Reproducible Survival Prediction with SEER Cancer Data

Stefan Hegselmann STEFAN.HEGSELMANN@UNI-MUENSTER.DE
Institute of Medical Informatics
University of Münster, Germany

Leonard Greulich LEONARD.GREULICH@UNI-MUENSTER.DE
Institute of Medical Informatics
University of Münster, Germany

Julian Varghese JULIAN.VARGHESE@UNI-MUENSTER.DE
Institute of Medical Informatics
University of Münster, Germany

Martin Dugas DUGAS@UNI-MUENSTER.DE
Institute of Medical Informatics
University of Münster, Germany

Abstract

Survival prediction for cancer patients can increase the prognostic accuracy and might ultimately lead to better informed decision making. To this end, many studies apply machine learning to cancer data of the Surveillance, Epidemiology, and End Results (SEER) program. The first part of this report contains a literature review to obtain a systematic overview of these studies. We identify 34 publications and extract information about experimental setups and efforts to ensure reproducibility. The review shows that only one of the identified studies mentions reproducibility and no study contains straightforward reproducible results. This motivates the second part of this work. We demonstrate the feasibility of reproducible cohort selection and survival prediction with SEER cancer data. Experiments are performed for 1- and 5-year survival of breast and lung cancer with cases diagnosed between 2004 and 2009. We compare minimal data preprocessing with 1-n encoding of categorical inputs and apply logistic regression and multilayer perceptron (MLP) models. Encoding with 1-n vectors proves beneficial throughout all experiments. For lung cancer, MLP models show a slightly superior performance. Moreover, importance of input attributes is analyzed with logistic regression weights and ablation analysis for MLPs.

1. Introduction

Cancer is the second leading cause of death in the United States. Most common types are breast and lung cancer with 268,670 and 234,030 expected new cases in 2018 (Siegel et al. (2018)). Applying machine learning for survival prediction, i.e. predicting whether a patient will survive a given period of time after diagnosis, can increase the prognostic accuracy and might ultimately lead to better informed decision making. The Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute collects cancer incidence and survival information covering over 30% of the population in the U.S. (Howlader et al. (2017)). Due to its broad coverage and comprehensive data collection, SEER data serves as basis for many survival prediction experiments with machine learning.

Reproducibility is a key requirement to obtain comparable results allowing a critical evaluation of new approaches in machine learning. In many disciplines, such as computer vision it is ensured through publicly available datasets and open code. Fully reproducible biomedical research is scarce (Collins and Tabak (2014)). A recent study examining reproducibility of mortality prediction in critical care based on the publicly accessible and well documented MIMIC-III dataset reports several problems in reproducing study cohorts based on textual descriptions. The authors suggest public dissemination of source code along with public datasets for an iterative improvement in the field (Johnson et al. (2017)).

Reproducibility for survival prediction experiments with SEER data can be separated into *reproducible cohort selection* and *reproducible results*. During cohort selection an experiment specific subset of all original cases is extracted. For instance, to perform survival prediction cases with benign tumors and certain unknown or missing attributes are commonly excluded. However, in contrast to datasets in most machine learning disciplines this experiment specific selections cannot be published due to privacy restrictions. Hence, cohort selections must be reproducible so that experiments can be conducted on the same data. Reproducible results require same cohorts but also allow to verify the experimental outcomes. For instance, by publishing the source code of the experiments.

Clinical Relevance Machine learning might improve prognostic models. However, survival prediction is a sensitive domain, so technology-based approaches should be reproducible for transparency and to establish trust. The first part of this report contains a literature review to identify efforts to ensure reproducibility in past studies. In the second part, we demonstrate fully reproducible survival prediction experiments with SEER data.

Technical Significance We do not present a new method for survival prediction, instead we use a simple setup with techniques from recent studies and focus on reproducibility. First, we will introduce our cohort selection, data extraction, and feature choice. We consider 1- and 5-year survival for breast and lung cancer based on cases diagnosed between 2004 and 2009. We compare minimal data preprocessing with 1-n encoding. Second, survival prediction is performed with logistic regression and multilayer perceptron (MLP) models. Lastly, we analyze the input attribute importance with logistic regression weights and ablation analysis for MLPs. By providing SEER*Stat session files and source code along with instructions, we ensure both reproducible cohort selection and reproducible results.

2. Literature Review

There are many studies applying machine learning for survival prediction with SEER cancer data. To obtain a systematic overview over existing publications and to assess their reproducibility, we perform a literature review. We extract information of the experimental setups and try to identify efforts to ensure reproducible experiments.

2.1. Study Selection

Figure 3 in the Appendix shows a flow diagram of inclusion and exclusion criteria for the literature review. Search was performed with PubMed and Google Scholar. Different queries are used, since the PubMed query yields more than 3,500 results in Google Scholar exceeding the scope of this review. Using a more specific query for Google Scholar lead

to 719 results that we cleaned from patents and citations ($N = 47$). Furthermore, titles without association with SEER, machine learning, or an application to survival prediction were excluded ($N = 538$). In general, uncertain candidates were kept for later and more in-depth analysis. We merged the resulting 134 candidate publications with the eleven results from PubMed and removed duplicates ($N = 5$) resulting in a total of 140 publications for content review. During this review, we further excluded studies inaccessible through our university access, bachelor/master/dissertation theses, and work with duplicate content or different spelling of the title or authors ($N = 44$). We screened the remaining 96 studies for clear indications of applying machine learning for survival prediction with SEER data and excluded studies based on abstract ($N = 35$) and full text review ($N = 27$) leaving us with 34 studies. In the last step, we also removed studies with ambiguous descriptions of the target variable and experiments that combined the SEER dataset with other data.

2.2. Information Extraction

We extracted information about the general experimental setups and try to identify efforts made for reproducibility in the 34 identified studies. To ensure correct data, two individuals performed information extraction separately. In case they extracted different information, the publication was consulted together to reach an agreement.

Information of the experimental setup includes cancer type, time range of SEER data, survival period, applied machine learning models, number of cases and attributes used as input, and the best reported result (see columns of Table 1). A dash indicates missing, a question mark uncertain information. We assigned models into common categories to simplify the notation. Abbreviations of those categories are given in the Appendix. Models that do not fit into a category or custom modifications are marked with an asterisk. Bracketing and plus symbols indicate model combinations. A single result is reported for each publication and cancer type. For this, we prefer the best model and 5-year and 1-year survival periods, because they are most common. The model and survival period for a reported result are indicated by mentioning them first in their respective columns. For instance, the result reported for Dimitoglou et al. (2012) (ACC=0.944) is for 5-year survival and a decision tree model. Number of attributes refers to the number of original SEER attributes that were used. We report results as Area Under the Receiver Operating Characteristic Curve (AUC), F1 score (F1), and accuracy (ACC) in this order. Two experiments carry out regression for which we report root mean squared error (RMSE).

To identify efforts made for reproducibility in the selected studies, we used text comprehension gained during the extraction of experimental setups. In addition to that, we searched for the terms *code* and *program* to find hints for published code and the terms *repro* and *repeat* to find statements about reproducibility. We scanned figures, tables, footnotes, and appendices for hints of reproducible experiments. The last column in Table 1 indicates whether a study published the source code of their experiments.

2.3. Experiments

Table 1 summarizes the information extracted from the 34 identified studies. Six of these papers were published in 2017 illustrating the topic's relevance. Most of the studies (N = 23) are concerned with breast cancer, which is the most common type of cancer in the

Table 1: Experimental setups from 34 studies identified in the literature review. A single result is reported for each publication and cancer type. Survival period and model for reported results are indicated by mentioning them at the first position.

Publication	Cancer	Time range	Surv. years	Models	Cases	Attr.	Result	Code
Afshar et al. (2015)	Breast	1999-2004	5	SVM, BN, DT	22,763	18	ACC=0.967	
Al-Allak et al. (2010)	Breast	1990-1997	10	DT	50,895	3	ACC=0.693	
Bellaachia and Guven (2006)	Breast	1988-2002	5	DT, MLP, NB	151,886	16	ACC=0.867	
Burke et al. (1997)	Breast	1977-1982	10	MLP	6,787	3	AUC=0.730	
Delen et al. (2005)	Breast	1973-2000	5	DT, MLP, LogR	202,932	16	ACC=0.936	
Dooling et al. (2016)	Breast	1973-2012	5/0.5/1	RF, MLP	329,949	66	AUC=0.844	X
Edeki and Pandya (2012)	Breast	1990-1997	10	(RF, MLP, LogR, DT, SVM) + BOO, BAG	15,194	20 ?	ACC=0.75	
Endo et al. (2008)	Breast	1992-1997	5	LogR, MLP, NB, BN, DT	37,256	10	ACC=0.858	
Jahanbazi and Nadimi (2016)	Breast	1973-2012	5	SMOTE + InfoGain + DT*	12,000	13	ACC=0.871	
Kate and Nadig (2017)	Breast	2004-2008	5	LogR, DT, NB	174,518	16	AUC=0.850	
Khan et al. (2008)	Breast	1973-2003	5	DT	40,600	16	AUC=0.77	
Kibis et al. (2017)	Breast	1973-2013	10	LogR, MLP, BN, DT	52,825	28	AUC=0.808	
Kim and Shin (2013)	Breast	1973-2003	5	SSL Co-train*, SVM, SSL*, MLP	50,000	16	AUC=0.81	
Miri Rostami and Ahmadzadeh (2017)	Breast	2004-2007	5	((SMOTE, DSO*) + PSO* +CFS*) + DT, BN, LR	23,512	10	AUC=0.939	
Nam and Shin (2013)	Breast	-	5	SSL Co-train*, SVM, SSL*, DT, MLP	50,000	16	AUC=0.81	
Park et al. (2013)	Breast	1973-2003	5	SVM, SSL*, MLP	162,500	16	AUC=0.80	
Shin and Nam (2014)	Breast	1973-2003	5	SSL Co-train*, SVM, SSL*, DT, MLP	50,000	16	AUC=0.81	
Shukla et al. (2018)	Breast	1973-2012	5/3/7	Clustering + MLP	85,189	25	ACC=0.692	
Solti and Zhai (2013)	Breast	1973-2009	10	DT, LogR, NB	657,711	12	AUC=0.852	
Street (1998)	Breast	1977-1982	10	MLP	>38,000	5	-	
Wang et al. (2012)	Breast	1973-2002	5	LogR, DT	215,375	6	AUC=0.829	
Wang et al. (2013)	Breast	1988-2002	5	(sampling, attr. selection*) + LogR, DT	215,221	9	AUC=0.829	
Wang et al. (2014)	Breast	1973-2007	5	(SMOTE + PSO*) + DT, LogR, 1-NN	215,221	20	ACC=0.943	
Al-Bahrani et al. (2017)	Colon	1988-2005	5/1/2	MLP	147,644	14	AUC=0.87	
Dooling et al. (2016)	Colon	1973-2012	5/0.5/1	MLP, RF	113,072	102	AUC=0.841	X
Gao et al. (2012)	Colon	1998-2000	5	ANFIS, MLP*, SVM, LogR, BN, DT, NB	10,000	14	AUC=0.821	
Noohi et al. (2013)	Colon	1969-2010	(<1,1-5,>5)	MLP, BN, DT	5,276	8 ?	ACC=0.716	
Silva et al. (2016)	Colon	2004-2012 ?	5/1/2/3/4	(DT, k-NN, RF, NB) + EV, ST, BOO, BAG	38,592	18/6	AUC=0.994	
Stojadinovic et al. (2013)	Colon	2000-2006	5/1/2/3	BN	77,402	13 ?	AUC=0.85	
Agrawal et al. (2012)	Lung	1998-2001	5/0.5/0.75/1/2	(DT, BOO, RF, MLP, SVM) + EV	57,254	63	AUC=0.940	
Dimitoglou et al. (2012)	Lung	1988-2003	5/7/10	DT, NB	174,491	14	ACC=0.944	
Dooling et al. (2016)	Lung	1973-2012	5/0.5/1	MLP, RF	177,089	114	AUC=0.875	X
Fradkin et al. (2006)	Lung	1988-2001	0.66	SVM, LogR	217,558	15	sens./spec.	
Lynch et al. (2017a)	Lung	2004-2009	6 (regression)	(GBM, LinR, RF, DT, SVM) + EV	10,442	18	RMSE=15.30	
Lynch et al. (2017b)	Lung	2004-2009	6 (regression)	(SOM, HC, MBC, k-means, NNMF, PCA) + LinR	10,442	8	RMSE=15.59	
Delen (2009)	Prostate	1988-2001	5	SVM, MLP, DT, LogR	>120,000	77	ACC=0.929	

U.S. One paper analyzes more than one cancer type (Dooling et al. (2016)). Stated start and end years vary between 1969 and 2013. Later start dates are chosen to include only recent data or to make use of attributes introduced at a later stage. The end date is usually determined by the SEER data available at experiment time with an additional offset to include all necessary follow-ups. A survival period of five years is used in many experiments (N = 25), followed by one (N = 8) and ten years (N = 7). There are two experiments which perform regression to predict survival months (Lynch et al. (2017a), Lynch et al. (2017b)). Plenty of different models are applied, sometimes in combination with sampling techniques or methods for input size reduction. Common models are decision trees (DT, N = 24) and multilayer perceptrons (MLP, N = 20), followed by logistic regression (LogR, N = 12) and support vector machines (SVM, N = 10). Cohort selection and, hence, the resulting number of cases varies a lot between experiments. Some studies simply use all cases available, others exclude cases based on empty attributes or specific filter criteria. Data preprocessing includes normalization methods, data imputation for missing values, and semantic mappings of attributes. In some experiments, especially where case numbers are round, cases were excluded to account for a target label imbalance or due to performance reasons. Input attributes used for prediction also differ a lot. Many experiments justify their selection with clinical relevance or a separate data analysis. AUC is given as performance measure in 18 studies. Whenever AUC is not given, F1 is also missing, so we report ACC.

2.4. Reproducibility

We identified several efforts to ensure reproducibility. Nearly all studies provide information about their experimental setups and textual descriptions of cohort selection (see Table 1). Several studies provide an overview of input attributes used for prediction. Some papers include overviews of the selected model parameters, names of SEER input files, or pseudocode for critical parts of their programs such as target label generation. Experiments with decision trees often include textual description of generated decision rules.

Only a single study (Dooling et al. (2016)) explicitly mentions reproducibility and provides Github repositories[1] containing source code for experiments and a web application based on the prediction algorithm. However, both repositories provide no meaningful instructions. Our attempts to understand the provided code failed due to missing explanatory comments and several complex data mappings. Executing the program and reproducing results was impossible due to undocumented external dependencies.

We cannot conclude with absolute certainty that none of the identified studies contains reproducible cohorts or results. To verify this statement, a manual review of every experiment would be necessary, which laid outside the scope of our review. However, we observed that 33 out of 34 studies only come with textual descriptions of their experiments that already proved insufficient to reproduce cohorts (Johnson et al. (2017)) and contain no clear hints for reproducibility. Only a single paper mentions reproducibility and provides source code. However, this approach also proved insufficient due to missing instructions and bad code quality. Hence, we can conclude that none of the identified studies is *straightforward reproducible*. Straightforward in the sense that a publication contains clear and simple instructions to reproduce it. Moreover, the literature review revealed that no study conducted

1. https://github.com/doolingdavid/{PAPERDATA,colon-cancer-nn-errors}

by different researchers or institutions is based on the same input data. This supports our assumption that it is difficult to reproduce cohorts of existing experiments. This motivates the following work introducing fully reproducible cohort selection and reproducible results for survival prediction with SEER data.

3. Cohort

SEER collects cancer incidence data from population-based cancer registries covering over 30% of the population in the U.S. The data contains detailed information on patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment. In addition to that, SEER registries follow up with patients for vital status to provide survival information (Howlader et al. (2017)). SEER releases data submissions annually, containing new incidences and updated information for existing cases. Our study is based on the November 2016 submission with data from 1974 to 2014 (SEER (2017)).

3.1. Cohort Selection

Cohort selection for our experiments is performed with SEER*Stat, a dedicated software for the analysis of SEER data. It allows authorized users to investigate the collected cancer cases and to produce statistics based on them. In addition to that, it offers (1) the specification of a SEER database, (2) predefined exclusion criteria, (3) a custom query builder with a sophisticated selection mechanism allowing to combine several attribute conditions, and (4) an export functionality for selected cases. Most importantly, the complete SEER*Stat configuration can be stored into a session file and re-used by others to fully reproduce the selected cohort. It is possible to publish session files since they include no patient information but only the necessary selection parameters.

We perform predictions for 1- and 5-year survival of breast and lung cancer, which are the most common tasks in past experiments. Only cases diagnosed since 2004 are considered, because at that time many new attributes were included and to use only recent data. We chose 2009 as end date to ensure a 5-year follow-up period for all incidences. Selecting a cohort for survival prediction with SEER data is an ambiguous task. There is no consent about inclusion and exclusion criteria in existing studies. We use a SEER*Stat exercise session provided by the SEER program[2] as basis of our selection. The exercise description claims to "represent the standard selections most commonly used for a survival analysis", which was confirmed by the SEER staff. It includes cases with malignant tumors and known age and excludes entries with unknown or missing cause of death, deaths only confirmed by autopsy or death certificate, and cases that are marked as alive with no survival time. It also excludes cases based on an expected survival table for U.S. citizens. We adjust these predefined cohort selection only regarding cancer type and year of diagnosis. The resulting SEER*Stat queries are given in Figure 1. The cohorts consist of 275,167 cases for breast and 229,011 cases for lung cancer. SEER*Stat session files for our cohort selections are contained in this paper's repository[3]. We refer to these files for further details regarding the cohort. By loading them into SEER*Stat, the selection can be reproduced.

2. https://seer.cancer.gov/seerstat/tutorials/survival3
3. https://github.com/stefanhgm/MLHC2018-reproducible-survival-seer/tree/master/cohort

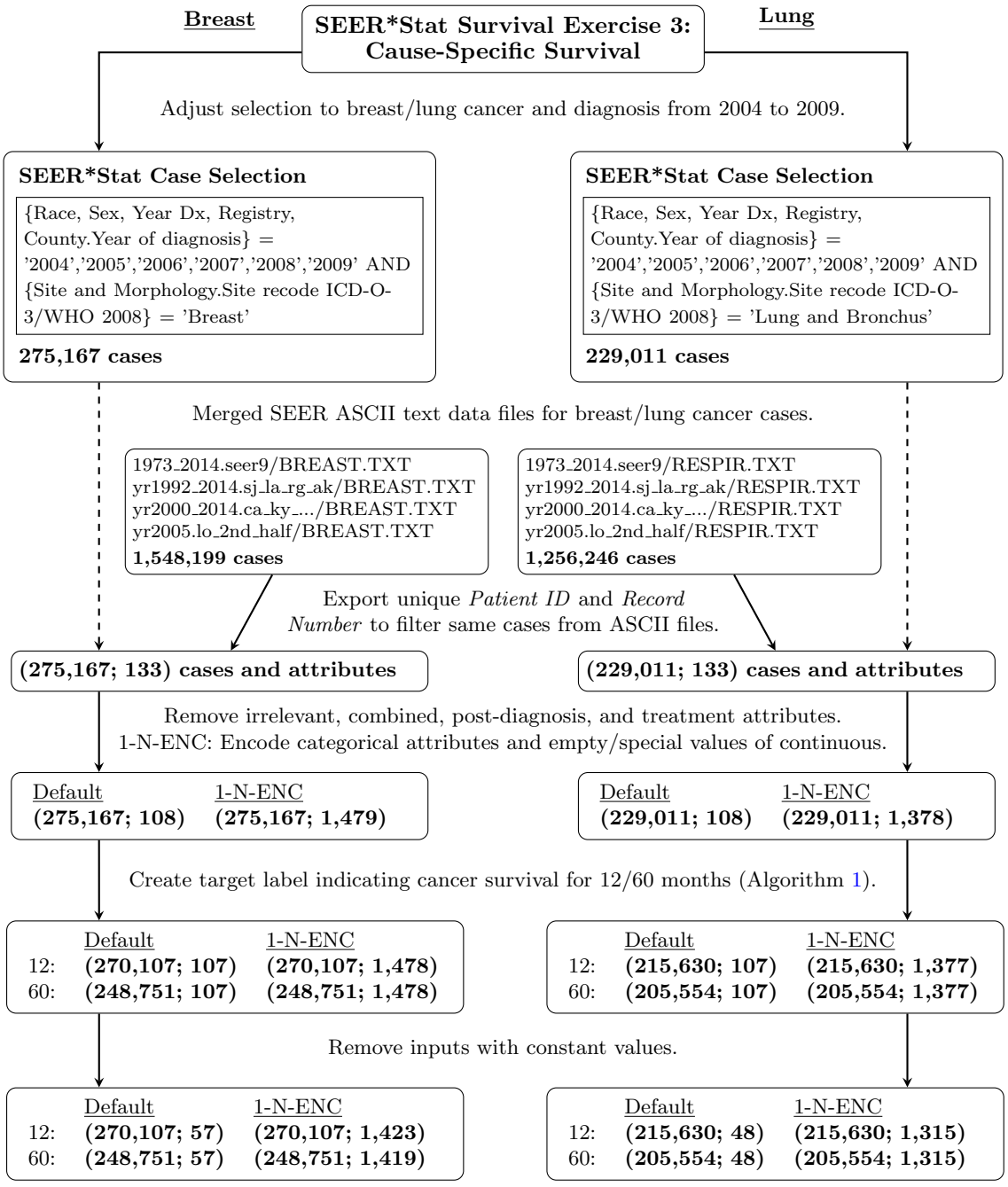

Figure 1: Flow chart for cohort selection, data extraction, and feature choice. The last boxes indicate the final number of cases and features for our experiments with breast and lung cancer, different input representations (Default and 1-N-ENC) and two survival periods (12 and 60 months).

3.2. Data Extraction

It is possible to export cases from SEER*Stat for further analysis. For our experiments, however, we use the SEER ASCII files since they contain numerical encodings of all attributes. The ASCII data for the 2016 submission (SEER (2017)) are available from SEER on request. Merging all incidences for breast and lung cancer results in 1,548,199 and 1,256,246 cases (see Figure 1). We export unique identifiers (patient ID and case number) from SEER*Stat and use them in our software to filter the ASCII data for our cohort. This yields the same cases in the ASCII format consisting of 133 attributes for each case.[4] Irrelevant fields such as patient ID and compound fields that combine several information are removed in our software. Moreover, in consultation with SEER staff we removed treatment information commonly excluded from survival analyses and attributes added after the initial diagnosis leaking future information. This decreased the number of attributes to 108.

3.3. Feature Choices and Target Label Generation

We perform experiments with two different input representations. The first representation treats all inputs as continuous that are normalized with respect to mean and variance. Hence, the number of resulting inputs remains 108 (see Default in Figure 1). However, many SEER attributes encode categorical (e.g. sex) or non-interval values (e.g. stage). Moreover, some attributes contain codes for special or unknown values. For instance, a tumor size in the interval 1 to 988 encodes size in millimeters whereas 999 indicates unknown and 991 less than 1 cm. In addition to that, a custom value of -1 was introduced to represent missing data. Treating this variable as continuous can prove insufficient. This motivates an alternative input representation where we divide attributes into categorical and continuous. Categorical attributes are encoded as 1-n vectors where each entry in the vector represents a value of the original attribute. For continuous attributes, values are mean and variance normalized as for the first representation except for values representing an unknown or special code. In this case, the value is set to zero and an additional 1-n vector for special codes is used instead. Figure 2 illustrates both input representations without mean and variance normalization. Three rows represent three exemplary cases with attributes size (continuous) and stage (categorical). The values 999 for size represent *unknown* and -1 for stage *missing*. Model a) uses the first representation with attributes as single inputs. Model b) and c) use the second input representation. Size is used as single input except for the special code 999 which is 1-n encoded. Stage, a categorical variable, is completely encoded as 1-n vector. The 1-n encoding for the second input representation increases the inputs to 1,479 and 1,378 (see 1-N ENC in Figure 1).

We create a target label indicating whether a person died within one or five years due to cancer or survived for this period of time. Algorithm 1 contains the according pseudocode. For each case, the target label *Survived cancer for n months* is set to zero if *Survival months* is less than 12 or 60 and *SEER cause of death classification* is equals to one indicating that a person died of their cancer.[5] If a person survived at least 12 or 60 months *Survived cancer for n months* is set to one. Remaining cases are removed. These are either right-censored or died within 12 or 60 months but due to another cause. For 60-months more cases are

4. https://seer.cancer.gov/manuals/read.seer.research.nov2016.sas
5. https://seer.cancer.gov/data-software/documentation/seerstat/nov2016/TextData.FileDescription.pdf

Algorithm 1 Create binary target label indicating survival of cancer for n months.

Input: Dataset *data* and number of survival months n
Output: Dataset *data* with new binary column *Survived cancer for n months*

```
Create new attribute column Survived cancer for n months
foreach case ∈ data do
    if Survival months < n and SEER cause of death classification = 1 then
        set column Survived cancer for n months to 0 (died within n months of cancer)
    else if Survival months ≥ n then
        set column Survived cancer for n months to 1 (survived n months)
    else
        remove case from data (died within n months of other cause or right-censored)
    end
end
return data
```

affected and removed. We exclude them from the experiments since our modeling approach cannot handle them. To generate the target label two attributes are merged into *Survived cancer for n months* decrementing the number of attributes by one (see Figure 1).

In a last step, constant attributes useless for discrimination are removed resulting in the final case and attribute counts given in Figure 1. Tables 4, 5, 6, and 7 in the Appendix summarize the resulting SEER attributes for 1- and 5-year breast and lung cancer survival predictions. They contain 57 and 48 inputs resulting from data preprocessing without 1-n encoding. Attribute names correspond to the ASCII documentation file. A minimum value of -1 indicates empty values, which are only present for three attributes. The column empty shows the exact number of empty values that are of significant size for *Insurance Recode (2007+)* and *Recode ICD-O-2 to 10*. These two attributes are either collected since 2007 or contain only empty and unknown values. A maximum value consisting only of nines, usually a special value for unknown, appears for many attributes. Attributes with the highest number of different values are *State-county recode*, FIPS code for state and county, *CS Tumor size*, tumor size in millimeters, and *Histologic Type ICD-0-3*, histologic coding of tumors. The last row contains information about the binary target label.

4. Methods

We perform survival prediction with logistic regression and MLP models. They are the second and third most common models used in past studies (see Section 2.3). Logistic regression serves as a simple model to provide a baseline for prediction results. MLPs are applied frequently in machine learning since the rising popularity of artificial neural networks and have been successfully applied to complex tasks. In particular, we want to investigate whether MLPs prove beneficial for many input attributes and whether they can deal with sparse inputs resulting from 1-n encoding. MLP models for the first and second input representations are illustrated as models a) and b) in Figure 2. In addition to that, we use a slightly modified MLP with the second input representation that performs an embedding of 1-n encoded attributes in its first hidden layer (MLPEmb). This approach is inspired from word representations in natural language processing (Mikolov et al. (2013)).

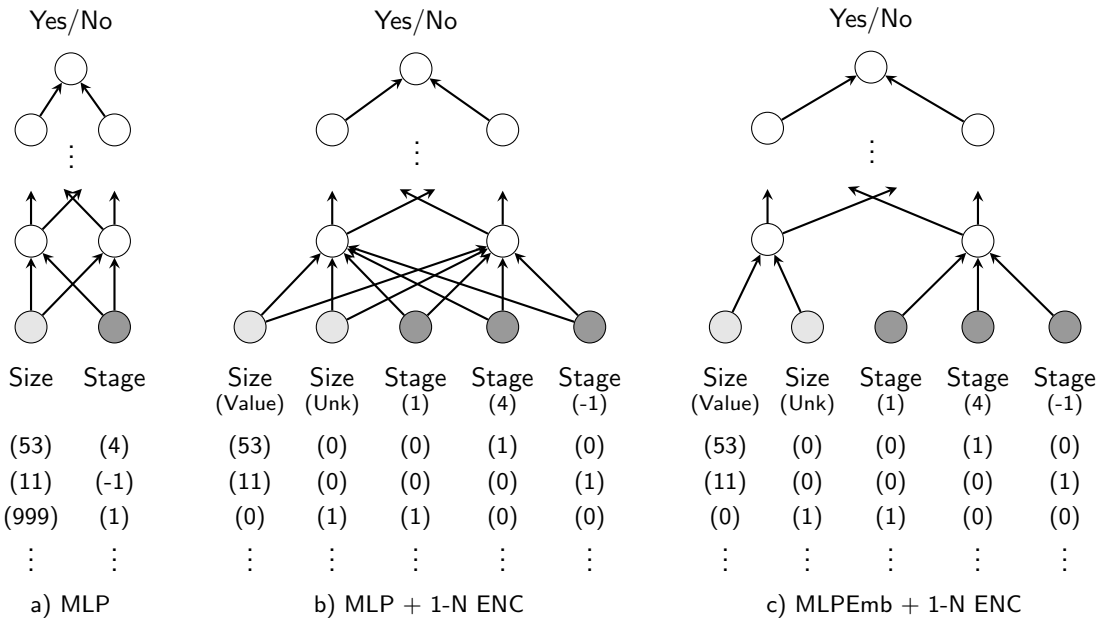

Figure 2: Illustration of MLP models used for experiments. Inputs for b) and c) are 1-n encoded (1-N ENC) and c) contains an embedding in the first layer (MLPEmb).

The underlying intention is to learn a representation of each input attribute in the first layer before combining them into a fully connected layer (see model c) in Figure 2).

The data resulting from cohort selection and data preprocessing is randomly split up into training, validation, and testing sets with the ratios 80%, 10%, and 10%. Evaluation is performed with Area Under the Receiver Operating Characteristic Curve (AUC) and F1 score (F1). We choose the best model based on the maximum sum of AUC and F1. We report accuracy (ACC) only for comparison with past studies. Hyperparameters are tuned on the validation set. A single run for each model and optimal parameter configuration is performed on the test set, which is the final score we are reporting. For logistic regression, regularization parameter C is tuned with $1 \cdot 10^x$ for x in $-2, -1, ..., 10$. For MLP models, the number of hidden layers $(1, 2, 3, 4)$, nodes per hidden layer $(20, 50, 100, 200)$, dropout $(0.0, 0.1, ..., 0.5)$, and training epochs $(20, 50, 100)$ are tuned. MLPEmb has an additional hyperparameter, the number of nodes used for attribute embedding in the first hidden layer (a single one for model c) in Figure 2). This value is tuned with $(3, 5, 10)$. As a result, there are 13, 288, and 864 parameter configurations for the logistic regression, MLP, and MLPEmb models. Tuning is performed on the HPC cluster of the University of Münster.

In addition to the survival prediction performance, we analyze the input importance of SEER attributes for the best performing logistic regression, MLP, and MLPEmb models. For logistic regression absolute values of the inputs weights are used. In case of 1-n-encoding, absolute weights for all inputs belonging to a specific attribute are summed. For MLP and MLPEmb ablation analysis is performed. For this, all inputs of a SEER attribute are set to zero and the absolute difference of the output serves as importance measurement.

Table 2: Survival prediction performance for breast cancer using a dummy classifier (BASE RATE), logistic regression (LOG REG), and MLP models (see Figure 2).

	1-year survival			5-year survival		
Model	AUC	F1	ACC	AUC	F1	ACC
BASE RATE	0.5	0.9849	0.9703	0.5	0.9318	0.8724
LOG REG	0.9381	0.9869	0.9744	0.8808	0.9481	0.9066
LOG REG + 1-N ENC	**0.9440**	**0.9873**	0.9752	0.9023	**0.9518**	0.9136
MLP	0.9384	0.9871	0.9748	0.9002	0.9516	0.9131
MLP + 1-N ENC	0.9393	0.9872	0.9749	0.9039	0.9517	0.9130
MLPEmb + 1-N ENC	0.9397	0.9872	0.9749	**0.9062**	0.9517	0.9134

Table 3: Survival prediction performance for lung cancer using a dummy classifier (BASE RATE), logistic regression (LOG REG), and MLP models (see Figure 2).

	1-year survival			5-year survival		
Model	AUC	F1	ACC	AUC	F1	ACC
BASE RATE	0.5	0.0	0.5609	0.5	0.0	0.8404
LOG REG	0.8209	0.6942	0.7552	0.8932	0.6180	0.8933
LOG REG + 1-N ENC	0.8419	0.7162	0.7696	**0.9100**	0.6394	0.9001
MLP	0.8347	0.7207	0.7639	0.9026	0.6381	0.8948
MLP + 1-N ENC	0.8446	0.7265	0.7681	0.9076	0.6492	0.8996
MLPEmb + 1-N ENC	**0.8476**	**0.7351**	0.7736	0.9078	**0.6539**	0.8961

5. Results

SEER*Stat session files and the code to reproduce our experiments along with an example are contained in a public Github repository[6]. Experimental results for 1- and 5-year survival prediction are contained in Table 2 for breast and Table 3 for lung cancer. Base rates of a dummy classifier predicting the most frequent label are given for comparison.

For breast cancer, the prior probability for surviving is already high as indicated by the base rates for F1 and ACC. Applying logistic regression or MLPs can only slightly improve these values in case of 1-year survival. Improvements are larger for 5-year survival. With regard to the optimization criteria AUC and F1, logistic regression with 1-n encoding almost always gives the best performance. Only AUC for 5-year survival is higher with the MLPEmb model. However, it is worth mentioning that prediction results for logistic regression with 1-n encoding and MLP models only show marginal differences making it difficult to determine a clearly superior approach. Input representation with 1-n encoding always improves AUC and F1 values. Best performing MLP models for breast cancer experiments contain only one or two hidden layers with 20 to 100 nodes per layer.

6. https://github.com/stefanhgm/MLHC2018-reproducible-survival-seer

In case of lung cancer, base rates are lower and model differences are larger. The base rate for F1 is zero since the negative label prevails. The MLPEmb model yields the best performance regarding AUC and F1. Only exception is the AUC value for 5-year survival, which is higher for logistic regression with 1-n encoding. The MLP with 1-n encoding outperforms logistic regression on the same scores as the MLPEmb model, only by a smaller margin. This suggests that embedding 1-n encoded attributes in MLPEmb is advantageous. Top F1 scores are much lower compared to lung cancer. Additional experiments on a local machine show a low recall of 0.6936 and 0.6495 for 1- and 5-year survival with MLPEmb. Just as for breast cancer, using 1-n encoding always gives an improved performance. However, best MLP models for lung cancer experiments are larger with two to four hidden layers and up to 200 nodes per layer. This suggests that lung cancer data contains more complex relationships useful for survival prediction.

Figures 4, 5, 6, and 7 in the Appendix illustrate the relative attribute importance for 1- and 5-year breast and lung cancer survival prediction. The diagrams are limited to the top ten attributes according to their summed importance across all models. For both cancer types *State-county recode* receives a very high importance for logistic regression with 1-n encoding (note the y-axis break). One possible explanation might be that *State-county recode* contains the highest number of different values leading to many logistic regression weights responsible for the 1-n encoded attribute. In addition to that, AJCC codes, staging, and breast cancer specific attributes are most important for breast cancer survival prediction. For lung cancer, the most important attributes include information about the tumor extension, histology, staging, and metastases. Moreover, attribute importance for 1- and 5-year predictions are very similar sharing eight and nine attributes out of the top ten for breast and lung cancer.

6. Discussion

A literature review for survival prediction with SEER cancer data was performed and we identified 34 relevant studies. Six of them were published in 2017. These publications aim to increase prognostic accuracy with methods from machine learning. Application of machine learning is justified with the success in other disciplines and the objective to incorporate more information and more complex relationships than common statistical approaches. However, only one out of 34 studies mentions reproducibility and publishes source code. Due to missing explanations and bad code quality, this approach also proved insufficient. Hence, we conclude that none of the identified studies is straightforward reproducible.

Moreover, there are no studies from different researchers based on the same cohort. Efforts to compare experimental outcomes exist. For instance, Al-Bahrani et al. (2017) refers to Stojadinovic et al. (2013) and Wang et al. (2012) try to compare results with the much-cited work of Delen et al. (2005). However, since predictions are not based on the same input data, comparability of those studies is limited. In our experiments, we have experienced that slight changes of the study cohort can lead to significant changes of the results. The current situation prevents transparent benchmarking and reliable statements about machine learning approaches for survival prediction with SEER data. Instead of performing experiments with different cohorts, data preprocessing methods, and machine learning models, future research should instead put a greater emphasize on reproducibility.

In addition to that, we reckon that existing studies present their results too positively. Only few publications report target prior probabilities or base rates of prediction scores. This results in misleading impressions of prediction performances. For instance, our results for 1-year breast cancer survival seem very accurate, but the improvements over base rates are only marginal. Results in the reviewed experiments are often described as good or usable and some studies demand a clinical application. From our point of view, these assessments are wrong in many cases. Overly optimistic reports might raise unrealistic expectations leading to a loss of confidence into the field of machine learning in health care.

Our experiments show that 1- and 5-year survival prediction for breast cancer with logistic regression and MLP models only yield small improvements over base rates. The situation is different for lung cancer. MLP models slightly outperform logistic regression and base rates. However, top F1 scores are still relatively low. We think these results do not suggest the application in a clinical use-case and several obstacles concerning integration, usability, and interpretability must be overcome to realize a clinical decision support system based on SEER data. We consider our experiments as proof-of-concept that (1) demonstrates the feasibility of reproducible survival prediction by providing SEER*Stat session files and open code and (2) gives a baseline for survival prediction performances of modern machine learning methods.

This work has limitations. We tried to identify reproducibility efforts in the reviewed studies solely based on hints in the text, figures, tables, footnotes, and appendices. Some publication might not contain such indications but still allow reproduction of cohorts and results. To identify these studies, a manual review of every experiment would be necessary. Second, our cohort selection is based on a common survival analysis selection provided by SEER, which we consider suitable for survival prediction. However, there might be reasons to choose another cohort, for instance to remove cases with missing attributes. Moreover, we excluded cases from the original cohort that died within one or five years due to other reasons than cancer, since no target label could be assigned to them. This could bias the experimental outcomes. Third, we excluded irrelevant and combined attributes and in consultation with SEER staff also attributes that could leak prognosis information. Relevant attributes might have been excluded or leaking attributes overseen in this attribute selection process. Fourth, we performed survival prediction only with logistic regression and MLP models and two simple input representations. Our methods did not account for the class imbalance occurring in the experimental data. We are aware that this imbalance results in high performance metrics. However, by providing base rates of a dummy classifier and prior probabilities of our target label, we aim to set these values into perspective. This paper does not focus on the methods for survival prediction but the reproducibility of experiments with approaches of recent studies. Moreover, more extensive parameter tuning was performed for MLPs, which might be the reason for their superior performance. Lastly, attribute importance based on logistic regression weights and ablation analysis can only give hints of attribute importance and should not be confused with full model interpretabilty.

Our work opens up possibilities for future work. First of all, we hope that future research builds upon our results and some of the existing methods for survival prediction will be evaluated on our cohort. Even if another cohort is used, reproducible analysis of SEER data is feasible and therefore should be standard practice. Future work can consider incorporation of censored data that we removed for our experiments. Dooling et al. (2016)

tackle this problem by predicting a survival function that uses survival months as additional input variable; Alaa and van der Schaar (2017) by using a competing risk model with a designated event for censoring. Moreover, SEER data might be used in combination with other data sources such as the SEER-Medicare Linked Database. Another possibility would be a combination with clinical data elements that proved superior to SEER data alone for predicting short-term mortality (Elfiky et al. (2017)). Future work could investigate whether SEER data can actually lead to a clinical outcome. To this end, clinical decision support systems utilizing SEER data should be compared with established methods for assessing survival of cancer patients.

7. Conclusion

We identified 34 studies that apply machine learning for survival prediction with SEER cancer data in a literature review. We extracted information of their experimental setups and scanned them for efforts to ensure reproducibility. This review showed that past experiments were performed with many different setups but contain no straightforward reproducible cohorts and results. Moreover, there are no studies from different institutions that are based on the exact same input data preventing transparent benchmarking. We show that reproducible analysis with SEER data is feasible and present fully reproducible survival prediction experiments for breast and lung cancer with logistic regression and MLPs. We encourage future studies to build upon our results and follow an open data approach to foster reproducible research.

Acknowledgments

We thank the SEER staff for providing great support for questions related to the SEER dataset and SEER*Stat software.

References

Hadi Lotfnezhad Afshar, Maryam Ahmadi, Masoud Roudbari, and Farahnaz Sadoughi. Prediction of breast cancer survival through knowledge discovery in databases. *Global journal of health science*, 7(4):392, 2015.

Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary. Lung cancer survival prediction using ensemble data mining on seer data. *Scientific Programming*, 20(1):29–42, 2012.

A Al-Allak, R Leonard, and PD Lewis. Abstract p4-09-19: The naïve-bayes decision tree (nbtree) classifierpredicting the probability of survival in breast cancer, 2010.

Reda Al-Bahrani, Ankit Agrawal, and Alok Choudhary. Survivability prediction of colon cancer patients using neural networks. *Health informatics journal*, page 1460458217720395, 2017.

Ahmed M Alaa and Mihaela van der Schaar. Deep multi-task gaussian processes for survival analysis with competing risks. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

Abdelghani Bellaachia and Erhan Guven. Predicting breast cancer survivability using data mining techniques. *Age*, 58(13):10–110, 2006.

Harry B Burke, Philip H Goodman, David B Rosen, Donald E Henson, John N Weinstein, Frank E Harrell, Jeffrey R Marks, David P Winchester, and David G Bostwick. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79(4):857–862, 1997.

Francis S Collins and Lawrence A Tabak. Nih plans to enhance reproducibility. *Nature*, 505(7485):612, 2014.

Dursun Delen. Analysis of cancer data: a data mining approach. *Expert Systems*, 26(1): 100–112, 2009.

Dursun Delen, Glenn Walker, and Amit Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2): 113–127, 2005.

George Dimitoglou, James A Adams, and Carol M Jim. Comparison of the c4. 5 and a naïve bayes classifier for the prediction of lung cancer survivability. *arXiv preprint arXiv:1206.1121*, 2012.

David Dooling, Angela Kim, Barbara McAneny, and Jennifer Webster. Personalized prognostic models for oncology: A machine learning approach. *arXiv preprint arXiv:1606.07369*, 2016.

Charles Edeki and Shardul Pandya. Comparative study of data mining and statistical learning techniques for prediction of cancer survivability. *Mediterranean journal of Social Sciences*, 3(14), 2012.

Aymen Elfiky, Maximilian Pany, Ravi Parikh, and Ziad Obermeyer. A machine learning approach to predicting short-term mortality risk in patients starting chemotherapy. *bioRxiv*, page 204081, 2017.

Arihito Endo, Takeo Shibata, and Hiroshi Tanaka. Comparison of seven algorithms to predict breast cancer survival. *International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association*, 13(2): 11–16, 2008.

Dmitriy Fradkin, Dona Schneider, and Ilya Muchnik. Machine learning methods in the analysis of lung cancer survival data. *DIMACS Technical Report 2005–35*, 2006.

Peng Gao, Xin Zhou, Zhen-ning Wang, Yong-xi Song, Lin-lin Tong, Ying-ying Xu, Zhen-yu Yue, and Hui-mian Xu. Which is a more accurate predictor in colorectal survival analysis? nine data mining algorithms vs. the tnm staging system. *PLoS One*, 7(7):e42015, 2012.

N Howlader, AM Noone, M Krapcho, D Miller, K Bishop, CL Kosary, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, HS Chen, EJ Feuer, and KA Chronin. Seer cancer statistics review, 1975-2014. 2017. *Bethesda, MD: National Cancer Institute*, 2017.

Turan Jahanbazi and Mohammad H Nadimi. Seer cancer statistics review. *International Journal of Computer Applications*, 155(8), 2016.

Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376, 2017.

Rohit J Kate and Ramya Nadig. Stage-specific predictive models for breast cancer survivability. *International journal of medical informatics*, 97:304–311, 2017.

Umer Khan, Hyunjung Shin, Jong Pill Choi, and Minkoo Kim. wfdt: weighted fuzzy decision trees for prognosis of breast cancer survivability. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pages 141–152. Australian Computer Society, Inc., 2008.

Eyyub Y Kibis, . Esra Byktahtakn, and Ali Dag. Data analytics approaches for breast cancer survivability: comparison of data mining methods. In *Proceedings of the 2017 Industrial and Systems Engineering Conference*, pages 591–596. Institute of Industrial and Systems Engineers (IISE), 2017.

Juhyeon Kim and Hyunjung Shin. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association*, 20(4):613–618, 2013.

Chip M Lynch, Behnaz Abdollahi, Joshua D Fuqua, R Alexandra, James A Bartholomai, Rayeanne N Balgemann, Victor H van Berkel, and Hermann B Frieboes. Prediction of lung cancer patient survival via supervised machine learning classification techniques. *International journal of medical informatics*, 108:1–8, 2017a.

Chip M Lynch, Victor H van Berkel, and Hermann B Frieboes. Application of unsupervised analysis techniques to lung cancer patient data. *PloS one*, 12(9):e0184370, 2017b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

S Miri Rostami and M Ahmadzadeh. Extracting predictor variables to construct breast cancer survivability model with class imbalance problem. *Journal of AI and Data Mining*, 2017.

Yonghyun Nam and Hyunjung Shin. A hybrid cancer prognosis system based on semi-supervised learning and decision trees. In *International Conference on Neural Information Processing*, pages 640–648. Springer, 2013.

Narges Alizadeh Noohi, Marzieh Ahmadzadeh, and M Fardaer. Medical data mining and predictive model for colon cancer survivability. *International Journal of Innovative Research in Engineering & Science*, 2, 2013.

Kanghee Park, Amna Ali, Dokyoon Kim, Yeolwoo An, Minkoo Kim, and Hyunjung Shin. Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, 26(9):2194–2205, 2013.

SEER. Surveillance, Epidemiology, and End Results Program (www.seer.cancer.gov). SEER*Stat Database: Incidence - SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2016 Sub (2000-2014) <Katrina/Rita Population Adjustment>, 1969-2015 Counties, based on the November 2016 submission. Technical report, National Cancer Institute, DCCPS, Surveillance Research Program, April 2017.

Hyunjung Shin and Yonghyun Nam. A coupling approach of a predictor and a descriptor for breast cancer prognosis. *BMC medical genomics*, 7(1):S4, 2014.

Nagesh Shukla, Markus Hagenbuchner, Khin Than Win, and Jack Yang. Breast cancer data analysis for survivability studies and prediction. *Computer Methods and Programs in Biomedicine*, 155:199–208, 2018.

Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.

Ana Silva, Tiago Oliveira, José Neves, and Paulo Novais. Treating colon cancer survivability prediction as a classification problem. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 5(1):37–50, 2016.

David Solti and Haijun Zhai. Predicting breast cancer patient survival using machine learning. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, page 704. ACM, 2013.

Alexander Stojadinovic, Anton Bilchik, David Smith, John S Eberhardt, Elizabeth Ben Ward, Aviram Nissan, Eric K Johnson, Mladjan Protic, George E Peoples, Itzhak Avital, et al. Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model. *Annals of surgical oncology*, 20(1):161–174, 2013.

W Nick Street. A neural network model for prognostic prediction. In *ICML*, pages 540–546, 1998.

Kung-Jeng Wang, Bunjira Makond, and Kung-Min Wang. An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. *BMC medical informatics and decision making*, 13(1): 124, 2013.

Kung-Jeng Wang, Bunjira Makond, Kun-Huang Chen, and Kung-Min Wang. A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing*, 20:15–24, 2014.

Kung-Min Wang, Bunjira Makond, Wei-Li Wu, KJ Wang, and YS Lin. Optimal data mining method for predicting breast cancer survivability. *International Journal of Innovative Management, Information & Production*, 3(2):28–33, 2012.