Research Survey for *A review of statistical and machine learning methods for modeling cancer risk using structured clinical data*

Richard Bosso

I. Introduction

THE goal of this research survey is to dissect and analyze some of the unique contributions made by *A review of statistical and machine learning methods for modeling cancer risk using structured clinical data*, within the broader context of the challenges, methodologies, limitations, and potential future research for using machine learning and structured clinical data to help improve cancer risk prediction [1]. For this reason, given that *A review of statistical and machine learning methods for modeling cancer risk using structured clinical data* was published by Richter and Khoshgoftaar in 2018 [1], this analysis will digest a variety of different machine learning methods that have been applied to the domain of cancer risk prediction since 2018, with special consideration given to future directions.

II. Background

According to the GLOBOCAN (Global Cancer Observatory) database, the year of 2020 alone saw approximately 19.3 million cancer cases worldwide, with projections suggesting that by 2040 there may be as many as 28.4 million new cases of cancer occurring on a yearly basis [2]. Cancer is typically the direct cause of roughly 10 million deaths recorded for every year, which comprises of almost 1 out of 6 deaths recorded yearly for the entire world, with some recent studies suggesting that the worldwide economic burden resulting from all types of cancer may amount to a total societal cost of $25.2 trillion from the next 3 decades to come [3]. For patients who successfully recover from cancer, it is not uncommon for anxiety and fear over the possibility of a cancer relapse to have a deeply pernicious impact on their mental health [4], and prevention of recurrence is a crucial element of many cancer treatments that can greatly impact a patient's prognosis and overall quality of life [5].

Studies have shown that early detection of cancer can significantly improve the likelihood of survival [6]. However, even though early screening is generally recognized as important for improving cancer survival rates, findings by Wender et al. [7] suggest that there may be some limitations to current cancer screening practices, some of which may be partially due to inconsistencies in demographic access to cancer screening caused by a variety of different factors. Though an analysis of the causes and effects behind these societal limitations are beyond the scope of this paper, Wender et al. make several recommendations for improving cancer screening practices, one of which suggests that more effective, risk-based cancer screening strategies should be developed to allow for more targeted cancer screening based on different types of patient information [7].

With regard to developing techniques for gauging the risk of cancer based on patient data, many different statistical modeling methods have been considered that use a wide variety of different types of patient data, all for the express purpose of learning how to improve rates of early detection for different types of cancer. One important caveat to note with their research, among other aspects, is that Richter and Khoshgoftaar [1] focused only on clinical data that satisfied certain criteria. For example, though there is a large body of research that shows how cancer modeling built with genetic and biochemical data can produce very effective results [1], [8], [9], [10], [11], Richter and Khoshgoftaar chose to exclude cancer risk prediction models built with genetic data from their literature survey, based on the argument that the practical implementation of cancer risk prediction models built with genetic or biochemical data is inherently limited by the fact that there is not enough genetic data recorded from enough people for the models to be meaningfully representative of the general population [1]. Instead, the kind of cancer risk prediction models that Richter and Khoshgoftaar do focus on are models built with a variety of other types of structured clinical data, including data from blood test results, histopathologic data detailing former cases of cancer, and even data reflecting demographics and behavior [1].

Of the 22 papers relevant to cancer risk prediction modeling that were analyzed by Richter and Khoshgoftaar in 2017 [1], 7 of these papers pertained to the prediction of cancer for patients who haven't had cancer previously, whereas the other 15 papers pertained to the problem of predicting the likelihood of cancer occurring again, suggesting that much of the research for cancer risk prediction considers both the initial risk and the recurrence of cancer to be very important events to predict [1]. The motivation for our work here is to analyze these findings made by Richter and Khoshgoftaar in 2017 [1] to then consider how structured clinical research data has been used for cancer risk prediction since then, such that we can make recommendations for future research.

III. Current Approaches for Cancer Risk Prediction with Structured Clinical Data

As outlined by Richter and Khoshgoftaar [1], commonly used traditional methods of cancer prediction have included

the TNM Classification of Malignant Tumors standard developed by the American Joint Committee on Cancer (AJCC) for cancer stage prediction [12] and the use of nomograms to predict the likelihood of cancer occurring for a patient based on multiple inputs [13]. Richter and Khoshgoftaar [1] noted several different works that were able to construct nomograms with better classification performance than was possible with the TMN Classification standard, when tested with similar data [14], [15], [16]. Generally, though Richter and Khoshgoftaar [1] found a wide variety of other types of classification models when comparing each of the works in their survey, such as Support Vector Machine (SVM) [17], [18], [19], Logistic Regression [20], [21], Random Forest [22], [23], C4.5 Decision Tree [24], [19], Survival Analysis with Cox Proportional Hazards [25], [18], and Artificial Neural Networks (ANN) [26], [27], only 8 out of the 22 research works that Richter and Khoshgoftaar could find actually compared the classification performance of multiple different models against each other given the same data [1]. Beyond model accuracy alone, different types of classification models can have different strengths and weaknesses that may be inherent to their structure and underlying algorithm, such that various types of data with differing levels of complexity might be better suited for some classification models more than others [28]. For these reasons, among others, the ability to test a variety of models with the same or similar dataset can help develop a better understanding of model performance [29].

Along with the models used in the research papers cited by Richter and Khoshgoftaar [1], feature reduction methods were also noted, since they can sometimes keep less important features in the data from causing problems with overfitting while also making certain models easier to interpret. Out of the 22 works cited, 17 did apply a feature reduction technique, with univariate analysis [30] being the most frequently used type of feature reduction method across 12 of the works cited [1].

All of the works cited by Richter and Khoshgoftaar [1] predominantly used either testing accuracy or Area Under ROC Curve (AUC) [31] as the performance metrics used to assess the quality of the models tested for cancer risk prediction. Perhaps due in part to the fact that almost a third of the works cited by them tested the efficacy of Survival Analysis with Cox Proportional Hazards [25] combined with univariate analysis [30], in spite of several instances where Survival Analysis was consistently found to be worse than other methods [18], [23], [27], Richter and Khoshgoftaar [1] generally found that the performance for many of the models in these works cited did not seem to measure up well with model performance that is often found in other medical and bioinformatics datasets [32]. Assuming that the performance of each model tested, in the works they cited, can be reliably assessed based on the provided scores for AUC [31] (this assumption may not always hold under certain conditions, particularly if a model is trained with datasets containing significant class imbalance [33]), it would seem that the highest performing cancer risk prediction model using structured clinical data that they found (with an AUC of 0.96) was a logistic regression model [20] that used an ensemble of mRMR [34], Information Gain [35], and ReliefF [36] feature ranking techniques, implemented by Cirkovic et al. [21], [1].

Before 2018, when the survey conducted by Richter and Khoshgoftaar was first published, they spent several months revising their work in 2017, so the most recent research works cited by them were published in 2016 [1]. Since 2016, a variety of different methods for using structured clinical data to generate cancer risk predictions have been noted in the research literature, and the methods used for our investigations here led us to discover much of the following research works on Google Scholar,[1] with some of this research following a criteria similar to what was defined by Richter and Khoshgoftaar for their review [1]. For our purposes, we have chosen to highlight several unique research works to understand and verify some of the more current developments in cancer risk prediction that use structured clinical data.

In 2018, Patrício et al. [37] investigated the effectiveness of using Support Vector Machine (SVM) [17], Logistic Regression [20], and Random Forest [22] models for the purpose of breast cancer detection. Monte Carlo Cross Validation [38] was used to train and test these models to generate predictions for binary classification, using a dataset consisting of patient data collected by the University Hospital Centre of Coimbra from 154 women who consented to being study participants from 2009 to 2013 [39], [37]. Univariate and multivariate feature selection techniques [30] were also applied to the training data to influence the number of features that were used by the models, and various ROC metrics [31] were used to assess which models performed the best when given different numbers of features. Based on the performance metrics and feature selection techniques that were used, they found that the SVM model generally seemed to outperform the Logistic Rogression and Random Forests models regardless of the number of features used during training, and from a total of 9 clinical features it was found that only 4 features were needed for the most optimal performance to attain an AUC of [0.87, 0.91] [37].

Since the dataset used by Patrício et al. has been made publicly available on the UC Irvine Machine Learning Repository[2] [40], we were able to access this dataset and generate our own experiments using machine learning models and performance metric evaluations that we implemented with scikit-learn [41]. For the purpose of validating and further expanding upon the work conducted by Patrício et al. [37], we compared the performance of 8 different classification models (Decision Tree [42], SVM [17], Random Forests [22], Logistic Regresssion [20], Multi Layer Perceptron (MLP) [26], AdaBoost [43], [44], k - Nearest Neighbors (kNN) [45], and Naive Bayes [46], [47]) across different numbers of features selected from the Coimbra dataset, such that 3 different feature selection methods (Mutual Information [48], Chi-square [49], and ANOVA F-value [50]) were used to

[1]https://scholar.google.com/

[2]https://archive.ics.uci.edu/

Features Selected	Decision Tree	SVM	Random Forests	Logistic Regression	Naive Bayes	Multi Layer Perceptron	kNN	AdaBoost
1	0.4913 (+/-) 0.0	0.5251 (+/-) 0.072	0.469 (+/-) 0.0217	0.5243 (+/-) 0.074	0.4788 (+/-) 0.060	0.4471 (+/-) 0.081	0.4525 (+/-) 0.079	0.448 (+/-) 0.0
2	0.4587 (+/-) 0.030	0.604 (+/-) 0.0460	0.4904 (+/-) 0.023	0.6057 (+/-) 0.043	0.6338 (+/-) 0.043	0.522 (+/-) 0.0650	0.4537 (+/-) 0.082	0.5734 (+/-) 0.0
3	0.5433 (+/-) 0.039	0.7405 (+/-) 0.039	0.637 (+/-) 0.0236	0.7498 (+/-) 0.035	0.7356 (+/-) 0.033	0.6654 (+/-) 0.054	0.4765 (+/-) 0.084	0.6747 (+/-) 0.0
4	0.4862 (+/-) 0.040	0.7692 (+/-) 0.043	0.6682 (+/-) 0.024	0.7698 (+/-) 0.041	0.7479 (+/-) 0.034	0.6841 (+/-) 0.053	0.4747 (+/-) 0.084	0.7347 (+/-) 0.006
5	0.5024 (+/-) 0.041	0.7594 (+/-) 0.048	0.6659 (+/-) 0.023	0.7678 (+/-) 0.041	0.7461 (+/-) 0.034	0.6827 (+/-) 0.057	0.4764 (+/-) 0.084	0.7363 (+/-) 0.008
6	0.6343 (+/-) 0.035	0.8028 (+/-) 0.037	0.7291 (+/-) 0.027	0.8098 (+/-) 0.037	0.7572 (+/-) 0.035	0.7587 (+/-) 0.048	0.4813 (+/-) 0.084	0.7393 (+/-) 0.023
7	0.6444 (+/-) 0.057	0.7972 (+/-) 0.038	0.7713 (+/-) 0.026	0.804 (+/-) 0.0389	0.7855 (+/-) 0.034	0.7968 (+/-) 0.059	0.4966 (+/-) 0.079	0.7909 (+/-) 0.003
8	0.6444 (+/-) 0.055	0.7897 (+/-) 0.040	0.7492 (+/-) 0.028	0.7945 (+/-) 0.039	0.7675 (+/-) 0.039	0.8036 (+/-) 0.052	0.481 (+/-) 0.0811	0.7614 (+/-) 0.003
9	0.6515 (+/-) 0.053	0.7833 (+/-) 0.042	0.7555 (+/-) 0.025	0.7884 (+/-) 0.042	0.7626 (+/-) 0.040	0.7854 (+/-) 0.052	0.4837 (+/-) 0.082	0.7633 (+/-) 0.007

TABLE I: Mean AUC recorded for each model trained with features selected from Chi-Square ranking, showing range of values within 2 standard deviations (95% confidence interval) from the mean for 500 randomized trials of 5 - fold cross validation

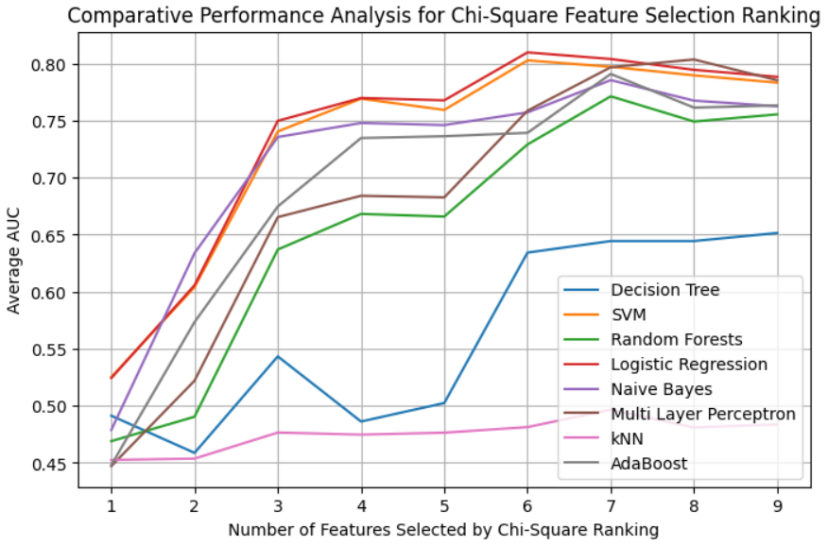

Fig. 1: Chi-Square Ranking Performance Comparison

Features	Chi_Squared_Ranking
MCP.1	214.917039
Insulin	89.203820
Glucose	88.125373
Resistin	61.949833
HOMA	45.656784
BMI	1.847119
Age	0.988417
Adiponectin	0.200949
Leptin	0.001849

Fig. 2: Chi-Square Feature Ranking

select and rank the features in order of importance. To determine average AUC performance, we generated multiple sets of 500 rounds of 5-fold cross validation and calculated the mean AUC and standard deviation of AUC scores measured from each combination of model and feature selection ranking, as shown in Tables 1 - 3 for each feature selection method. For anyone who wishes to reproduce our experimental results, the code we have used is available at the link provided below.[3] Figures 2, 5, and 6 show the features selected by each feature selection method tested, listed in the order of importance for each feature that was calculated by each ranking, and Figures 1, 3, and 4 visually represent how the AUC performance compared between each model given different numbers of feature sets.

From looking at basic trends in AUC performance for the majority of models, as can be seen in Figures 1, 3, and 4, most models generally seemed to perform significantly worse when trained with just the first 1 - 2 features recommended by Chi-Square feature selection, when compared to both Mutual Information and ANOVA F-Value feature selection.

[3]https://github.com/RBosso/A_review_of_statistical_and_machine_learning

Features Selected	Decision Tree	SVM	Random Forests	Logistic Regression	Naive Bayes	Multi Layer Perceptron	kNN	AdaBoost
1	0.6098 (+/-) 0.0	0.4625 (+/-) 0.114	0.6195 (+/-) 0.014	0.4675 (+/-) 0.113	0.6764 (+/-) 0.050	0.7042 (+/-) 0.043	0.6769 (+/-) 0.055	0.6877 (+/-) 0.0
2	0.6888 (+/-) 0.025	0.7492 (+/-) 0.032	0.7645 (+/-) 0.017	0.7636 (+/-) 0.031	0.783 (+/-) 0.0323	0.825 (+/-) 0.0296	0.7564 (+/-) 0.044	0.7753 (+/-) 0.0
3	0.6708 (+/-) 0.041	0.7744 (+/-) 0.039	0.8415 (+/-) 0.017	0.7818 (+/-) 0.038	0.8003 (+/-) 0.037	0.8156 (+/-) 0.054	0.8278 (+/-) 0.043	0.8296 (+/-) 0.0
4	0.663 (+/-) 0.0467	0.7699 (+/-) 0.040	0.7873 (+/-) 0.020	0.7776 (+/-) 0.039	0.8045 (+/-) 0.034	0.7803 (+/-) 0.049	0.8282 (+/-) 0.043	0.8198 (+/-) 0.014
5	0.6917 (+/-) 0.032	0.79 (+/-) 0.03994	0.8011 (+/-) 0.022	0.7987 (+/-) 0.038	0.7963 (+/-) 0.037	0.7954 (+/-) 0.044	0.8053 (+/-) 0.043	0.789 (+/-) 0.0136
6	0.6673 (+/-) 0.031	0.7885 (+/-) 0.042	0.7638 (+/-) 0.023	0.7953 (+/-) 0.039	0.791 (+/-) 0.0363	0.791 (+/-) 0.0463	0.795 (+/-) 0.0421	0.8071 (+/-) 0.018

TABLE II: Mean AUC recorded for each model with features selected from Mutual Information ranking, for range of values within 2 standard deviations (95% confidence interval) from the mean for 500 randomized trials of 5 - fold cross validation

Features Selected	Decision Tree	SVM	Random Forests	Logistic Regression	Naive Bayes	Multi Layer Perceptron	kNN	AdaBoost
1	0.5791 (+/-) 0.0	0.7644 (+/-) 0.025	0.604 (+/-) 0.0196	0.7644 (+/-) 0.025	0.7304 (+/-) 0.027	0.7596 (+/-) 0.029	0.6742 (+/-) 0.057	0.6998 (+/-) 0.0
2	0.5676 (+/-) 0.026	0.7582 (+/-) 0.029	0.6243 (+/-) 0.020	0.7629 (+/-) 0.030	0.7689 (+/-) 0.027	0.7585 (+/-) 0.032	0.682 (+/-) 0.0530	0.673 (+/-) 0.0
3	0.5591 (+/-) 0.029	0.7509 (+/-) 0.037	0.6185 (+/-) 0.018	0.7609 (+/-) 0.031	0.7682 (+/-) 0.027	0.7465 (+/-) 0.043	0.714 (+/-) 0.0504	0.6876 (+/-) 0.0
4	0.51 (+/-) 0.0299	0.7767 (+/-) 0.044	0.6601 (+/-) 0.021	0.7817 (+/-) 0.038	0.7697 (+/-) 0.033	0.7339 (+/-) 0.061	0.759 (+/-) 0.0425	0.7486 (+/-) 0.009
5	0.6385 (+/-) 0.035	0.815 (+/-) 0.0330	0.7284 (+/-) 0.025	0.8242 (+/-) 0.033	0.7816 (+/-) 0.033	0.7884 (+/-) 0.051	0.7742 (+/-) 0.040	0.7358 (+/-) 0.008
6	0.6368 (+/-) 0.035	0.8029 (+/-) 0.037	0.7284 (+/-) 0.026	0.8098 (+/-) 0.037	0.7572 (+/-) 0.035	0.7591 (+/-) 0.049	0.4813 (+/-) 0.084	0.7399 (+/-) 0.023
7	0.6453 (+/-) 0.057	0.7973 (+/-) 0.039	0.7722 (+/-) 0.025	0.804 (+/-) 0.0388	0.7855 (+/-) 0.034	0.7959 (+/-) 0.052	0.4966 (+/-) 0.079	0.7909 (+/-) 0.003
8	0.6437 (+/-) 0.054	0.7896 (+/-) 0.040	0.7497 (+/-) 0.027	0.7945 (+/-) 0.039	0.7675 (+/-) 0.039	0.8045 (+/-) 0.049	0.481 (+/-) 0.0811	0.7613 (+/-) 0.003
9	0.6526 (+/-) 0.056	0.7833 (+/-) 0.043	0.756 (+/-) 0.0267	0.7884 (+/-) 0.042	0.7626 (+/-) 0.040	0.7861 (+/-) 0.053	0.4837 (+/-) 0.082	0.7636 (+/-) 0.007

TABLE III: Mean AUC recorded for each model with features selected from ANOVA F-Value ranking, for range of values within 2 standard deviations (95% confidence interval) from the mean for 500 randomized trials of 5 - fold cross validation

Whereas Figure 1 seems to show that none of the models that trained with only the one top feature ranked by Chi-Square (MCP.1, as shown in Figure 2) managed to have an average AUC that was greater than 0.55, Figures 3 and 4 show that almost all models trained with just the single top ranked features recommended by Mutual Information (Age, as shown in Figure 5) and ANOVA F-Value (Glucose, as shown in Figure 6) did have average AUC values greater than 0.55, with half of all models attaining average AUC scores greater than 0.70 when trained with just the Glucose feature recommended by ANOVA F-Value feature selection. Observing all models trained with just 2 features, similarly poor performance is seen for all models trained with the 2 features recommended by Chi-Square (MCP.1 and Insulin, as shown in Figure 2) when compared to the 2 feature AUC performance noted for Mutual Information (Age and Glucose,

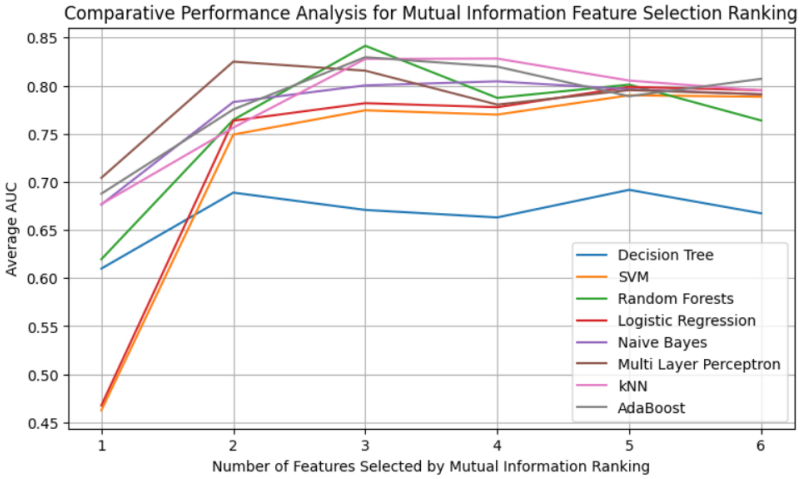

Fig. 3: Mutual Information Performance Comparison

Features	Mutual_Info_Ranking
Age	0.121419
Glucose	0.093410
Resistin	0.038398
HOMA	0.031227
Insulin	0.002506
Leptin	0.002395
BMI	0.000000
Adiponectin	0.000000
MCP.1	0.000000

Fig. 5: Mutual Information Feature Ranking

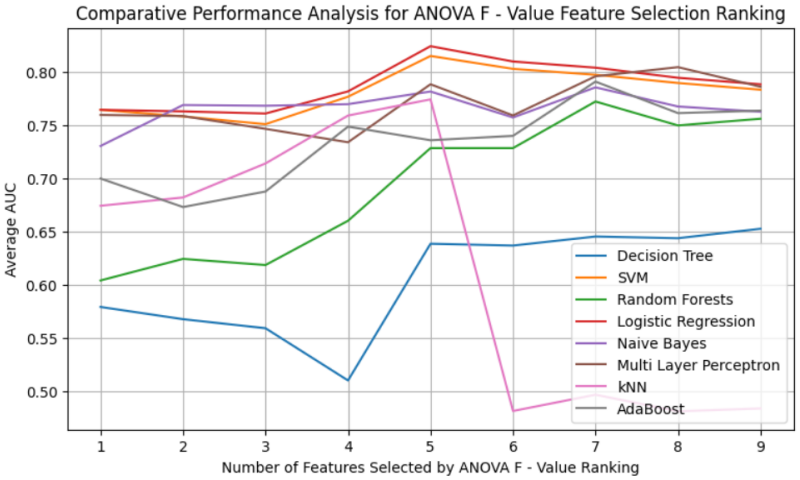

Fig. 4: ANOVA F-Value Performance Comparison

Features	ANOVA_F_Ranking
Glucose	19.755454
HOMA	10.002398
Insulin	9.459504
Resistin	6.211315
BMI	2.039862
MCP.1	0.959962
Age	0.216670
Adiponectin	0.043322
Leptin	0.000133

Fig. 6: ANOVA F-Value Feature Ranking

as shown in Figure 5) and ANOVA F-Value (Glucose and HOMA, as shown in Figure 6). These findings may suggest that Chi-Square feature selection (at least when ranking features for these 8 models tested, when trained with datasets that may be similar to the Coimbra dataset tested here) may not perform as well at ranking features, compared to Mutual Information and ANOVA F-Value feature selection, though further experiments with a larger variety of different datasets would likely be needed to support such a claim. Regardless, this seems to suggest that MCP.1 and Insulin may be less important than features like Glucose and Age for cancer risk prediction model training, which seems to align somewhat with conclusions on feature importance made by Patrício et al. [37].

Regarding model performance, there seemed to be variation between feature selection methods, with some trends in performance that seemed unique to certain models. For example, though kNN seemed to consistently show the weakest performance of any other model when trained with dataset features ranked by Chi-Square, kNN managed to show significantly better performance when compared to other models trained with features ranked by Mutual Information, and Figure 4 seems to show that kNN trained with ANOVA F-Value ranked features saw a significant drop in performance from 5 features (having an average AUC of roughly 0.77) to 6 features (an average AUC of roughly 0.48), suggesting that kNN might have had more difficulty with maintaining a stable response to larger sets of noisy/unimportant features than was seen with other models. When trained with Mutual Information features, part of the reason why kNN doesn't see this drop off in performance might be partly due to Mutual Information only generating feature rankings for a total of 6 features (note the 0 ranking values for the bottom 3 values shown in Figure 5).

While Decision Tree consistently seemed to be one of the worst performing models regardless of the feature selection method used, for most models the feature selection method seemed to have some influence on how the performance of certain models compared to other models. For instance, though SVM and Logistic Regression often showed the highest performance compared to other models that where

trained using Chi-Square and ANOVA F-Value feature selection, as can been in Figures 1 and 4, the results shown by Figure 3 seem to suggest that SVM and Logistic Regression never seemed to have the highest average AUC when trained with features selected by Mutual Information. Regardless of why this occurred, this may suggest that different models can perform very differently when given the same features in their training data. Looking at the performance results for each model tested across different sets of features in every feature selection method implemented here, Table 1 seems to show that the highest average AUC recorded using Chi-Square feature selection was attained by Logistic Regression when trained using 6 features (an average AUC of 0.8098, marginally higher than what was seen with SVM for the same set of features), Table 2 seems to show that the highest average AUC recorded using Mutual Information feature selection was attained by Random Forests when trained with 3 features (an average AUC of 0.8415), and Table 3 seems to show that the highest average AUC recorded using ANOVA F-Value feature selection was attained by Logistic Regression when trained with 5 features (an average AUC of 0.8242).

Beyond the context of our comparisons of AUC performance between these models, MLP, Random Forests, and AdaBoost had some noteworthy performance, but their training times were considerably longer than for every other model. For real world usage, where training datasets may be considerably larger, longer training times may prevent these models from being as easily scalable, which may make some of the other models tested here more desirable when training with similar structured clinical data for cancer prediction.

In 2020, Islam et al. [51] used the Wisconsin Breast Cancer Dataset, consisting of clinical data with 9 different histopathological features [52], to determine how well a variety of classification models can predict whether a tumor will remain benign or become malignant. They compared the performance of SVM [17], Logistic Regression [20], Random Forest [22], ANN [26], and kNN (k - Nearest Neighbors) [45] models to each other, and Islam et al. [51] derived their comparisons using several different performance metrics for each model tested, including AUC [31], Area Under Precision - Recall Curve (AUPRC) [53], and F1 - Score [54]. Additionally, to the best of the author's knowledge, this work done by Islam et al. [51] seems to be the only cancer risk prediction methodology among these works cited that also used the Matthews Correlation Coefficient (MCC) performance metric [55]. First developed formally by Udny Yule [56], MCC has been recognized as a reliable performance metric with some unique advantages over other methods for performance evaluation [57]. From their performance evaluations, it was found that the ANN model seemed to have the highest level of performance when compared to the other models tested [51]. As such, in addition to this work being an example of how just a few histopathological clinical features might be used to predict the likelihood of tumor malignancy, this work seems unique for their relatively diverse usage of different performance metrics [51], which may be somewhat in contrast with the majority of research works in cancer risk prediction with clinical data that were found by Richter and Khoshgoftaar [1]. It is generally important to understand the underlying assumptions that are made by different performance metrics, since different performance metrics can provide distinct interpretations of model performance, so incorrect assumptions about performance metrics can lead to incorrect conclusions when using them to assess the performance of different models [58], [59].

Published in 2021, a study conducted by Alfayez et al. [60] surveyed several examples of cancer risk prediction models being applied with structured clinical data, with a specific focus on identifying research works that dealt with generating predictions for early cancer detection using ordinary clinical data from the general population. Though all of their works cited tested different models that were specific to different types of cancer, which made meaningful conclusions about optimal model performance difficult to make given the limited number of research papers that they were able to cite, it might be valuable to note that all of the works that they discovered seemed to use datasets with relatively high levels of class imbalance [60]. The estimated number of newly reported cancer cases in the United States was approximately 2 million in 2023 [61] within a total population of over 339 million people [62], so it is reasonable to expect that the typical population of patients will most often turn up negative when being tested for any new cases of cancer, and that a large dataset labeled with these patient cancer results may oftentimes be highly imbalanced. For example, Alfayez et al. [60] noted one work (published by Richter and Khoshgoftaar [63] in 2019) that tested several models with a large dataset of clinical features from over 4 million individuals, where only 10,129 of these individuals were labeled as having confirmed cases of melanoma [60], [63]. Highly imbalanced data can present unique challenges when using predictive models [64], but real-world cancer screening data may sometimes be highly imbalanced [60], so training models with such highly imbalanced cancer prediction datasets might be essential for optimal performance with cancer risk prediction [65].

Even for cancer prediction models that do not make use of structured clinical data by itself, there may be some justification for trying to incorporate such data with other varieties of training data. Published in 2023, Kayikci and Khoshgoftaar [66] used the METABRIC dataset [67] and the TCGA-BRCA dataset [68] to develop a breast cancer prediction model that uses a multimodal combination of genetic and clinical data. The proposed model was constructed as multiple sets of Convolutional Neural Network (CNN) models [69] that have attention-based mechanisms [70] built into their CNN layer architecture, in order for the separate modalities / types of data to help generate accurate predictions [66]. They compared their proposed model with a multimodal deep neural network using multidimensional data (MDNNMD) [71] and found that their proposed model seemed to have better classification performance than MDNNMD overall, showing how standard clinical data incorporated into multimodal data can be used effectively by models that are built to process different types of data [66].

IV. Limitations with Current Approaches

For cancer risk prediction generated by machine learning methods in general, there have been numerous examples of research works that have used medical diagnostic images as model training data, such as MRI scans, CT scans, and other visual lab testing imagery, and the highest performing methods applied for image classification are oftentimes deep learning models and neural network models [72], [73], [74], [75], [76], [77], [78]. When consulting comprehensive research summaries of machine learning methods that have been applied to many different types of cancer prediction, however, there generally seem to be relatively fewer models that actually use basic clinical, behavioral, and historical patient data [79]. In the survey done by Alfayez et al. [60], given the criteria of only highlighting research that featured machine learning methods being used with commonplace behavioral and clinical data, in order to generate cancer risk predictions for adults in the general population who lacked any symptoms for cancer, they queried hundreds of papers on PubMed[4] that covered any study leading up to November of 2020 and were only able to narrow down 10 papers that satisfied their criteria [80], [81], [82], [83], [84], [85], [63], [86], [87], [88]. Among all 10 of these papers, they discovered that only one cancer risk prediction model among these research works was actually adopted by a healthcare organization at that time [89], [81], [83], and they ultimately concluded that the supply of research they could find to satisfy their conditions was not enough to make any useful performance comparisons for the different models cited [60]. A 2023 survey published by Burnett et al. [90], which used a specific set of queries to check roughly 22,000 research papers for any works that focused on using commonplace clinical data to predict colorectal cancer, could only uncover 14 research works that were actually relevant to their criteria, and they also concluded that there was too much uncertainty to suggest if one model paradigm inherently performed better than the others overall [90].

With regard to making structured clinical data more accessible in general, Richter and Khoshgoftaar [1] noted that the historical lack of engagement that clinical professionals have often had with Electronic Health Records (EHR) [91] has likely exacerbated the problem of making complete clinical data readily available. Addressing the fact that most of the clinical datasets used by the works they cited were typically from years - long clinical studies that would end half a decade or more before the datasets from these clinical studies would actually be published, Richter and Khoshgoftaar noted that this distance of time between data collection and publishing time is likely due to a variety of different factors related to complex administrative processes and policy matters, but they expressed concern that medical practices and clinical guidelines may change too quickly for models trained with such old data to generalize well with modern-day expectations for cancer treatment [1].

Beyond the limitations already noted with the relative rarity of commonly recorded types of structured clinical datasets that we have previously discussed [60], [90], there are limitations inherent to testing machine learning models with most healthcare data, and these limitations often prevent model performance results from being reproducible with real - world data [92]. Compared to most other research domains within machine learning research, McDermott et al. [92] noted several factors which cause reproducibility problems whenever machine learning methods are applied to healthcare data, finding that machine learning research papers using healthcare data usually don't make their implementation code visible to the public, often won't use data that is openly licensed for public access, and will typically not test the performance of their machine learning models with more than one healthcare dataset per paper [92]. Datasets are not always openly accessible between different healthcare organizations [1], [92], and the lack of machine learning research that independently validates and verifies model performance with other structured clinical datasets likely limits how well these models can generalise with real data [60].

The majority of the works noted here typically targeted certain types of cancer, such as breast cancer, but relatively few works that used structured clinical data actually tested models that screened for the risk of just getting any type of cancer at all [1], [60], [90]. Though machine learning models that predict the risk of one type of cancer can have good performance for that classification task [1], [90], early screening for cancer may benefit from having models that may generate predictions for a wide spectrum of different types of cancer. Survival rates for cancer are generally higher when cancer is caught at earlier stages [60], so preemptive screening might be limited if the models are built for screening only for specific types of cancer.

V. Conclusion

Conquering the societal burden of cancer is one of the most significant challenges posed to the modern healthcare system. Cancer will likely continue to threaten the lives of millions of people for years to come. Whether left untreated or detected too late, cancer can be a fatal illness, and it is one of the main causes of death worldwide. Research has clearly shown that earlier detection of cancer can often improve the chance of survival, and for this reason the facilitation of early screening for cancer is a matter of life and death. The problem is that resources are limited, and wasted resources are also a matter of life and death. For early cancer screening practices to have the best impact on humanity, screening can not proceed blindly, as the most effective screening needs to be justified by efficient foresight. Machine learning has been considered as one method for providing this foresight, but we have found that facilitating early screening with this methodology may require structured clinical data consisting of the type of basic clinical, behavioral, and demographic information that medical organizations constantly collect from the general population.

For this overview of the research literature tackling these challenges, beginning with an examination of the unique survey published by Richter and Khoshgoftaar in 2018 [1], we have analyzed the ways in which structured clinical data have helped build machine learning models to generate cancer risk

[4]https://pubmed.ncbi.nlm.nih.gov/

predictions since that time. We have noted several limitations that may be somewhat unique to this subdomain of machine learning, much of which seems connected to the limited supply of up-to-date structured clinical data that has typically been made available. Nevertheless, the research that exists has shown how a wide variety of different types of predictive models can potentially provide good methods for predicting cancer risk effectively, we have seen how a variety of different feature selection methods can help improve the performance of these models, we have seen how different performance metrics have been used to help judge the performance of these models overall, we have seen that even imbalanced structured clinical data has built various cancer prediction models with some notable success, and we have been able to verify ourselves how different machine learning models can be tested with different feature selection methods to generate cancer risk predictions with the data collected by the University Hospital Centre of Coimbra [39], [37].

As discussed, there have been several limitations within this domain of research, but with these limitations in mind we can make recommendations for future research. Presently, several surveys have suggested that there need to be more structured clinical datasets made available by clinical institutions in general, in order for the models being built to generalize as well as possible with real-world data. There need to be more machine learning models being tested with structured clinical data that has been collected as recently as possible, so that the quality of data matches more accurately with what modern-day expectations for cancer prognosis actually would be. Though some research works have been noted here that did test machine learning efficacy using highly imbalanced data [80], [81], [82], [83], [84], [85], [63], [86], [87], [88], there should be more future research that seeks to leverage and validate more imbalanced structured clinical data, particularly when such data can be collected from the general population. With regard to model testing in future works, wider varieties of machine learning models should be compared to each other more often, such that more thorough comparisons of model performance can be made, along with more comparative testing of different feature selection methods using a thorough range of different performance metrics within individual studies. Future studies should also examine how models might be developed to generate the risk of cancer more generally, rather than only focusing solely on individual types of cancer, such that early screening for cancer might have a better chance of being preemptively facilitated for unexpected or lesser known types of cancer.

REFERENCES

[1] A. N. Richter and T. M. Khoshgoftaar, "A review of statistical and machine learning methods for modeling cancer risk using structured clinical data," *Artificial intelligence in medicine*, vol. 90, pp. 1–14, 2018.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] S. Chen, Z. Cao, K. Prettner, M. Kuhn, J. Yang, L. Jiao, Z. Wang, W. Li, P. Geldsetzer, T. Bärnighausen *et al.*, "Estimates and projections of the global economic cost of 29 cancers in 204 countries and territories from 2020 to 2050," *JAMA oncology*, vol. 9, no. 4, pp. 465–472, 2023.

[4] N. M. Tauber, M. S. O'Toole, A. Dinkel, J. Galica, G. Humphris, S. Lebel, C. Maheu, G. Ozakinci, J. Prins, L. Sharpe *et al.*, "Effect of psychological intervention on fear of cancer recurrence: a systematic review and meta-analysis," *Journal of clinical oncology*, vol. 37, no. 31, p. 2899, 2019.

[5] D. A. Mahvi, R. Liu, M. W. Grinstaff, Y. L. Colson, and C. P. Raut, "Local cancer recurrence: the realities, challenges, and opportunities for new therapies," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 488–505, 2018.

[6] J. J. Ott, A. Ullrich, and A. B. Miller, "The importance of early symptom recognition in the context of early detection and cancer survival," *European Journal of Cancer*, vol. 45, no. 16, pp. 2743–2748, 2009.

[7] R. C. Wender, O. W. Brawley, S. A. Fedewa, T. Gansler, and R. A. Smith, "A blueprint for cancer screening and early detection: Advancing screening's contribution to cancer control," *CA: a cancer journal for clinicians*, vol. 69, no. 1, pp. 50–79, 2019.

[8] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *PLoS one*, vol. 8, no. 4, p. e61318, 2013.

[9] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[10] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[11] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: a review," *Bioengineering*, vol. 10, no. 2, p. 173, 2023.

[12] S. B. Edge, A. C. S. American Joint Committee on Cancer *et al.*, *AJCC cancer staging handbook: from the AJCC cancer staging manual*. Springer, 2010, vol. 19.

[13] V. P. Balachandran, M. Gonen, J. J. Smith, and R. P. DeMatteo, "Nomograms in oncology: more than meets the eye," *The lancet oncology*, vol. 16, no. 4, pp. e173–e180, 2015.

[14] O. Cahlon, M. F. Brennan, X. Jia, L.-X. Qin, S. Singer, and K. M. Alektiar, "A postoperative nomogram for local recurrence risk in extremity soft tissue sarcomas after limb-sparing surgery without adjuvant radiation," *Annals of surgery*, vol. 255, no. 2, pp. 343–347, 2012.

[15] M. R. Weiser, R. G. Landmann, M. W. Kattan, M. Gonen, J. Shia, J. Chou, P. B. Paty, J. G. Guillem, L. K. Temple, D. Schrag *et al.*, "Individualized prediction of colon cancer recurrence using a nomogram," *Journal of clinical oncology*, vol. 26, no. 3, pp. 380–385, 2008.

[16] W. A. See, "Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer: International bladder cancer nomogram consortium, bochner bh, kattan mw, vora kc, department of urology, memorial sloan-kettering cancer center, kimmel center for prostate and urologic tumors, new york, ny," in *Urologic Oncology: Seminars and Original Investigations*, vol. 25, no. 3. Elsevier, 2007, p. 275.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[18] W. Kim, K. S. Kim, J. E. Lee, D.-Y. Noh, S.-W. Kim, Y. S. Jung, M. Y. Park, and R. W. Park, "Development of novel breast cancer recurrence prediction model using support vector machine," *Journal of breast cancer*, vol. 15, no. 2, p. 230, 2012.

[19] L. G. Ahmad, A. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A. Razavi *et al.*, "Using three machine learning techniques

for predicting breast cancer recurrence," *J Health Med Inform*, vol. 4, no. 124, p. 3, 2013.

[20] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[21] B. R. A. Cirkovic, A. M. Cvetkovic, S. M. Ninkovic, and N. D. Filipovic, "Prediction models for estimation of survival rate and relapse for breast cancer patients," in *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2015, pp. 1–6.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[23] A. G. Singal, A. Mukherjee, J. B. Elmunzer, P. D. Higgins, A. S. Lok, J. Zhu, J. A. Marrero, and A. K. Waljee, "Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma," *Official journal of the American College of Gastroenterology— ACG*, vol. 108, no. 11, pp. 1723–1730, 2013.

[24] S. L. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," 1994.

[25] D. Cox and D. Oakes, *Analysis of Survival Data*, ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1984. [Online]. Available: https://books.google.com/books?id=Y4pdM2soP4IC

[26] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of microbiological methods*, vol. 43, no. 1, pp. 3–31, 2000.

[27] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and E. Alba-Conejo, "A combined neural network and decision trees model for prognosis of breast cancer relapse," *Artificial intelligence in medicine*, vol. 27, no. 1, pp. 45–63, 2003.

[28] E. W. Steyerberg, T. van der Ploeg, and B. Van Calster, "Risk prediction with machine learning and regression methods," *Biometrical Journal*, vol. 56, no. 4, pp. 601–606, 2014.

[29] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint arXiv:1811.12808*, 2018.

[30] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. Ieee, 2015, pp. 1200–1205.

[31] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[32] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "A comparative evaluation of feature ranking methods for high dimensional bioinformatics data," in *2011 IEEE International Conference on Information Reuse & Integration*. IEEE, 2011, pp. 315–320.

[33] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data–recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 245–251.

[34] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," vol. 3, 09 2003, pp. 523– 528.

[35] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Icml*, vol. 97, no. 412-420. Nashville, TN, USA, 1997, p. 35.

[36] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Ninth International Workshop on Machine Learning*, D. H. Sleeman and P. Edwards, Eds. Morgan Kaufmann, 1992, pp. 249–256.

[37] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo, "Using resistin, glucose, age and bmi to predict the presence of breast cancer," *BMC cancer*, vol. 18, pp. 1–8, 2018.

[38] Q.-S. Xu and Y.-Z. Liang, "Monte carlo cross validation," *Chemometrics and Intelligent Laboratory Systems*, vol. 56, no. 1, pp. 1–11, 2001.

[39] J. Crisostomo, P. Matafome, D. Santos-Silva, A. L. Gomes, M. Gomes, M. Patrício, L. Letra, A. B. Sarmento-Ribeiro, L. Santos, and R. Seiça, "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," *Endocrine*, vol. 53, pp. 433–442, 2016.

[40] P. J. C. J. M. P. S. R. Patrcio, Miguel and F. Caramelo, "Breast Cancer Coimbra," UCI Machine Learning Repository, 2018, DOI: https://doi.org/10.24432/C52P59.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[42] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[43] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[44] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[45] E. Fix and J. L. Hodges, "Discriminatory analysis: Nonparametric discrimination: Small sample performance," 1952.

[46] C. S. Variances, "Updating formulae and a pairwise algorithm for computing sample variances tf chan, yale university, new haven, usa gh golub and rj leveque, stanford university, usa," in *COMPSTAT 1982 5th Symposium held at Toulouse 1982: Part I: Proceedings in Computational Statistics*. Physica, 1982, p. 30.

[47] T. Bayes, "Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s," *Philosophical transactions of the Royal Society of London*, no. 53, pp. 370–418, 1763.

[48] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[49] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of 7th IEEE international conference on tools with artificial intelligence*. Ieee, 1995, pp. 388–391.

[50] S. Shakeela, N. S. Shankar, P. M. Reddy, T. K. Tulasi, and M. M. Koneru, "Optimal ensemble learning based on distinctive feature selection by univariate anova-f statistics for ids," *International Journal of Electronics and Telecommunications*, pp. 267–275, 2021.

[51] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast cancer prediction: a comparative study using machine learning techniques," *SN Computer Science*, vol. 1, pp. 1–14, 2020.

[52] W. Wolberg, "Breast Cancer Wisconsin (Original)," UCI Machine Learning Repository, 1992, DOI: https://doi.org/10.24432/C5HP4Z.

[53] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[54] C. v. Rijsbergen, *Information retrieval*. Butterworth-Heinemann, 1979.

[55] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.

[56] G. U. Yule, "On the methods of measuring association between two attributes," *Journal of the Royal Statistical Society*, vol. 75, no. 6, pp. 579–652, 1912.

[57] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, pp. 1–13, 2020.

[58] M. McDermott, L. H. Hansen, H. Zhang, G. Angelotti, and J. Gallifant, "A closer look at auroc and auprc under class imbalance," *arXiv preprint arXiv:2401.06091*, 2024.

[59] N. R. Cook, "Use and misuse of the receiver operating characteristic curve in risk prediction," *Circulation*, vol. 115, no. 7, pp. 928–935, 2007.

[60] A. A. Alfayez, H. Kunz, and A. G. Lai, "Predicting the risk of cancer in adults using supervised machine learning: a scoping review," *BMJ open*, vol. 11, no. 9, p. e047755, 2021.

[61] R. L. Siegel, K. D. Miller, N. S. Wagle, A. Jemal *et al.*, "Cancer statistics, 2023," *Ca Cancer J Clin*, vol. 73, no. 1, pp. 17–48, 2023.

[62] *The World Factbook 2021*. Washington, DC: Central Intelligence Agency, 2021. Available: https://www.cia.gov/the-world-factbook/field/population/country-comparison.

[63] A. N. Richter and T. M. Khoshgoftaar, "Efficient learning from big data for cancer risk modeling: a case study with melanoma," *Computers in biology and medicine*, vol. 110, pp. 29–39, 2019.

[64] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[65] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[66] S. Kayikci and T. M. Khoshgoftaar, "Breast cancer prediction using gated attentive multimodal deep learning," *Journal of Big Data*, vol. 10, no. 1, p. 62, 2023.

[67] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.

[68] National Cancer Institute. The Cancer Genome Atlas (TCGA). https://portal.gdc.cancer.gov/projects/TCGA-BRCA. Accessed April 2024.

[69] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: analysis, applications, and prospects," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 12, pp. 6999–7019, 2021.

[70] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[71] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 3, pp. 841–850, 2018.

[72] S. Dixit, A. Kumar, and K. Srinivasan, "A current review of machine learning and deep learning models in oral cancer diagnosis: Recent technologies, open challenges, and future research directions," *Diagnostics*, vol. 13, no. 7, p. 1353, 2023.

[73] Y. Kumar, S. Gupta, R. Singla, and Y.-C. Hu, "A systematic review of artificial intelligence techniques in cancer prediction and diagnosis," *Archives of Computational Methods in Engineering*, vol. 29, no. 4, pp. 2043–2070, 2022.

[74] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosis-a survey," *Pattern Recognition*, vol. 83, pp. 134–149, 2018.

[75] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges," *Journal of Infection and Public Health*, vol. 13, no. 9, pp. 1274–1289, 2020.

[76] M. Dildar, S. Akram, M. Irfan, H. U. Khan, M. Ramzan, A. R. Mahmood, S. A. Alsaiari, A. H. M. Saeed, M. O. Alraddadi, and M. H. Mahnashi, "Skin cancer detection: a review using deep learning techniques," *International journal of environmental research and public health*, vol. 18, no. 10, p. 5479, 2021.

[77] S. L. Goldenberg, G. Nir, and S. E. Salcudean, "A new era: artificial intelligence and machine learning in prostate cancer," *Nature Reviews Urology*, vol. 16, no. 7, pp. 391–403, 2019.

[78] J. Calderaro, T. P. Seraphin, T. Luedde, and T. G. Simon, "Artificial intelligence for the prevention and clinical management of hepatocellular carcinoma," *Journal of hepatology*, vol. 76, no. 6, pp. 1348–1361, 2022.

[79] D. Painuli, S. Bhardwaj *et al.*, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," *Computers in Biology and Medicine*, vol. 146, p. 105580, 2022.

[80] G. F. Stark, G. R. Hart, B. J. Nartowt, and J. Deng, "Predicting breast cancer risk using personal health data and machine learning models," *Plos one*, vol. 14, no. 12, p. e0226765, 2019.

[81] M. C. Hornbrook, R. Goshen, E. Choman, M. O'Keeffe-Rosetti, Y. Kinar, E. G. Liles, and K. C. Rust, "Early colorectal cancer detected by machine learning model using gender, age, and complete blood count data," *Digestive diseases and sciences*, vol. 62, pp. 2719–2727, 2017.

[82] Y.-H. Wang, P. A. Nguyen, M. M. Islam, Y.-C. Li, H.-C. Yang *et al.*, "Development of deep learning algorithm for detection of colorectal cancer in ehr data." *MedInfo*, vol. 264, pp. 438–441, 2019.

[83] J. L. Schneider, E. Layefsky, N. Udaltsova, T. R. Levin, and D. A. Corley, "Validation of an algorithm to identify patients at risk for colorectal cancer based on laboratory test and demographic data in diverse, community-based population," *Clinical Gastroenterology and Hepatology*, vol. 18, no. 12, pp. 2734–2741, 2020.

[84] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Scientific reports*, vol. 6, no. 1, pp. 1–10, 2016.

[85] G. R. Hart, D. A. Roffman, R. Decker, and J. Deng, "A multi-parameterized artificial neural network for lung cancer risk prediction," *PLoS One*, vol. 13, no. 10, p. e0205264, 2018.

[86] D. Roffman, G. Hart, M. Girardi, C. J. Ko, and J. Deng, "Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network," *Scientific reports*, vol. 8, no. 1, p. 1701, 2018.

[87] H.-H. Wang, Y.-H. Wang, C.-W. Liang, and Y.-C. Li, "Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer," *JAMA dermatology*, vol. 155, no. 11, pp. 1277–1283, 2019.

[88] D. Zhao and C. Weng, "Combining pubmed knowledge and ehr data to develop a weighted bayesian network for pancreatic cancer prediction," *Journal of biomedical informatics*, vol. 44, no. 5, pp. 859–868, 2011.

[89] Barts health using AI to prioritise care for colon cancer patients, 2020. Available: https://www.bartshealth.nhs.uk/news/barts-health-using-ai-to-prioritise-care-for-colon-cancer-patients-8867.

[90] B. Burnett, S.-M. Zhou, S. Brophy, P. Davies, P. Ellis, J. Kennedy, A. Bandyopadhyay, M. Parker, and R. A. Lyons, "Machine learning in colorectal cancer risk prediction from routinely collected data: a review," *Diagnostics*, vol. 13, no. 2, p. 301, 2023.

[91] G. Paré, L. Raymond, A. O. de Guinea, P. Poba-Nzaou, M.-C. Trudel, J. Marsan, and T. Micheneau, "Electronic health record usage behaviors in primary care medical practices: a survey of family physicians in canada," *International Journal of Medical Informatics*, vol. 84, no. 10, pp. 857–867, 2015.

[92] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi, "Reproducibility in machine learning for health research: Still a ways to go," *Science Translational Medicine*, vol. 13, no. 586, p. eabb1655, 2021.