applied sciences

MDPI

Article

Enhancing Breast Cancer Risk Prediction with Machine Learning: Integrating BMI, Smoking Habits, Hormonal Dynamics, and BRCA Gene Mutations—A Game-Changer Compared to Traditional Statistical Models?

Luana Conte [1,2], Emanuele Rizzo [3], Emanuela Civino [4], Paolo Tarantino [5], Giorgio De Nunzio [1,2,*,†] and Elisabetta De Matteis [4,†]

[1] Laboratory of Biomedical Physics and Environment, Department of Mathematics and Physics "E. De Giorgi", University of Salento, 73100 Lecce, Italy; luana.conte@unisalento.it
[2] Laboratory of Advanced Data Analysis for Medicine (ADAM) at the Laboratory of Interdisciplinary Research Applied to Medicine, University of Salento, Local Health Authority of Lecce, 73100 Lecce, Italy
[3] Department of Biological and Environmental Sciences and Technologies, University of Salento, 73100 Lecce, Italy; emanuele.rizzo1@unisalento.it
[4] Oncological Screenings Unit, Local Health Authority of Lecce, 73100 Lecce, Italy; civino.emanuela84@gmail.com (E.C.); dr.dematteis.elisabetta@gmail.com (E.D.M.)
[5] Medical Genetics Unit, "Vito Fazzi" Hospital, 73100 Lecce, Italy
* Correspondence: giorgio.denunzio@unisalento.it
† Co-Last Authors: Giorgio De Nunzio and Elisabetta De Matteis.

Abstract: The association between genetics and lifestyle factors is crucial when determining breast cancer susceptibility, a leading cause of deaths globally. This research aimed to compare the body mass index, smoking behavior, hormonal influences, and BRCA gene mutations between affected patients and healthy individuals, all with a family history of cancer. All these factors were then utilized as features to train a machine learning (ML) model to predict the risk of breast cancer development. Between 2020 and 2023, a total of 1389 women provided detailed lifestyle and risk factor data during visits to a familial cancer center in Italy. Descriptive and inferential statistics were assessed to explore the differences between the groups. Among the various classifiers used, the ensemble of decision trees was the best performer, with a 10-fold cross-validation scheme for training after normalizing the features. The performance of the model was evaluated using the receiver operating characteristic (ROC) curve and its area under the curve (AUC), alongside the accuracy, sensitivity, specificity, precision, and F1 score. Analysis revealed that individuals in the tumor group exhibited a higher risk profile when compared to their healthy counterparts, particularly in terms of the lifestyle and genetic markers. The ML model demonstrated predictive power, with an AUC of 81%, 88% sensitivity, 57% specificity, 78% accuracy, 80% precision, and an F1 score of 0.84. These metrics significantly outperformed traditional statistical prediction models, including the BOADICEA and BCRAT, which showed an AUC below 0.65. This study demonstrated the efficacy of an ML approach in identifying women at higher risk of breast cancer, leveraging lifestyle and genetic factors, with an improved predictive performance over traditional methods.

Keywords: breast cancer risk; risk factors; machine learning; statistical models

Citation: Conte, L.; Rizzo, E.; Civino, E.; Tarantino, P.; De Nunzio, G.; De Matteis, E. Enhancing Breast Cancer Risk Prediction with Machine Learning: Integrating BMI, Smoking Habits, Hormonal Dynamics, and BRCA Gene Mutations—A Game-Changer Compared to Traditional Statistical Models? *Appl. Sci.* **2024**, *14*, 8474. https://doi.org/10.3390/app14188474

Academic Editors: Asiya Khan and Gloria Iyawa

Received: 6 August 2024
Revised: 11 September 2024
Accepted: 15 September 2024
Published: 20 September 2024

1. Introduction

Breast cancer is the prevalent form of tumor affecting women globally [1,2]. In Italy alone, in 2020, there were 55,700 new cases, representing 30% of all cancer diagnoses and the leading cause of cancer death in the country [3]. Despite the increased awareness of breast cancer [4–6], significant public focus, and improvements in breast imaging techniques for early detection [7–9], death rates due to breast cancer continue to be worryingly high worldwide.

While screenings are a fundamental pillar of the early diagnosis of tumors in the general population, it is now possible to investigate subjects with an elevated risk of tumor development even before the cancer has developed. In a portion of the population, certain genetic mutations can predispose individuals to the development of tumors. While cancer typically arises from the gradual and random accumulation of multiple genetic mutations, there are instances where it is linked to specific driver mutations or monogenic changes. These genetic factors can substantially increase the likelihood of developing certain types of cancer in comparison to the general population. These cancers, commonly known as hereditary–familial tumors, represent roughly 15% of all cancer cases [10] and occur in individuals with an inherited genetic variant that directly contributes to the onset of the disease.

Among these identified genes, the mutation of the BRCA1 and 2 genes significantly elevates the risk of developing breast and ovarian cancers [11]. High-risk individuals can be identified through current breast cancer screening risk stratification models, based on statistics, which pinpoint women eligible for additional screening or preventive interventions [12–14]. The risk models widely used in clinical settings are typically based on statistics and primarily rely on factors such as genetics, family history of breast cancer, prior benign breast diseases, and reproductive history. However, these variables may not always be readily available during routine screening procedures. Additionally, some risk models may exhibit limited discriminatory capability, with area under the curve (AUC) values below 0.7 [15–20]. Among them is the BCRAT, also known as the Gail model, which is designed as a breast cancer risk assessment tool to predict the probability of a woman developing invasive breast cancer within the next 5 years and up to the age of 90. This tool considers various risk factors, including age at menarche, age at first childbirth, ethnicity, race, history of breast cancer and the number of breast biopsy examinations. Another model, the BOADICEA, predicts individual risks for breast and ovarian cancer based on both family history and BRCA1 and BRCA2 mutations. Unlike the Gail model, the BOADICEA takes into account a detailed family history, encompassing first-, second-, and third-degree relatives, and can also integrate genetic test results considering other rare yet highly penetrant genes, such as PALB2, CHEK2, and ATM, employing a polygenic model [21]. Consequently, the BOADICEA indicates a more substantial decrease in the relative risk as the age of the affected relative increases, aligning more closely with the existing literature [22].

Other commonly used models include the BRCAPRO, which assesses the risk of having BRCA1/2 mutations (https://projects.iq.harvard.edu/bayesmendel/brcapro, accessed on 15 April 2024), and the IBIS, which calculates the breast cancer risk using the Tyrer–Cuzick index in percentage form (https://ibis-risk-calculator.magview.com/, accessed on 15 April 2024). The BRCAPRO calculates the breast cancer risk primarily by assessing the probability of carrying major genes based on personal and family histories of breast, ovarian, or related cancers associated with these genes [23]. It incorporates pedigree information, including age at onset, and the outcomes of genetic testing for BRCA1 and BRCA2 in both women and their relatives. However, these models are less suitable for assessing the breast cancer risk in the general population, as they do not account for other unknown genetic factors.

The Tyrer–Cuzick model [24], instead, combines two widely used sub-models for assessing the breast cancer risk. It integrates a genetic segregation model for familial risk with a proportional hazards regression model for other risk factors. The risk factors in the model can be broadly categorized into five main groups: i. family history and highly penetrant dominant genetic mutations; ii. factors related to estrogen exposure, such as age at first childbirth, age at menopause, menarche (onset of periods), use of hormone replacement therapy, height, and weight; iii. specific types of prior benign breast disease; iv. breast imaging features observed on mammograms, notably the amount of dense tissue; and v. common but individually less impactful genetic differences, specifically single nucleotide polymorphisms (SNPs). These SNPs, comprising several hundred relatively common genetic variants, each with a modest impact on disease, collectively contribute significantly

to the overall risk assessment through a "polygenic risk score" [25]. Additionally, there are other potential risk factors that are challenging to quantify but may enhance the model performance.

Other models, including the Breast Cancer Surveillance Consortium (BCSC) [20,26] and Rosner–Colditz models [27,28], rely solely on a regression function. The Rosner–Colditz model includes terms for the same broad risk factors as the Tyrer–Cuzick model, along with additional factors such as alcohol consumption, adolescent body somatotype, and hormone levels. However, these models differ in the assumptions regarding the risks and prevalence of risk factors, as well as the utilization of interaction terms [22].

Despite the widespread use of these risk models, each of them presents significant limitations that can affect their predictive accuracy and generalizability. The Gail model, for example, exhibits low AUC values in non-white populations, particularly among African American and Asian women, as it was initially developed using data from predominantly white women [12,29]. Furthermore, it fails to incorporate crucial high-risk factors such as genetic mutations, including BRCA1/2, or a more extensive family history, and it overlooks important risk modifiers like breast density and lifestyle factors. Similarly, the BOADICEA, while more inclusive of a detailed family history and genetic factors, suffers from its complexity, making it less accessible for widespread clinical use. Moreover, its reliance on data from European populations limits its applicability to other ethnic groups [30,31]. The BRCAPRO model, while focusing on BRCA1/2 mutations, does not account for other significant risk factors and heavily depends on a detailed and accurate family history, which may not always be available or reliable [32,33]. The Tyrer–Cuzick model (IBIS), on the other hand, is known to overestimate the risk in certain populations, such as women with benign breast disease or those using hormone replacement therapy, and requires complex, detailed data inputs [22,24]. Additionally, the Breast Cancer Surveillance Consortium (BCSC) model is limited by its reliance on U.S. mammography screening data, making it less relevant for international populations, while its focus on breast density often overlooks other genetic or non-genetic risk factors [34,35]. Lastly, the Rosner–Colditz model, though comprehensive, is data-intensive and challenging for routine clinical use, while it underestimates the genetic risk by prioritizing modifiable risk factors [36,37]. These limitations highlight the need for more adaptable and inclusive models that integrate a wider array of genetic, environmental, and lifestyle factors to enhance the breast cancer risk prediction accuracy.

In conclusion, although traditional statistical tools are widely used for identifying target populations for screening, they often show weak correlations with actual screening outcomes. More advanced methods, such as machine learning (ML) algorithms trained on the same data, may offer superior predictive capacity [38], as suggested by successful ML applications in various domains, underscoring the potential of these methods to enhance the predictive accuracy across different areas, including healthcare [39]. Several factors explain why ML models may outperform traditional statistical models. Firstly, ML handles complex, non-linear relationships by incorporating a vast range of risk factors (genetic, lifestyle, and environmental), without oversimplifying their relationships. Algorithms like random forests or neural networks can model complex non-linear interactions between variables, which traditional models often miss. For example, [40] demonstrated that ML models, due to their inherent non-linearity, achieved equivalent (and superior) performance across different ethnicities compared with the Tyrer–Cuzick model, which showed significant differences in the AUC values for white and African American women. The same paper also showed another strong point of ML models, i.e., the capability of incorporating high-dimensional data, such as extensive genomic information, providing a more comprehensive risk prediction. Another advantage of ML is the ability to address data completeness issues [41]: ML models are better equipped to handle missing or incomplete data through advanced imputation techniques, by the definition of dummy variables, and by architectures natively capable of facing missing data issues, such as decision-tree-based approaches. This makes ML models more robust in real-world settings where perfect data are rarely available. ML can also be continuously updated as new data are collected,

ensuring they remain accurate or improve over time and adapt to changes in population health dynamics [42]. Additionally, ML can work with unstructured data, such as images, text, or audio, and deep learning (DL) methods can automatically identify and construct features from raw data, reducing the need for manual feature extraction and selection. Overall, ML models and other artificial intelligence (AI) approaches may indeed show significant potential for improving the accuracy of classifying women with and without breast cancer.

In this scenario, the aim of our study was firstly to thoroughly examine the differences between healthy subjects and those with diagnosed breast cancer, all with a family history for cancer. Similar to traditional statistical models, we focused on lifestyle factors such as the BMI, voluntary habits like smoking, hormonal influences and the mutational status of the BRCA genes. This approach allowed for the collection and utilization of the majority of predictive variables known to increase the risk of developing breast cancer. Secondly, utilizing the capabilities of AI, risk factors served [42] as features to predict subjects at high risk of breast cancer. Unlike traditional statistical models, the proposed model harnesses more advanced predictive models in the AI domain, enabling increasingly personalized preventive interventions and treatments.

2. Materials and Methods

2.1. Design

This retrospective study spanned from January 2020 to December 2023, focusing on individuals with cancer familiarity who attended the familial cancer center of the local health authority of Lecce, Italy. In this study, only women were included as subjects, given that the investigation also focused on hormonal and reproductive factors, which are crucial for a comprehensive understanding of the breast cancer risk and prevention. Participants were referred to the clinic by specialists or general practitioners, or attended spontaneously. They were asked to complete the questionnaire independently before the oncogenetic counseling session and submit it at the time of the consultation. The questionnaire was developed to include several sections, each targeting different aspects of lifestyle and medical history that could influence the breast cancer risk. These sections encompassed socio-demographic information, reproductive history, and family history of cancer, among others. Details of the survey are provided in the next section.

In adherence with the criteria for conducting genetic testing, some women underwent blood sampling to determine the presence of mutations in the BRCA1 and BRCA2 genes, as well as variants of uncertain significance (VUSs).

2.2. Inclusion and Exclusion Criteria

The inclusion criteria were as follows: women of reproductive age as well as those who were postmenopausal were selected, with an age cutoff of 45 years to distinguish between women of childbearing age and those beyond it. Only women with a family history of cancer were included in this study, as the objective was to analyze the risk in genetically predisposed individuals. Women with specific medical conditions that could confound the study outcomes were excluded. Additionally, patients with other concomitant tumors or metastases were excluded, as only primary breast cancer was considered for this analysis.

2.3. Survey Instrument

Our survey was organized into two main sections to gather comprehensive data. The first section focused on demographics and exposure to risk factors. This section covered topics such as age, body weight, height, body mass index (BMI), and smoking habits. The second section aimed to collect information on specific female characteristics. This included hormonal dynamics details, including menarche, pregnancies, abortions, breastfeeding practices, and information related to menopause.

2.4. Ethical Considerations

Ethical clearance for this retrospective study was secured from the Bioethical Committee of the IRCCS Tumor Institute "Giovanni Paolo II", Bari, Italy, under the protocol number 1695/CEL, dated 10 June 2024. The ethical framework governing this study was thoroughly outlined at the outset of the questionnaire, ensuring alignment with the guidelines set forth by the Italian Data Protection Authority. We underscored to potential participants that their involvement in the study was entirely voluntary, allowing the freedom to withdraw at any stage without consequence. An informed consent document was provided to all interested individuals, reiterating the voluntary basis of their participation, alongside assurances regarding the confidentiality and anonymity of the data collected. To further safeguard participant anonymity, all the collected responses were subjected to a de-identification process.

2.5. Statistical Analysis

The dataset encompassing the responses from a total of 1389 participants was evaluated by employing a blend of descriptive and inferential statistical techniques. The cohort was divided into two distinct groups for the analysis: Group A, comprising 473 healthy subjects with a family history of cancer, and Group B, consisting of 916 participants diagnosed with breast cancer. Continuous variables were summarized using the mean values and standard deviations, while categorical variables were quantified through the frequencies and percentage distributions. The Mann–Whitney test was used to assess the differences in responses between the two groups. Additionally, for hormonal dynamics variables, statistical differences according to the Mann–Whitney test were also calculated by dividing the women from both groups into those younger or older than 45 years of age. The age of 45 was chosen as a cutoff to distinguish between women of childbearing age and those who have presumably surpassed the age of having children, aiming to further refine the analysis based on potential differences in the risk factors or genetic predisposition associated with age. The Mann–Whitney non-parametric test was used instead of a parametric one because both the Shapiro–Wilk and Kolmogorov–Smirnov tests showed significant p-values, indicating that the data were not normally distributed.

The statistical analyses were executed using the MATLAB software, 2023b version, adhering to a significance threshold of $p < 0.05$ to validate the reliability and significance of the findings.

2.6. Machine Learning

A supervised ML approach was employed to predict the likelihood of developing breast cancer based on the risk factors collected. These risk factors, serving as features, were used to train several ML classifiers, including decision tree, support vector machine (SVM), naive Bayes, ECOC, discriminant analysis, linear models, ensemble of trees, artificial neural networks (ANNs), and k-nearest neighbors (KNNs). Among these, the ensemble of decision trees emerged as the best performer. This model was chosen not only for its superior performance but also for its ability to handle missing values effectively and to combine multiple weak learners into a stronger predictive model, which is particularly advantageous given the nature of the dataset. The training process used a 10-fold cross-validation scheme to partition the data into training and validation sets. To optimize the performance of the best model, we employed the MATLAB automatic hyperparameter optimization feature. This process automatically tuned the parameters of the ensemble of decision tree models to find the optimal configuration for our dataset. The optimization procedure selected the AdaBoostM1 method, configured with 11 as the number of trees and a learning rate of 0.42. The number of trees and the learning rate were determined as the most effective parameters to balance the trade-off between model complexity and performance. The ensemble method, AdaBoost, enhanced the performance of these decision trees by iteratively focusing on difficult-to-classify instances, thus improving the overall accuracy of the model. AdaBoost's capability to combine weak learners into a strong

classifier and its inherent robustness against overfitting made it the preferred choice for this study. For completeness, it is important to mention that other classifiers, including artificial neural networks with missing data imputation, were also used, but their results were inferior. To evaluate the classifier performance and to determine the optimal threshold for predictions, the receiver operating characteristic (ROC) curve was employed. It plots the true positive rate (sensitivity) on the Y-axis against the false positive rate (1—specificity) on the X-axis across different threshold values. The curve provides a comprehensive overview of the trade-off between sensitivity and specificity at various thresholds. Based on the optimal working point of the ROC curve, a binary classifier was derived. The ROC curve was utilized to assess the classifier performance with the area under the curve (AUC) figure of merit and to determine an "optimal" prediction threshold that maximizes the accuracy. At this optimal threshold, the sensitivity, specificity, and accuracy of the classifier were calculated. The sensitivity, representing the true positive rate, and specificity, representing the true negative rate, were derived directly from the confusion matrix at this threshold. The accuracy was computed as the proportion of true results (both true positives and true negatives) relative to the total number of cases examined. To further assess the performance of the model, particularly in the context of imbalanced classes, the precision–recall (PR) curve was employed. The PR curve plots the precision (the proportion of true positive results relative to all positive predictions) against the recall (the proportion of true positive results relative to all actual positives) at various threshold levels.

The F1 score, which is the harmonic mean of the precision and recall, was calculated at specific thresholds to provide a balanced measure of the model's accuracy. The F1 score offers insight into how well the model balances the trade-off between precision and recall, especially when the goal is to minimize both false positives and false negatives.

To ensure a robust and unbiased classification, the features were normalized to a range of 0–1 using min–max normalization on the training dataset [43–45]. The same normalization parameters were then applied to the validation set samples. This procedure ensured that each feature contributed equally to the model training process. This scaling method is crucial for preventing features with larger ranges from dominating the decision-making process of the ensemble decision trees used in the AdaBoost model. For instance, without normalization, features such as age, which can vary significantly, might have disproportionately influenced the model, overshadowing other important features like the BMI or number of pregnancies. Similar considerations hold for the z-score standardization, which scales features to have a mean of 0 and a standard deviation of 1. The z-score normalization is typically more appropriate for datasets where the distribution of features is approximately normal, and in the presence of outliers because extreme values do not disproportionately affect the range of transformed data. Given the non-Gaussian distributions of our dataset's features, min–max normalization was more appropriate to maintain consistency and improve model performance.

The computation was performed with MATLAB software, 2023b license.

3. Results

3.1. Baseline Characteristics and Exposure to Risk Factors

A total of 1389 women chose to participate. The responses from the group with a family history of cancer were separately analyzed for two groups: Group A, which included women who had not been diagnosed with cancer, and Group B, consisting of those who had received a cancer diagnosis. Table 1 presents the demographics and risk factor exposure for the participants. The median age in Group A was 46 years old compared to 54 in the tumor group ($p < 0.001$). The BMI further differentiated the groups ($p = 0.01$).

Table 1. Socio-demographic characteristics and risk factor exposure. *$p < 0.05$; ***$p < 0.001$.

	Group A Healthy Subjects with a Family History of Breast Cancer (n = 473) N (%)	**Group B Subjects with Cancer (n = 916) N (%)**	***p*-Value**
Age			
Range	18–93	26–95	<0.001
Median	47.00	55.53	***
STD	16.47	12.44	
Indicate body weight (kg)			
Range	40–130	42–154	0.26
Median	64.00	67.00	
STD	15.44	14.22	
Indicates height (cm)			
Range	146–183	144–188	0.02 *
Median	165.00	162.00	
STD	6.20	6.42	
Body mass index (BMI)			
Range	16–48	17–69	0.01 **
Median	23.45	25.87	
STD	6.08	5.34	
Being a smoker			
No	130 (27.5)	455 (49.7)	
Yes	36 (7.6)	100 (10.9)	0.28
Ex-smoker	24 (5.1)	106 (11.6)	
Missing	283 (59.8)	255 (27.8)	
Smoking duration (years)			
Range	1–46	2–61	0.009 **
Median	15.00	20.00	
STD	11.99	11.90	
Number of cigarettes smoked			
Range	1–30	1–40	0.91
Median	10.00	10.00	
STD	6.26	6.56	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Similar results were found in both groups regarding smoking habits, with more smokers and ex-smokers in the tumor group. Group A also had a lower median of long-term smokers (17.02) compared to Group B (21.52), ($p = 0.009$). No difference was found in the number of cigarettes smoked. As shown in Table 1, the demographic characteristics of the study participants revealed key differences between the healthy group and those diagnosed with breast cancer. The mean age of participants in Group B was significantly higher than that in Group A, which aligns with the understanding that the breast cancer risk increases with age. Additionally, the BMI distribution suggests a potential link between a higher body mass and the breast cancer incidence ($p = 0.01$). Figure 1 shows that the BMI was slightly higher in Group B, supporting this association. Smoking habits, while not significantly different in terms of the number of cigarettes smoked, showed a trend toward a longer smoking duration in Group B, which could indicate a cumulative risk factor.

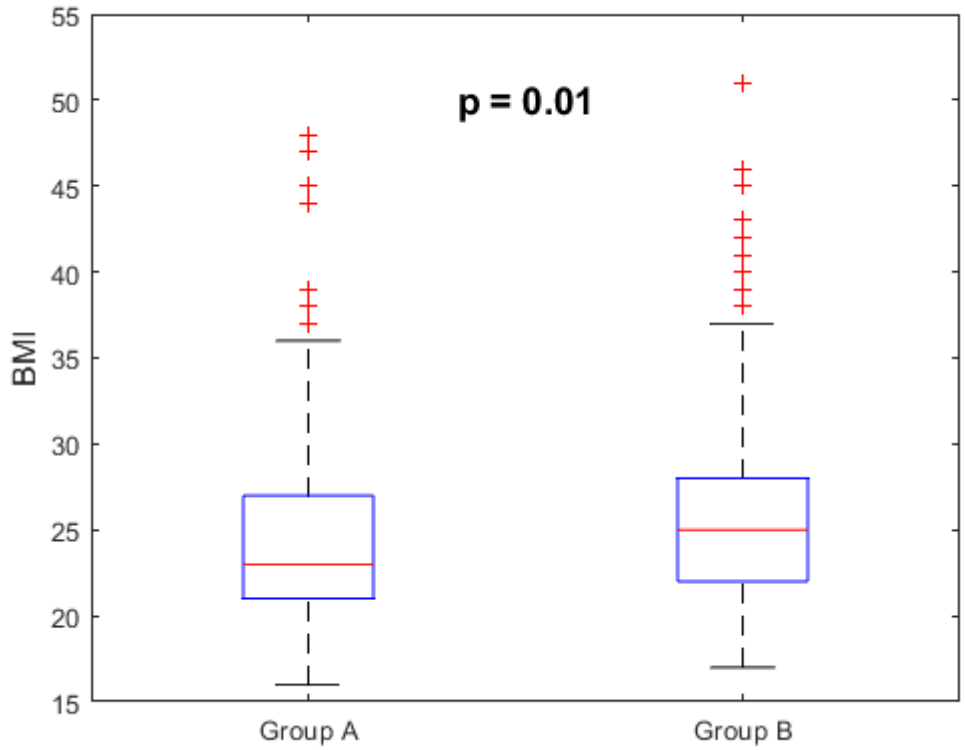

Figure 1. Boxplot comparing the body mass index (BMI) between Group A (healthy) and Group B (cancer-affected). The red horizontal line inside each box represents the median BMI for each group, while the edges of the box indicate the first and third quartiles. The whiskers extend to the most extreme values that are not considered outliers (with respect to a 1.5 × IQR threshold, where IQR is the interquartile range). Outliers are represented by the + symbols. The reported *p*-value ($p = 0.01$) indicates a statistically significant difference between the two groups.

Table 2 provides an in-depth look at the hormonal dynamics and reproductive histories of the participants.

Table 2. Hormonal dynamics among respondents. A cutoff of 45 years old was utilized to distinguish between women of childbearing age and those potentially beyond childbearing age, nearing menopause.

	Group A Healthy Subjects with a Family History of Breast Cancer (n = 473) N (%)	**Group B Subjects with Cancer (n = 916) N (%)**	***p*-Value**
Age at menarche			
<45			
Range	9–16	9–16	
Median	12.00	12.00	0.40
STD	1.43	1.53	
≥45			
Range	9–17	9–18	
Median	12.41	12.44	0.79
STD	12.00	12.00	

Table 2. *Cont.*

	Group A Healthy Subjects with a Family History of Breast Cancer (n = 473) N (%)	Group B Subjects with Cancer (n = 916) N (%)	*p*-Value
Number of pregnancies			
<45			
Range	0–3	0–4	
Median	1.00	1.00	0.65
STD	0.96	0.96	
≥45			
Range	0–11	0–9	
Median	2.00	2.00	0.28
STD	1.48	1.31	
Age at first pregnancy			
<45			
Range	18–38	13–42	
Median	29.00	29.00	0.55
STD	4.78	5.80	
≥45			
Range	16–44	13–55	
Median	25.00	26.00	0.20
STD	5.93	6.01	
Number of abortions			
<45			
Range	0–3	0–4	
Median	0.29	0.39	0.62
STD	0.56	0.76	
≥45			
Range	0–6	0–10	
Median	0.71	0.72	0.74
STD	1.07	1.20	
Did you breastfeed your children?			
<45			
No	182 (85)	108 (57.8)	
Yes	32 (15)	79 (42.2)	<0.001 ***
≥45			
No	180 (74.1)	342 (47.9)	
Yes	63 (25.9)	372 (52.1)	<0.001 ***
Missing	16 (3.4)	15 (1.6)	
If you answered yes to the previous question, please indicate the duration in months			
<45			
Range	1–44	1–66	
Median	11.00	9.00	0.74
STD	13.75	13.87	
≥45			
Range	1–50	1–60	
Median	6.00	8.00	0.49
STD	11.42	9.31	

Table 2. *Cont.*

	Group A Healthy Subjects with a Family History of Breast Cancer (n = 473) N (%)	Group B Subjects with Cancer (n = 916) N (%)	*p*-Value
Are you in the age of menopause?			
<45			
No	212 (99.1)	148 (79.1)	<0.001 ***
Yes	2 (0.9)	39 (20.9)	
≥45			
No	183 (75.3)	323 (45.2)	<0.001 ***
Yes	60 (24.7)	391 (54.8)	
Missing	16 (3.4)	15 (1.6)	
Indicate the age at menopause			
<45			
Range	39–41	33–44	
Median	40.00	39.00	0.71
STD	1.41	2.93	
≥45			
Range	30–59	33–66	
Median	50.00	50.00	0.12
STD	5.10	4.47	
Contraceptives assumption			
<45			
No	189 (88.3)	135 (72.2)	<0.001 ***
Yes	25 (11.7)	52 (27.8)	
≥45			
No	224 (92.2)	592 (82.9)	<0.001 ***
Yes	19 (7.8)	122 (17.1)	
Missing	16 (3.4)	15 (1.6)	
Hormonal stimulation for assisted reproduction (PMA)			
<45			
No	211 (98.3)	178 (95.2)	0.05 *
Yes	3 (1.4)	9 (4.8)	
≥45	242 (99.6)	700 (98)	0.98
No	1 (0.4)	114 (2)	
Yes			
Missing	16 (3.4)	15 (1.6)	
Hormonal replacement therapy			
<45			
No	212 (99.1)	172 (92)	<0.001 ***
Yes	2 (0.9)	15 (8)	
≥45			
No	241 (99.2)	668 (93.6)	<0.001 ***
Yes	2 (0.8)	46 (6.4)	
Missing	16 (3.4)	15 (1.6)	

* $p < 0.05$; *** $p < 0.001$.

The average age at which menstruation began and the age of menopause showed no difference between the groups. However, Group B reported more women in menopause compared to Group A ($p < 0.001$).

There was no great difference in the number of pregnancies between the groups. Group B more frequently reported breastfeeding their children (42.2% vs. 15% under 45 years old and 52.1% vs. 25.9 over 45 years old, $p < 0.001$). A similar duration of breastfeeding was found between the two groups.

Regarding contraceptive usage, a higher percentage of the healthy subjects reported using contraceptives compared to the affected group (11.7% vs. 27.8% for under 45 years old and 7.8% vs. 17.1% for over 45 years old, $p < 0.001$).

Group B also reported a slightly higher hormonal stimulation for assisted reproduction (PMA) compared to the healthy group ($p = 0.05$). A significant difference was noted in the hormonal replacement therapy usage, with Group A being less likely to have used this therapy compared to the affected group ($p < 0.001$).

The genetic status of the participants, focusing on the outcomes of the BRCA1 and BRCA2 mutations, was assessed for subjects who underwent blood sampling in adherence with the criteria for conducting genetic testing (Table 3). Positive mutation statuses were reported in 41% (n = 194) of t Group A and in a lower percentage of 12.6% (n = 115) within the affected group. Also, VUSs were observed in 1.7% (n = 8) of the healthy group, contrasting with 5.5% (n = 50) in Group B ($p < 0.001$). The specific mutation analysis revealed that 58.4% of the healthy subjects and 45.4% of the affected group had BRCA1 mutations, while BRCA2 mutations were identified in 41% of Group A and 52.7% of the affected group; a small fraction (0.5% vs. 1.8%) had mutations in both BRCA1 and BRCA2 ($p < 0.001$).

Table 3. Mutational status of the respondents *** $p < 0.001$.

	Group A Healthy Subjects with a Family History of Breast Cancer (n = 473) N (%)	**Group B Subjects with Cancer (n = 916) N (%)**	***p*-Value**
Mutation outcomes			
Negative	162 (34.2)	640 (69.9)	
Positive	194 (41.0)	115 (12.6)	<0.001 ***
VUSs	8 (1.7)	50 (5.5)	
Not screened	109 (23)	111 (12.1)	
Specific found mutations			
BRCA1	118 (58.4)	75 (45.4)	<0.001 ***
BRCA2	83 (41.0)	87 (52.7)	
BRCA1/2	1 (0.5)	3 (1.8)	

3.2. Machine Learning Predictive Model

The risk factors collected were employed to train several classifiers with the aim of predicting the risk of developing cancer. Figure 2 presents the ROC curves of all the classifiers, along with their corresponding AUC values. The ensemble of decision trees was the best performer. Key metrics such as the AUC, accuracy, specificity, and sensitivity provide a comprehensive overview of the model effectiveness. An average AUC of 81% was found in distinguishing between those at risk of developing cancer and those not at risk, with 88% sensitivity, 57% specificity, 78% accuracy, 80% precision, and an F1 score of 0.84, as computed at the optimal cutoff point, which maximizes the classifier's accuracy. To assess the stability and reliability of our machine learning model, we conducted a 10-fold cross-validation, calculating the AUC for each fold. As depicted in Figure 3, the AUC values demonstrated consistent performance across all the folds, with only minor fluctuations

observed. The AUC values ranged from approximately 0.78 to 0.88, indicating a robust and reliable model performance.

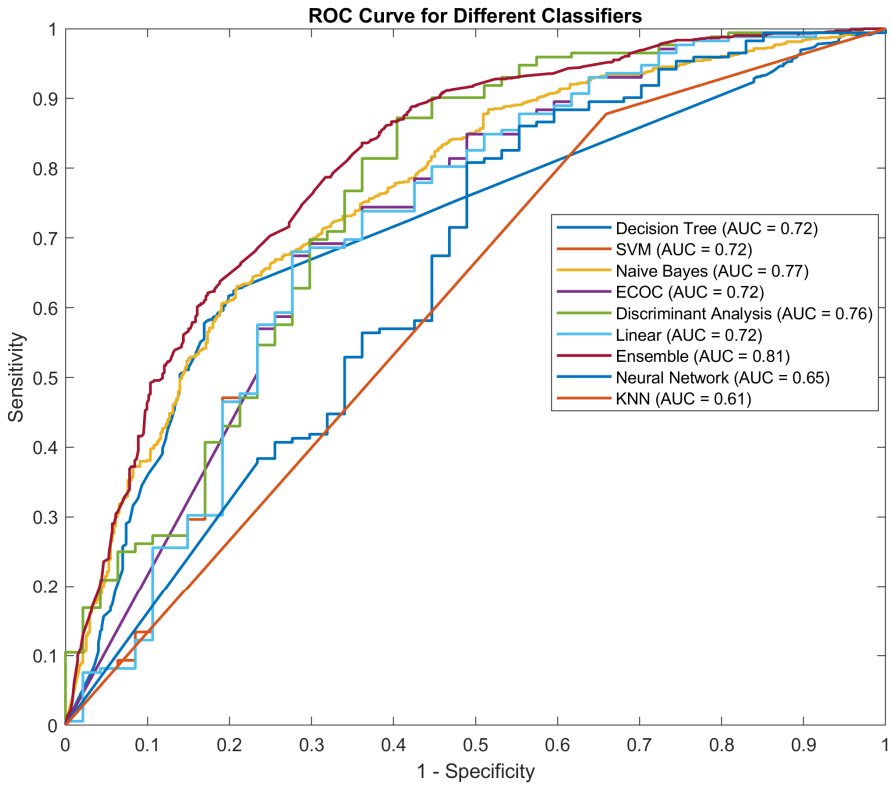

Figure 2. ROC curves for the different classifiers used to predict the risk of developing cancer. The classifiers tested include decision tree, support vector machine (SVM), naive Bayes, ECOC, discriminant analysis, linear, ensemble of trees, neural network, and k-nearest neighbor (KNN). The AUC values for each classifier are displayed in the legend, with the ensemble of trees achieving the highest AUC of 0.81, indicating the best predictive performance among the models tested.

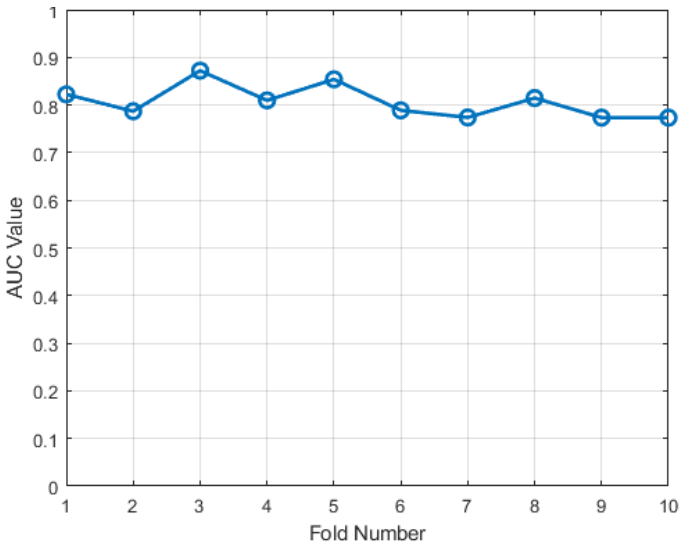

Figure 3. Distribution of the AUC values across the 10-fold cross-validation.

The ROC and the PR curves, as depicted in Figure 4, are a graphical representation of the best classifier performance.

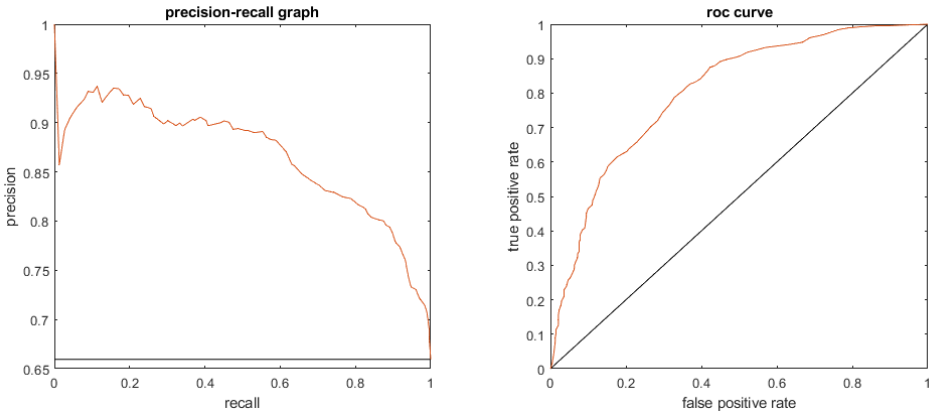

Figure 4. Precision–recall (PR) curve on the left and receiver operating characteristic (ROC) curve on the right, representing the performance of the ensemble of decision trees classifier in predicting the breast tumor risk.

Using the predictor importance function from the decision tree ensemble, we identified the most significant predictors of the breast cancer risk. The analysis highlighted that the genetic mutation status (e.g., BRCA), menopausal status, BMI, age, years of smoking, and breastfeeding history were the top factors influencing the model's predictions.

4. Discussion

This study provides significant insights into the risk factors associated with breast cancer, particularly within a population of women with a family history of the disease. By focusing on women of reproductive age and postmenopausal women, this study specifically targeted a group where hormonal dynamics play a critical role in breast cancer development. Participants were categorized into two distinct groups: Group A, made up of individuals who had not been diagnosed with cancer, and Group B, comprising those who had a prior history of cancer diagnosis. The purpose of this study was to compare these two groups: we examined various aspects of lifestyle, such as the BMI and smoking, hormonal influences and the mutational status of the BRCA genes. These risk factors were used as features to develop an ML model capable of predicting individuals at an increased risk of developing breast cancer. The most influential predictors included the genetic mutation status (e.g., BRCA), menopausal status, BMI, age, years of smoking, and breastfeeding history. The identification of these factors as the top predictors aligns with existing research, highlighting the significant impact of both genetic predispositions and lifestyle factors on the breast cancer risk. The Group B participants were generally older than those in the general population group, highlighting a significant age-related susceptibility to breast cancer, which is consistent with existing research [46]. In the context of the investigated lifestyle factors, Group A and Group B showed a difference in the BMI. The relationship between the BMI and the breast cancer risk has been extensively studied [47–51]. Notably, obesity has been linked to an increased risk of breast cancer in postmenopausal women, a relationship that underscores the significance of body weight management in cancer prevention strategies. This association is thought to be mediated through various mechanisms, including hormonal changes, inflammation, and insulin resistance, which are known to influence cancer pathogenesis. Moreover, the adipose tissue in obese individuals is not merely a passive storage of fat but an active endocrine organ that secretes estrogen, adipokines, and inflammatory markers, all of which have been implicated in cancer development and progression [47]. The stark contrast in the BMI between the two groups highlights the importance of incorporating regular physical activity and a correct diet into daily routines as a potentially effective strategy for lowering the breast cancer risk. Furthermore, it underscores the need for targeted public health strategies aimed at increasing well-being levels, especially among populations at a higher risk of breast cancer.

The observed differences in smoking habits between the two groups highlight intriguing behavioral patterns that merit a deeper analysis within the context of breast cancer risk and survivorship. Specifically, the group with a history of breast cancer showed a greater incidence of long-term smoking compared to the healthy group. Smoking has been identified as a significant risk factor for breast cancer development, particularly for women who begin smoking in adolescence, around the time of menopause, or before having their first child, or those with a familial predisposition to the disease [52]. Smoking has also been linked to an increased likelihood of breast cancer recurrence [53]. It can negatively affect the outcomes of surgical procedures and heighten the chances of complications during breast reconstruction surgeries [54]. Offering smoking cessation support to breast cancer patients is crucial. This intervention can lower the risks tied to radiotherapy and potentially decrease mortality rates, emphasizing the importance of targeted support programs for these individuals [55].

Significant differences were also observed in relation to hormonal factors, including the age at menopause, PMA and use of hormonal replacement therapy. These results are in line with the literature, including the prolonged exposure to endogenous hormones in women with a later onset of menopause [56].

Surprisingly, the affected group reported a higher likelihood of breastfeeding compared to the healthy subjects. This result does not align with research suggesting the protective effects of full-term pregnancies and breastfeeding against breast cancer [57–61]. It is possible that this is a consequence of the higher number of children among the women in Group B compared to Group A. A history of medical abortions, along with the use of multiple contraceptive methods, has been associated with an increased risk of postmenopausal breast cancer [59]. Interestingly, women who have used intrauterine devices for contraception for over twenty years tend to have a lower likelihood of developing breast cancer compared to others in their age group [59]. However, research consistently shows that oral contraceptives can significantly elevate the breast tumor risk. Despite this increased risk, it is essential to recognize the benefits that oral contraceptives provide, which must be weighed against their potential risks [60].

The significant disparity in hormonal replacement therapy (HRT) usage between the two groups highlights the ongoing debate around HRT and the breast cancer risk. With the affected group more likely to have used HRT, this finding aligns with the literature suggesting a potential association between HRT, especially combined estrogen–progestogen therapies, and an increased breast cancer risk [62,63].

Lastly, the observed differences in the BRCA mutation statuses between Group A and Group B underline the critical role of genetic factors in the breast cancer risk. It is normal to observe a higher number of mutations among healthy patients, since for each affected patient, one or more healthy blood relatives are sent for genetic testing to determine their risk. The presence of VUSs in a larger portion of the affected group also highlights the challenges in genetic testing and interpretation, emphasizing the complexity of genetic contributions to breast cancer and the need for further research in this area.

By using factors that significantly differed between Group A and Group B as features, we trained different ML models that were able to predict subjects at an increased risk of developing a breast tumor. The best performer was the ensemble of decision trees. To assess the stability and reliability of the model's performance, we conducted a 10-fold cross-validation procedure and analyzed the distribution of the AUC values across each fold. As shown in Figure 2, the AUC values consistently ranged between approximately 0.78 and 0.88, with minor fluctuations. This consistency across the different folds indicates that the model's ability to discriminate between those at risk and those not at risk of developing breast cancer is robust and not overly dependent on any particular subset of the data. Building on these cross-validation results, the model achieved an overall AUC of 81%. An AUC of 0.5 indicates no discrimination ability, equivalent to random guessing, while an AUC of 1.0 represents a perfect classifier. In our model, the AUC of 81% suggests

a strong ability to distinguish between individuals with and without the risk of developing breast cancer.

To identify the optimal cutoff point, we analyzed the ROC curve to find the threshold that maximizes the classifier's accuracy. This optimal point balances the trade-off between sensitivity and specificity, aiming to maximize the model's effectiveness in correctly identifying true positives while minimizing false positives. At this threshold, the model achieved a sensitivity of 88%, meaning it correctly identified 88% of those at risk, and a specificity of 57%, indicating its moderate ability to correctly identify those not at risk. The overall accuracy at this point was 78%, reflecting the proportion of correct classifications (both true positives and true negatives) across all the cases examined.

In addition to these metrics, the model's precision at the optimal threshold was 80%, which indicates that 80% of the positive predictions were indeed correct. The F1 score, calculated as the harmonic mean of the precision and recall, was 0.84. This high F1 score demonstrates the balanced performance of the model in terms of the precision and recall, highlighting its strength in identifying true positives while maintaining a reasonable level of false positives.

These outcomes indicate the commendable capability of our model to differentiate between women at risk of developing breast cancer, although there is room for improvement in minimizing the false positives. Upon comparing our ML-based model with traditional statistical-based models referenced in the introduction, such as the Gail model, BOADICEA, BRCAPRO, and IBIS, we observed the notable superiority of our ML approach. Traditional models, such as the BOADICEA and BCRAT, have shown an AUC between 0.53 and 0.64 [16–20]. There is a 36 to 47% chance that these models will not identify high-risk women, while some low-risk women may receive unnecessary preventive treatments [19,38].

Our analysis underscores how our ML model surpasses these constraints, offering a more accurate and personalized risk prediction, potentially enhancing the identification of women who would benefit most from targeted screening or preventive interventions. Moreover, the inclusion of lifestyle variables, hormonal dynamics, and mutational status in our model emphasizes the importance of considering a broad spectrum of risk factors, beyond those traditionally used in statistical models.

Our study offers valuable insights, but it is important to acknowledge certain limitations that should be taken into account when interpreting the findings. Firstly, this study is focused on a single familial cancer center, limiting the generalizability of the findings to broader populations. Secondly, we lack information on diet and physical activity, which play a central role in lifestyle factors and especially in breast cancer [50,64]. In addition, the exclusion of patients with concomitant tumors or metastases ensures that the study focuses solely on primary breast cancer, which is crucial for understanding the specific risk factors associated with the initial onset of the disease. However, this exclusion may limit the generalizability of the findings to those with more complex cancer histories, where multiple malignancies could interact in ways not captured by this study. Furthermore, by excluding younger and older women, as well as those with other tumors or metastases, the generalizability of the findings is limited. This specific focus provides valuable insights into the breast cancer risk for the selected demographic but may not fully apply to broader populations. Consequently, while our findings are significant for understanding the risk in genetically predisposed women, further studies are needed to explore these factors in more diverse groups. Another significant limitation is the assessment of only one mutation (BRCA genes). Genetic factors are crucial to understanding the cancer risk; however, our work does not delve deeply into this aspect. Not collecting extensive genetic data on a large scale restricts our capacity to thoroughly investigate the relationship between genetic factors and lifestyle influences.

5. Conclusions

Early identification of women at exceptionally high risk of developing breast cancer is pivotal, offering them opportunities for risk-reducing surgery, preventive treatments, and tailored screening programs. While breast cancer screening risk stratification models exist, they primarily rely on factors such as genetics, family history, and reproductive factors. Moreover, these statistical models might lack real-time availability during routine screenings and could exhibit limited discriminatory capabilities. Many of these models focus on BRCA1 and BRCA2 mutations, which may introduce bias by overlooking other genetic variations that contribute to the breast cancer risk. This narrow focus could limit the model's applicability to individuals with less common mutations and reduce its generalizability.

Our study, which also incorporates lifestyle factors and hormonal influences, attempts to address some of these issues by providing a more nuanced understanding of the differences between healthy individuals and those already diagnosed with breast cancer. By leveraging the power of ML, we have surpassed the limitations of traditional statistical models. However, the reliance on BRCA mutations as a primary genetic factor may reduce the model's comprehensiveness. The AI-driven model developed in our study, unlike existing statistical tools such as the Gail model, BOADICEA, BRCAPRO, and IBIS, offers a more advanced and personalized approach to predicting high-risk patients. Still, to further enhance the accuracy, future research should consider the inclusion of additional genetic factors beyond the BRCA genes, such as PALB2, CHEK2, and other rare mutations. Moreover, incorporating a wider range of lifestyle and environmental variables could provide a more holistic understanding of the breast cancer risk. In conclusion, the integration of ML into breast cancer risk prediction offers significant advantages over traditional statistical methods. The ability of our model to integrate and analyze a vast array of risk factors, coupled with its superior accuracy and discriminative power, lays the groundwork for future developments in personalized medicine and breast cancer prevention.

Author Contributions: Conceptualization, L.C. and E.D.M.; methodology, L.C.; validation, L.C., E.D.M. and G.D.N.; formal analysis, L.C.; investigation, L.C.; data curation, L.C., E.R., E.C. and P.T.; writing—original draft preparation, L.C.; writing—review and editing, E.D.M. and G.D.N.; visualization, E.D.M. and G.D.N.; supervision, E.D.M. and G.D.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Bioethical Committee of the IRCCS Tumor Institute "Giovanni Paolo II", Bari, Italy, under the protocol number 1695/CEL, dated 10 June 2024.

Informed Consent Statement: Informed consent was obtained from all the subjects involved in the study. Written informed consent has been obtained from the patients to publish this paper.

Data Availability Statement: Data are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef] [PubMed]
2. Amuta, A.O.; Mkuu, R.S.; Jacobs, W.; Ejembi, A.Z. Influence of Cancer Worry on Four Cancer Related Health Protective Behaviors among a Nationally Representative Sample: Implications for Health Promotion Efforts. *J. Cancer Educ.* **2018**, *33*, 1002–1010. Available online: http://www.ncbi.nlm.nih.gov/pubmed/28251521 (accessed on 6 August 2024). [CrossRef] [PubMed]
3. AIOM-AIRTUM-Siapec-Iap. I Numeri del Cancro in Italia 2022. Available online: https://www.aiom.it/wp-content/uploads/2022/12/2022_AIOM_NDC-web.pdf (accessed on 10 October 2023).
4. Conte, L.; De Nunzio, G.; Lupo, R.; Mieli, M.; Lezzi, A.; Vitale, E.; Carriero, M.C.; Calabrò, A.; Carvello, M.; Rubbi, I.; et al. Breast Cancer Prevention: The Key Role of Population Screening, Breast Self-Examination (BSE) and Technological Tools. Survey of Italian Women. *J. Cancer Educ.* **2023**, *38*, 1728–1742. [CrossRef] [PubMed]

5. Conte, L.; Lupo, R.; Lezzi, A.; Sciolti, S.; Rubbi, I.; Carvello, M.; Calabrò, A.; Botti, S.; Fanizzi, A.; Massafra, R.; et al. Breast Cancer Prevention Practices and Knowledge in Italian and Chinese Women in Italy: Clinical Checkups, Free NHS Screening Adherence, and Breast Self-Examination (BSE). *J. Cancer Educ.* **2024**. [CrossRef] [PubMed]
6. Conte, L.; Lupo, R.; Sciolti, S.; Lezzi, A.; Rubbi, I.; Botti, S.; Carvello, M.; Fanizzi, A.; Massafra, R.; Vitale, E.; et al. Exploring the Landscape of Breast Cancer Prevention among Chinese Residents in Italy: An In-Depth Analysis of Screening Adherence, Breast Self-Examination (BSE) Practices, the Role of Technological Tools, and Misconceptions Surrounding Risk Factors and Sy. *Int. J. Environ. Res. Public. Health* **2024**, *21*, 308. Available online: https://www.mdpi.com/1660-4601/21/3/308 (accessed on 6 August 2024). [CrossRef]
7. Conte, L.; Tafuri, B.; Portaluri, M.; Galiano, A.; Maggiulli, E.; De Nunzio, G. Breast Cancer Mass Detection in DCE–MRI Using Deep-Learning Features Followed by Discrimination of Infiltrative vs. In Situ Carcinoma through a Machine-Learning Approach. *Appl. Sci.* **2020**, *10*, 6109. Available online: https://www.mdpi.com/2076-3417/10/17/6109 (accessed on 6 August 2024).
8. Tafuri, B.; Conte, L.; Portaluri, M.; Galiano, A.; Maggiulli, E.; De Nunzio, G. Radiomics for the Discrimination of Infiltrative vs. In Situ Breast Cancer. *Biomed. J. Sci. Tech. Res.* **2019**, *24*, 17890–17893.
9. Conte, L.; Rizzo, E.; Grassi, T.; Bagordo, F.; De Matteis, E.; De Nunzio, G. Artificial Intelligence Techniques and Pedigree Charts in Oncogenetics: Towards an Experimental Multioutput Software System for Digitization and Risk Prediction. *Computation* **2024**, *12*, 47. Available online: https://www.mdpi.com/2079-3197/12/3/47 (accessed on 6 August 2024). [CrossRef]
10. Hereditary Breast Cancer and BRCA Genes. Centers for Disease Control and Prevention 2023. Available online: https://www.cdc.gov/cancer/breast/young_women/bringyourbrave/hereditary_breast_cancer/index.htm (accessed on 6 August 2024).
11. Baretta, Z.; Mocellin, S.; Goldin, E.; Olopade, O.I.; Huo, D. Effect of BRCA germline mutations on breast cancer prognosis: A systematic review and meta-analysis. *Medicine* **2016**, *95*, e4975. Available online: http://www.ncbi.nlm.nih.gov/pubmed/27749552 (accessed on 6 August 2024). [CrossRef]
12. Gail, M.H.; Brinton, L.A.; Byar, D.P.; Corle, D.K.; Green, S.B.; Schairer, C.; Mulvihill, J.J. Projecting Individualized Probabilities of Developing Breast Cancer for White Females Who Are Being Examined Annually. *JNCI J. Natl. Cancer Inst.* **1989**, *81*, 1879–1886. Available online: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/81.24.1879 (accessed on 6 August 2024). [CrossRef]
13. Claus, E.B.; Risch, N.; Thompson, W.D. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast Cancer Res. Treat.* **1993**, *28*, 115–120. Available online: http://link.springer.com/10.1007/BF00666424 (accessed on 6 August 2024). [CrossRef] [PubMed]
14. Antoniou, A.C.; Cunningham, A.P.; Peto, J.; Evans, D.G.; Lalloo, F.; Narod, S.A.; A Risch, H.; E Eyfjord, J.; Hopper, J.L.; Southey, M.C.; et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions. *Br. J. Cancer* **2008**, *98*, 1457–1466. Available online: https://www.nature.com/articles/6604305 (accessed on 6 August 2024). [CrossRef] [PubMed]
15. Ahn, J.S.; Shin, S.; Yang, S.-A.; Park, E.K.; Kim, K.H.; Cho, S.I.; Ock, C.-Y.; Kim, S. Artificial Intelligence in Breast Cancer Diagnosis and Personalized Medicine. *J. Breast Cancer* **2023**, *26*, 405–435. Available online: http://www.ncbi.nlm.nih.gov/pubmed/37926067 (accessed on 6 August 2024). [CrossRef] [PubMed]
16. Wang, X.; Huang, Y.; Li, L.; Dai, H.; Song, F.; Chen, K. Assessment of performance of the Gail model for predicting breast cancer risk: A systematic review and meta-analysis with trial sequential analysis. *Breast Cancer Res.* **2018**, *20*, 18. Available online: http://www.ncbi.nlm.nih.gov/pubmed/29534738 (accessed on 6 August 2024). [CrossRef]
17. Amir, E.; Evans, D.G.; Shenton, A.; Lalloo, F.; Moran, A.; Boggis, C.; Wilson, M.; Howell, A. Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J. Med. Genet.* **2003**, *40*, 807–814. Available online: http://www.ncbi.nlm.nih.gov/pubmed/14627668 (accessed on 6 August 2024). [CrossRef]
18. Brentnall, A.R.; Harkness, E.F.; Astley, S.M.; Donnelly, L.S.; Stavrinos, P.; Sampson, S.; Fox, L.; Sergeant, J.C.; Harvie, M.N.; Wilson, M.; et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Res.* **2015**, *17*, 147. Available online: http://www.ncbi.nlm.nih.gov/pubmed/26627479 (accessed on 6 August 2024). [CrossRef]
19. Meads, C.; Ahmed, I.; Riley, R.D. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Res. Treat.* **2012**, *132*, 365–377. Available online: http://www.ncbi.nlm.nih.gov/pubmed/22037780 (accessed on 6 August 2024). [CrossRef]
20. Tice, J.A.; Cummings, S.R.; Smith-Bindman, R.; Ichikawa, L.; Barlow, W.E.; Kerlikowske, K. Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Ann. Intern. Med.* **2008**, *148*, 337–347. Available online: http://www.ncbi.nlm.nih.gov/pubmed/18316752 (accessed on 6 August 2024). [CrossRef]
21. Lee, A.J.; Cunningham, A.P.; Tischkowitz, M.; Simard, J.; Pharoah, P.D.; Easton, D.F.; Antoniou, A.C. Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. *Genet. Med.* **2016**, *18*, 1190–1198. Available online: https://linkinghub.elsevier.com/retrieve/pii/S1098360021014118 (accessed on 6 August 2024). [CrossRef]
22. Brentnall, A.R.; Cuzick, J. Risk Models for Breast Cancer and Their Validation. *Stat. Sci.* **2020**, *35*, 14–30. Available online: http://www.ncbi.nlm.nih.gov/pubmed/32226220 (accessed on 6 August 2024). [CrossRef]
23. Parmigiani, G.; Berry, D.A.; Aguilar, O. Determining Carrier Probabilities for Breast Cancer–Susceptibility Genes BRCA1 and BRCA2. *Am. J. Hum. Genet.* **1998**, *62*, 145–158. Available online: https://linkinghub.elsevier.com/retrieve/pii/S0002929707601323 (accessed on 6 August 2024). [CrossRef] [PubMed]

24. Tyrer, J.; Duffy, S.W.; Cuzick, J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **2004**, *23*, 1111–1130. [CrossRef] [PubMed]
25. Mavaddat, N.; Michailidou, K.; Dennis, J.; Fachal, L.; Lee, A.; Tyrer, J.P.; Chen, T.-H.; Wang, Q.; Bolla, M.K.; Yang, X.; et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am. J. Hum. Genet.* **2019**, *104*, 21–34. Available online: https://linkinghub.elsevier.com/retrieve/pii/S0002929718304051 (accessed on 6 August 2024). [CrossRef] [PubMed]
26. Tice, J.A.; Miglioretti, D.L.; Li, C.-S.; Vachon, C.M.; Gard, C.C.; Kerlikowske, K. Breast Density and Benign Breast Disease: Risk Assessment to Identify Women at High Risk of Breast Cancer. *J. Clin. Oncol.* **2015**, *33*, 3137–3143. [CrossRef]
27. Rice, M.S.; Tworoger, S.S.; Hankinson, S.E.; Tamimi, R.M.; Eliassen, A.H.; Willett, W.C.; Colditz, G.; Rosner, B. Breast cancer risk prediction: An update to the Rosner–Colditz breast cancer incidence model. *Breast Cancer Res. Treat.* **2017**, *166*, 227–240. Available online: http://link.springer.com/10.1007/s10549-017-4391-5 (accessed on 6 August 2024). [CrossRef]
28. Zhang, X.; Rice, M.; Tworoger, S.S.; Rosner, B.A.; Eliassen, A.H.; Tamimi, R.M.; Joshi, A.D.; Lindstrom, S.; Qian, J.; Colditz, G.A.; et al. Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study. Zheng W, editor. *PLoS Med.* **2018**, *15*, e1002644. [CrossRef]
29. Rockhill, B.; Spiegelman, D.; Byrne, C.; Hunter, D.J.; Colditz, G.A. Validation of the Gail et al. Model of Breast Cancer Risk Prediction and Implications for Chemoprevention. *JNCI J. Natl. Cancer Inst.* **2001**, *93*, 358–366. Available online: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/93.5.358 (accessed on 6 August 2024). [CrossRef]
30. Sharpe, M.; Strong, V.; Allen, K.; Rush, R.; Postma, K.; Tulloh, A.; Maguire, P.; House, A.; Ramirez, A.; Cull, A. Major depression in outpatients attending a regional cancer centre: Screening and unmet treatment needs. *Br. J. Cancer* **2004**, *90*, 314–320. Available online: https://www.nature.com/articles/6601578 (accessed on 6 August 2024). [CrossRef]
31. Gray, S.W.; Martins, Y.; Feuerman, L.Z.; Biesecker, B.B.; Christensen, K.D.; Joffe, S.; Rini, C.; Veenstra, D.; McGuire, A.L. Social and behavioral research in genomic sequencing: Approaches from the Clinical Sequencing Exploratory Research Consortium Outcomes and Measures Working Group. *Genet. Med.* **2014**, *16*, 727–735. Available online: https://linkinghub.elsevier.com/retrieve/pii/S1098360021031725 (accessed on 6 August 2024). [CrossRef]
32. Lazzeroni, L.C. Linkage Disequilibrium and Gene Mapping: An Empirical Least-Squares Approach. *Am. J. Hum. Genet.* **1998**, *62*, 159–170. Available online: https://linkinghub.elsevier.com/retrieve/pii/S0002929707601335 (accessed on 6 August 2024). [CrossRef]
33. Euhus, D.M.; Smith, K.C.; Robinson, L.; Stucky, A.; Olopade, O.I.; Cummings, S.; Garber, J.E.; Chittenden, A.; Mills, G.B.; Rieger, P.; et al. Pretest Prediction of BRCA1 or BRCA2 Mutation by Risk Counselors and the Computer Model BRCAPRO. *JNCI J. Natl. Cancer Inst.* **2002**, *94*, 844–851. Available online: https://academic.oup.com/jnci/article/94/11/844/2519752 (accessed on 6 August 2024). [CrossRef] [PubMed]
34. Jo, H.M.; Lee, E.H.; Ko, K.; Kang, B.J.; Cha, J.H.; Yi, A.; Jung, H.K.; Jun, J.K. Prevalence of Women with Dense Breasts in Korea: Results from a Nationwide Cross-sectional Study. *Cancer Res. Treat.* **2019**, *51*, 1295–1301. Available online: https://pubmed.ncbi.nlm.nih.gov/30699499/ (accessed on 6 August 2024). [CrossRef] [PubMed]
35. Sprague, B.L.; Gangnon, R.E.; Burt, V.; Trentham-Dietz, A.; Hampton, J.M.; Wellman, R.D.; Kerlikowske, K.; Miglioretti, D.L. Prevalence of Mammographically Dense Breasts in the United States. *JNCI J. Natl. Cancer Inst.* **2014**, *106*, dju255. Available online: https://academic.oup.com/jnci/article-lookup/doi/10.1093/jnci/dju255 (accessed on 6 August 2024). [CrossRef] [PubMed]
36. Rosner, B.; Colditz, G.A. Nurses' Health Study: Log-Incidence Mathematical Model of Breast Cancer Incidence. *JNCI J. Natl. Cancer Inst.* **1996**, *88*, 359–364. [CrossRef]
37. Habel, L.A.; Dignam, J.J.; Land, S.R.; Salane, M.; Capra, A.M.; Julian, T.B. Mammographic Density and Breast Cancer After Ductal Carcinoma In Situ. *JNCI J. Natl. Cancer Inst.* **2004**, *96*, 1467–1472. [CrossRef]
38. Ming, C.; Viassolo, V.; Probst-Hensch, N.; Chappuis, P.O.; Dinov, I.D.; Katapodi, M.C. Machine learning techniques for personalized breast cancer risk prediction: Comparison with the BCRAT and BOADICEA models. *Breast Cancer Res.* **2019**, *21*, 75. [CrossRef]
39. Rao, A.R.; Wang, H.; Gupta, C. Predictive Analysis for Optimizing Port Operations. Available online: http://arxiv.org/abs/2401.14498 (accessed on 25 January 2024).
40. Yala, A.; Lehman, C.; Schuster, T.; Portnoi, T.; Barzilay, R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* **2019**, *292*, 60–66. [CrossRef]
41. Weng, S.F.; Reps, J.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? Liu B, editor. *PLoS ONE* **2017**, *12*, e0174944. [CrossRef]
42. Feng, J.; Phillips, R.V.; Malenica, I.; Bishara, A.; Hubbard, A.E.; Celi, L.A.; Pirracchio, R. Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. *npj Digit. Med.* **2022**, *5*, 66. Available online: https://www.nature.com/articles/s41746-022-00611-y (accessed on 6 August 2024). [CrossRef]
43. Data Mining Elsevier; 2012. Available online: https://linkinghub.elsevier.com/retrieve/pii/C20090618195 (accessed on 6 August 2024).
44. Hastie, T.; Tibshirabni, R.; Jerome, F. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: Berlin/Heidelberg, Germany, 2009.
45. Han, J.; Kamber, M.; Pei, J. *Data Mining. Concepts and Techniques*, 3rd ed.; The Morgan Kaufmann Series in Data Management Systems; University of Illinois at Urbana-Champaign: Champaign, IL, USA, 2012.

46. Sun, Y.-S.; Zhao, Z.; Yang, Z.-N.; Xu, F.; Lu, H.-J.; Zhu, Z.-Y.; Shi, W.; Jiang, J.; Yao, P.-P.; Zhu, H.-P. Risk Factors and Preventions of Breast Cancer. *Int. J. Biol. Sci.* **2017**, *13*, 1387–1397. Available online: http://www.ncbi.nlm.nih.gov/pubmed/29209143 (accessed on 6 August 2024). [CrossRef]
47. Dehesh, T.; Fadaghi, S.; Seyedi, M.; Abolhadi, E.; Ilaghi, M.; Shams, P.; Ajam, F.; Mosleh-Shirazi, M.A.; Dehesh, P. The relation between obesity and breast cancer risk in women by considering menstruation status and geographical variations: A systematic review and meta-analysis. *BMC Womens Health* **2023**, *23*, 392. Available online: http://www.ncbi.nlm.nih.gov/pubmed/37496015 (accessed on 6 August 2024). [CrossRef] [PubMed]
48. Cortesi, L.; Galli, G.R.; Domati, F.; Conte, L.; Manca, L.; Berio, M.A.; Toss, A.; Iannone, A.; Federico, M. Obesity in Postmenopausal Breast Cancer Patients: It Is Time to Improve Actions for a Healthier Lifestyle. The Results of a Comparison Between Two Italian Regions With Different "Presumed" Lifestyles. *Front. Oncol.* **2021**, *11*, 769683. [CrossRef] [PubMed]
49. Chan, D.S.M.; Vieira, A.R.; Aune, D.; Bandera, E.V.; Greenwood, D.C.; McTiernan, A.; Rosenblatt, D.N.; Thune, I.; Vieira, R.; Norat, T. Body mass index and survival in women with breast cancer-systematic literature review and meta-analysis of 82 follow-up studies. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **2014**, *25*, 1901–1914. Available online: http://www.ncbi.nlm.nih.gov/pubmed/24769692 (accessed on 6 August 2024). [CrossRef] [PubMed]
50. Picon-Ruiz, M.; Morata-Tarifa, C.; Valle-Goffin, J.J.; Friedman, E.R.; Slingerland, J.M. Obesity and adverse breast cancer risk and outcome: Mechanistic insights and strategies for intervention. *CA Cancer J. Clin.* **2017**, *67*, 378–397. Available online: http://www.ncbi.nlm.nih.gov/pubmed/28763097 (accessed on 6 August 2024). [CrossRef]
51. Lee, K.; Kruper, L.; Dieli-Conwright, C.M.; Mortimer, J.E. The Impact of Obesity on Breast Cancer Diagnosis and Treatment. *Curr. Oncol. Rep.* **2019**, *21*, 41. Available online: http://link.springer.com/10.1007/s11912-019-0787-1 (accessed on 6 August 2024). [CrossRef]
52. Jones, M.E.; Schoemaker, M.J.; Wright, L.B.; Ashworth, A.; Swerdlow, A.J. Smoking and risk of breast cancer in the Generations Study cohort. *Breast Cancer Res.* **2017**, *19*, 118. Available online: http://www.ncbi.nlm.nih.gov/pubmed/29162146 (accessed on 6 August 2024). [CrossRef]
53. Bishop, J.D.; Killelea, B.K.; Chagpar, A.B.; Horowitz, N.R.; Lannin, D.R. Smoking and breast cancer recurrence after breast conservation therapy. *Int. J. Breast Cancer* **2014**, *2014*, 327081. Available online: http://www.ncbi.nlm.nih.gov/pubmed/24693439 (accessed on 6 August 2024). [CrossRef]
54. Padubidri, A.N.; Yetman, R.; Browne, E.; Lucas, A.; Papay, F.; Larive, B.; Zins, J. Complications of postmastectomy breast reconstructions in smokers, ex-smokers, and nonsmokers. *Plast. Reconstr. Surg.* **2001**, *107*, 342–349, discussion 350–351. Available online: http://www.ncbi.nlm.nih.gov/pubmed/11214048 (accessed on 6 August 2024). [CrossRef]
55. Taylor, C.; Correa, C.; Duane, F.K.; Aznar, M.C.; Anderson, S.J.; Bergh, J.; Dodwell, D.; Ewertz, M.; Gray, R.; Jagsi, R.; et al. Estimating the Risks of Breast Cancer Radiotherapy: Evidence From Modern Radiation Doses to the Lungs and Heart and From Previous Randomized Trials. *J. Clin. Oncol.* **2017**, *35*, 1641–1649. Available online: http://www.ncbi.nlm.nih.gov/pubmed/28319436 (accessed on 6 August 2024). [CrossRef]
56. Dall, G.V.; Britt, K.L. Estrogen Effects on the Mammary Gland in Early and Late Life and Breast Cancer Risk. *Front. Oncol.* **2017**, *7*, 110. Available online: http://www.ncbi.nlm.nih.gov/pubmed/28603694 (accessed on 6 August 2024). [CrossRef]
57. Chowdhury, R.; Sinha, B.; Sankar, M.J.; Taneja, S.; Bhandari, N.; Rollins, N.; Bahl, R.; Martines, J. Breastfeeding and maternal health outcomes: A systematic review and meta-analysis. *Acta Paediatr.* **2015**, *104*, 96–113. Available online: http://www.ncbi.nlm.nih.gov/pubmed/26172878 (accessed on 6 August 2024). [CrossRef] [PubMed]
58. Kim, S.; Ko, Y.; Lee, H.J.; Lim, J.-E. Menopausal hormone therapy and the risk of breast cancer by histological type and race: A meta-analysis of randomized controlled trials and cohort studies. *Breast Cancer Res. Treat.* **2018**, *170*, 667–675. Available online: http://www.ncbi.nlm.nih.gov/pubmed/29713854 (accessed on 6 August 2024). [CrossRef] [PubMed]
59. Yuan, X.; Yi, F.; Hou, C.; Lee, H.; Zhong, X.; Tao, P.; Li, H.; Xu, Z.; Li, J. Induced Abortion, Birth Control Methods, and Breast Cancer Risk: A Case-Control Study in China. *J. Epidemiol.* **2019**, *29*, 173–179. Available online: http://www.ncbi.nlm.nih.gov/pubmed/30101815 (accessed on 6 August 2024). [CrossRef] [PubMed]
60. Fitzpatrick, D.; Pirie, K.; Reeves, G.; Green, J.; Beral, V. Combined and progestagen-only hormonal contraceptives and breast cancer risk: A UK nested case-control study and meta-analysis. *PLoS Med.* **2023**, *20*, e1004188. Available online: http://www.ncbi.nlm.nih.gov/pubmed/36943819 (accessed on 6 August 2024). [CrossRef] [PubMed]
61. Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol.* **2012**, *13*, 1141–1151. Available online: https://linkinghub.elsevier.com/retrieve/pii/S1470204512704254 (accessed on 6 August 2024). [CrossRef]
62. Huber, D.; Seitz, S.; Kast, K.; Emons, G.; Ortmann, O. Hormone replacement therapy in BRCA mutation carriers and risk of ovarian, endometrial, and breast cancer: A systematic review. *J. Cancer Res. Clin. Oncol.* **2021**, *147*, 2035–2045. Available online: http://www.ncbi.nlm.nih.gov/pubmed/33885953 (accessed on 6 August 2024). [CrossRef]

63. Deli, T.; Orosz, M.; Jakab, A. Hormone Replacement Therapy in Cancer Survivors—Review of the Literature. *Pathol. Oncol. Res.* **2020**, *26*, 63–78. Available online: http://www.ncbi.nlm.nih.gov/pubmed/30617760 (accessed on 6 August 2024). [CrossRef]
64. Conte, L.; Lupo, R.; Lezzi, A.; Paolo, V.; Rubbi, I.; Rizzo, E.; Carvello, M.; Calabrò, A.; Botti, S.; De Matteis, E.; et al. A Nationwide Cross-Sectional Study Investigating Adherence to the Mediterranean Diet, Smoking, Alcohol and Work Habits, Hormonal dynamics between Breast Cancer Cases and Healthy Subjects. *Clin. Nutr. Open Sci.* **2024**, *55*, 1–19. Available online: https://linkinghub.elsevier.com/retrieve/pii/S2667268524000135 (accessed on 6 August 2024). [CrossRef]