n=415, n=104 and n=129). Comparing the different metrics of the prediction methods related to ML classification (table 2), boosting proved to be the best in the test sample, with an accuracy of 84.5% and a precision of 80%, while it showed a slightly poorer accuracy than the others in the validation sample.

We investigated the importance of the variables used as inputs in our algorithms. The feature importance is calculated by measuring the number of times a variable is used to split the data across all trees. This approach reflects the relative importance of each feature in the construction of the boosting, while for the random forest algorithm the feature importance is expressed as mean decrease in accuracy (ie, the change in model accuracy after permuting the feature values) and total increase in node purity (ie, the total reduction in impurity brought by a feature across all trees). The relative influence for the boosting classification is reported in table 3 and the feature importance metrics for the random forest and the

Table 1 Clinical–pathological characteristics of the study sample

	BRCA1/2 non-carriers (n=496)	*BRCA1/2* carriers (n=152)	
Features			**P value**
Age, mean (SD)	60.1 (11.6)	57.1 (11.1)	0.003*
Histotype, n (%)			<0.001†
Serous G2/G3	353 (71.2)	144 (94.7)	
Serous G1	16 (3.2)	1 (0.7)	
Endometrioid G2/G3	33 (6.7)	(3.9)	
Endometrioid G1	50 (10.1)	1 (0.7)	
Clear cells	42 (8.5)	0	
Rare histotypes	2 (0.4)	0	
Family history of OC, n (%)	29 (5.8)	53 (34.9)	<0.001†
Family history of BC, n (%)	169 (34.1)	78 (51.3)	<0.001†
Personal history of BC, n (%)	36 (7.3)	29 (19.1)	<0.001†
FIGO stage, n (%)			0.007†
I/II	154 (31.0)	30 (19.7)	
III/IV	342 (69.0)	122 (80.3)	

*t-test.
†χ^2 test.
BC, breast cancer; FIGO, Federation of Gynecology and Obstetrics; OC, ovarian cancer.

support vector machine are shown in table 4. Notably, in each algorithm, the most important input feature was the first-degree familiarity for OC, followed by histotype.

DISCUSSION

As artificial intelligence (AI) technologies continue to develop and improve, ML is finding more and more applications in medicine and oncology[10–12] and recommendations have been developed on how to use ML to develop predictive models.[13] Indeed, it is widely recognised that these methods offer many advantages over traditional approaches, through providing internal validation and testing of the algorithm.

In the context of OC research and care, ML approaches have been employed to facilitate diagnosis and prognosis based on histopathological data, enhance diagnostic accuracy by identifying specific patterns in imaging and biomarkers, enable the early detection of cancer through the integrated analysis of various data types, and predict the malignancy grade of tumours.[14–19] However, to the best of our knowledge, they have never been used to predict the likelihood of identifying germline *BRCA1/2* pathogenic variants based on the clinicopathological features of affected patients.

Using an ML approach, we found that the boosting algorithm proved to have the best performance in the test sample in terms of accuracy, precision and recall. Although accuracy and precision are in the range 80–90% and can be considered good, values>90% would be desirable for routine use in clinical practice; moreover, 30.8% of Recall indicated a still too low sensitivity. Therefore, our model should be validated on a larger and independent cohort and could further be refined with information from additional variables, such as molecular characteristics of tumour tissue.

In both models, the feature that resulted most influential in predicting the presence of a germline *BRCA1/2* pathogenic variant was the family history of OC, followed by the non-low-grade serous histotype, the age at diagnosis <50 years, the family history of BC and the personal history of BC. Regarding the family history of BC/OC, while in this and in the previous study it was necessary

Table 2 Evaluation metrics for the three ML algorithms

	Test (n=129)				Validation (n=104)
Algorithm	Accuracy	Precision (PPV)	Recall (TP)	AUC	Accuracy
Boosting	0.845	0.800	0.308	0.788	0.798
Random forest	0.744	0.556	0.147	0.713	0.808
Support vector machine	0.814	0.727	0.276	0.623	0.817

AUC, area under the curve; ML, machine learning; PPV, positive predictive value; TP, true positive.

Table 3 Relative influence scores of features included in the boosting algorithm

Feature importance metrics		
	Relative influence	**Mean dropout loss**
Family history of OC	52.888	0.283
Histotype	19.462	0.236
Personal history of BC	17.078	0.211
Age	8.394	0.218
Family history of BC	2.178	0.168
FIGO stage	0.000	0.159

Mean dropout loss (defined as 1—area under the curve is based on 50 permutations.
BC, breast cancer; FIGO, Federation of Gynecology and Obstetrics; OC, ovarian cancer.

to consider first and second degree of kinship together to have a sufficient number, in a future possibly a larger cohort, it will be useful to try to stratify based on the different degrees of kinship, for a more accurate prediction of the risk. The FIGO stage, on the other hand, was found to have practically no influence; this confirms the results of our previous logistic regression model, which showed that neither the main effect of FIGO stage nor its interaction with 'family history' was significant, thus justifying the exclusion of this variable from the risk score.[7 20]

Table 4 Relative influence scores of features included in the random forest and support vector machine algorithms

(A) Random forest			
Feature importance metrics			
	Mean decrease in accuracy	**Total increase in node purity**	**Mean dropout loss**
Family history of ovarian cancer	0.042	0.151	0.454
Histotype	0.03	0.046	0.35
Age	6.455×10^{-4}	0.034	0.346
Family history of breast cancer	–0.001	0.023	0.321
Personal history of breast cancer	0.009	0.022	0.337
FIGO stage	0.007	0.001	0.326

(B) Support vector machine	
Feature importance metrics	
	Mean dropout loss
Family history of ovarian cancer	0.491
Histotype	0.344
FIGO stage	0.317
Age	0.317
Family history of breast cancer	0.317
Personal history of breast cancer	0.316

Mean dropout loss is based on 50 permutations.
FIGO, Federation of Gynecology and Obstetrics.

These results confirmed that the variables previously identified using a traditional approach are indeed relevant for predictive purposes, supporting the usefulness of our predictive model in estimating the likelihood of patients with OC being carriers of germline pathogenic variants in the *BRCA1/2* genes. Moreover, they supported the reliability of the previous logistic regression-based model because the ML approaches allow the algorithms to be tested and validated on two different samples. However, although the model is very promising, it requires further validation on a larger independent cohort before it can be definitively transferred to daily clinical practice, and therefore we aim to validate it on other datasets.

Once validated, it could have important implications by helping patients to obtain more accurate and personalised pretest genetic counselling and, from the perspective of health services, by allowing genetic testing to be prioritised based on the likelihood of clinical benefit, where appropriate. It is worth noting that the clinical features included in these models are collected as part of routine clinical practice, thereby facilitating the potential integration of AI into OC screening.

Contributors Conceptualisation: GI, DT and PR. Methodology: GI, LG, DT and PR. Investigation and data curation: all authors. Writing—original draft and preparation: GI. Writing—review and editing: DT and PR. All authors have read and agreed to the published version of the manuscript. DT is the guarantor.

Funding The work reported in this publication was funded by the Italian Ministry of Health, RC-2025-2797528.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants and was approved by the ethics committee of 'Area Vasta Emilia Centro' of Emilia-Romagna region, Italy (490/2022 /Oss/AOUBo). Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request.

ORCID iD
Giovanni Innella https://orcid.org/0000-0002-6909-2412

REFERENCES

1 Sun Y, Wen B. Machine-learning diagnostic models for ovarian tumors. *Heliyon* 2024;10:e36994.
2 Kokori E, Aderinto N, Olatunji G, *et al*. Machine learning use in early ovarian cancer detection. *Discov Med* 2025;2:66.
3 Guo LY, Wu AH, Wang YX, *et al*. Deep learning-based ovarian cancer subtypes identification using multi-omics data. *BioData Min* 2020;13:10.
4 Sorayaie Azar A, Babaei Rikan S, Naemi A, *et al*. Application of machine learning techniques for predicting survival in ovarian cancer. *BMC Med Inform Decis Mak* 2022;22:345.
5 Yang Z, Zhang Y, Zhuo L, *et al*. Prediction of prognosis and treatment response in ovarian cancer patients from histopathology images using graph deep learning: a multicenter retrospective study. *Eur J Cancer* 2024;199:113532.

6 Xiong X, Cai L, Yang Z, *et al*. Multimodal data integration with machine learning for predicting PARP inhibitor efficacy and prognosis in ovarian cancer. *Front Oncol* 2025;15:1571193.
7 Innella G, Erini G, De Leo A, *et al*. Development of a risk score based on clinical-pathological features to predict the presence of germline BRCA1/2 pathogenic variants in ovarian cancer patients. *ESMO Open* 2025;10:105300.
8 Liu S, Wu X. n.d. Letter Re: 'Development of a risk score based on clinicalepathological features to predict the presence of germline BRCA1/2 pathogenic variants in ovarian cancer patients.
9 Innella G, Godino L, Erini G, *et al*. Factors predicting *BRCA1/2* pathogenic variants in patients with ovarian cancer: a systematic review with meta-analysis. *J Clin Pathol* 2023;76:510–7.
10 Coudray N, Ocampo PS, Sakellaropoulos T, *et al*. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018;24:1559–67.
11 Du-Harpur X, Watt FM, Luscombe NM, *et al*. What is AI? Applications of artificial intelligence to dermatology. *Br J Dermatol* 2020;183:423–30.
12 Lundberg SM, Erion G, Chen H, *et al*. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat Mach Intell* 2020;2:56–67.
13 Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385.
14 Breen J, Allen K, Zucker K, *et al*. Artificial intelligence in ovarian cancer histopathology: a systematic review. *NPJ Precis Oncol* 2023;7:83.
15 Ledger A, Ceusters J, Valentin L, *et al*. Multiclass risk models for ovarian malignancy: an illustration of prediction uncertainty due to the choice of algorithm. *BMC Med Res Methodol* 2023;23:276.
16 Mercioni MA, Holban S. Ovarian cancer detection with popular ai algorithms: a brief review. In: Costin HN, Magjarević R, Petroiu GG, eds. Advances in Digital Health and Medical Bioengineering. EHB 2023. IFMBE Proceedings; 2024
17 Hong M-K, Ding D-C. Early Diagnosis of Ovarian Cancer: A Comprehensive Review of the Advances, Challenges, and Future Directions. *Diagnostics (Basel)* 2025;15:406.
18 Naderi Yaghouti AR, Shalbaf A, Alizadehsani R, *et al*. Artificial Intelligence for Ovarian Cancer Detection with Medical Images: A Review of the Last Decade (2013–2023). *Arch Computat Methods Eng* 2025;32:4093–124.
19 Acosta-Jiménez S, Mendoza-Mendoza MM, Galván-Tejada CE, *et al*. Detection of ovarian cancer using a methodology with feature extraction and selection with genetic algorithms and machine learning. *Netw Model Anal Health Inform Bioinforma* 2025;14.
20 Innella G, Erini G, De Leo A, *et al*. Response to: Re: 'Development of a risk score based on clinical–pathological features to predict the presence of germline BRCA1/2 pathogenic variants in ovarian cancer patients'. *ESMO Open* 2025;10:105762.