significant number of relevant papers were extracted and are presented in this review. As mentioned above, from the initial group of papers we selected a representative list that follows a well-organized structure. Specifically, we selected these studies that make use of recognizable ML techniques and integrated data from heterogeneous sources in order to predict the desirable outcome. We focused mainly on studies that have been published the last 5 years as an aim to present the most recent state of the art in the field and their advances in comparison to older publications. Tables 1a, 1b, and 1c depict some of the publications presented in this review. Cancer type, ML method, number of patients, type of data as well as the overall accuracy achieved by each proposed method are presented. Each sub-table corresponds to studies regarding a specific scenario (i.e. cancer susceptibility prediction, cancer recurrence prediction and cancer survival prediction). It should be noted that in articles that more than one ML techniques are applied for prediction, we decided to present here the most accurate predictive model.

A detailed analysis of more recent studies revealed that there is a growing trend in risk assessment as well as the prediction of recurrence of a cancer type regardless the ML technique used. Many research groups have tried to predict the possibility of redeveloping cancer after remission and appeared to improve the accuracy of predictions compared to alternative statistical techniques. Moreover, the vast majority of these publications used molecular and clinical data in order to make their predictions. The use of such measurable features as input data is a growing trend based on the advent of HTTs.

In the following, we are going to discuss one case for each of the objectives of predicting (i) susceptibility, (ii) recurrence and (iii) survival, all by means of ML techniques. Each sub-section summarizes the representative studies we have selected based on their predictive outcomes. We only selected those publications that have been accepted the last 5 years and make use of distinguishable ML methods. We provide the readers with the appropriate details of the most recent techniques used for the prediction and prognosis of most frequent cancer types.

4.1. Prediction of cancer susceptibility

We performed a Scopus and a PubMed advanced search which was limited to the last 5 years. Out of these results one of the publications employs ML techniques for the prediction of susceptibility in a cancer type [55]. The authors perform a genetic epidemiology study of bladder cancer susceptibility in terms of Learning Classifying Systems (LCSs). We decided to exclude this work from the present case study as it deals with genetic information and examines further genetic problems. Based on these limitations we continued our search to the specific biomedical databases. Most of these titles neither referred to the specified keywords that are mentioned in the relevant survey nor used ML techniques for their predictions. Among the most recent publications that resulted after our limited literature search regarding the cancer risk assessment prediction [19,56–58], we selected a recent and very interesting study to present relevant to the breast cancer risk estimation by means of ANNs [19]. It is a different study among the others presented in this review article regarding the data type used. Although all of the publications selected make use of molecular, clinical or population-based data, this work encompasses mammographic findings and demographic characteristics to the model. Even though this work doesn't fit our general statement regarding our search criteria, we decided to include it in this case study because no other search result met our needs. We excluded this work from our general statement because no other search result met our needs. The major intense in developing decision-making tools that can discriminate among benign and malignant findings in breast cancer is commented by the authors. They also mention that when developing prediction models, risk stratification is of major interest. According to their knowledge, existing studies based on the use of computer models, have also utilized specific ML techniques, such as ANNs, in order to assess the risk of breast cancer patients. In their work, ANNs are employed in order to develop a prediction model that could classify malignant mammographic findings from benign. They built their model with a large number of hidden layers which generalizes better than networks with small number of hidden nodes. Regarding the collected data in this study, 48.774 mammographic findings as well as demographic risks factors and tumor characteristics were considered. All of the mammographic records were reviewed by radiologists and the reading information was obtained. This dataset was then fed as input to the ANN model. Its performance was estimated by means of ten-fold cross validation. Additionally, in order to prevent the case of overfitting the authors used the ES approach. This procedure, generally, controls the network error during training and stops it if overfitting occurs. The calculated AUC of their model was 0.965 following training and testing by means of ten-fold cross validation. The authors claimed that their model can accurately estimate the risk assessment of breast cancer patients by integrating a large data sample. They also declared that their model is unique among others if we consider that the most important factors they used to train the ANN model are the mammography findings with tumor registry outcomes. One very interesting characteristic in this study is the calculation of two main components of accuracy, namely discrimination and calibration. Discrimination is a metric that someone calculates in order to separate benign abnormalities from malignant ones, while calibration is a measurement used when a risk prediction model aims to stratify patients into high or low risk categories. The authors plotted (i) a ROC curve in order to evaluate the discriminative ability of their model and (ii) a calibration curve for comparing afterwards their model's calibration to the perfect calibration of predicting breast cancer risk. Apart from these findings, the authors also noted that the use of a mix of screening and diagnostic datasets cannot be reliably separated when feeding as input to the ANN. So, in order to overcome such limitations the authors should consider the purpose of preprocessing steps for transforming the raw data into appropriate formats for subsequent analysis.

4.2. Prediction of cancer recurrence

Based on our survey, we here present the most relevant and recent publications that proposed the use of ML techniques for cancer recurrence prediction. A work which studies the recurrence prediction of

Table 1a
Publications relevant to ML methods used for cancer susceptibility prediction.

Publication	Method	Cancer type	No of patients	Type of data	Accuracy	Validation method	Important features
Ayer T et al. [19]	ANN	Breast cancer	62,219	Mammographic, demographic	AUC = 0.965	10-fold cross validation	Age, mammography findings
Waddell M et al. [44]	SVM	Multiple myeloma	80	SNPs	71%	Leave-one-out cross validation	snp739514, snp521522, snp994532
Listgarten J et al. [45]	SVM	Breast cancer	174	SNPs	69%	20-fold cross validation	snpCY11B2 (+) 4536 T/C snpCYP1B1 (+) 4328 C/G
Stajadinovic et al. [46]	BN	Colon carcinomatosis	53	Clinical, pathologic	AUC = 0.71	Cross-validation	Primary tumor histology, nodal staging, extent of peritoneal cancer