Breast cancer data analysis for survivability studies and prediction

Nagesh Shukla[a], Markus Hagenbuchner[b], Khin Than Win[b], Jack Yang[c]

[a]School of Systems, Management and Leadership, Faculty of Engineering and Information Technology, University of Technology Sydney, NSW 2007, Australia
[b]School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2500, Australia
[c]SMART Infrastructure Facility, Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, NSW 2500, Australia

Abstract

Background: Breast cancer is the most common cancer affecting females worldwide. Breast cancer survivability prediction is challenging and a complex research task. Existing approaches engage statistical methods or supervised machine learning to assess/predict the survival prospects of patients.

Objective: The main objectives of this paper is to develop a robust data analytical model which can assist in (i) a better understanding of breast cancer survivability in presence of missing data, (ii) providing better insights into factors associated with patient survivability, and (iii) establishing cohorts of patients that share similar properties.

Methods: Unsupervised data mining methods viz. the self-organising map (SOM) and density-based spatial clustering of applications with noise (DBSCAN) is used to create patient cohort clusters. These clusters, with associated patterns, were used to train multilayer perceptron (MLP) model for improved patient survivability analysis. A large dataset available from SEER program is used in this study to identify patterns associated with the survivability of breast cancer patients. Information gain was computed for the purpose of variable selection. All of these methods are data-driven and require little (if any) input from users or experts.

Results: SOM consolidated patients into cohorts of patients with similar properties. From this, DBSCAN identified and extracted nine cohorts (clusters). It is found that patients in each of the nine clusters have different survivability time. The separation of patients into clusters improved the overall survival prediction accuracy based on MLP and revealed intricate conditions that affect the accuracy of a prediction.

Conclusions: A new, entirely data driven approach based on unsupervised learning methods improves understanding and helps identify patterns associated with the survivability of patient. The results of the analysis can be used to segment the historical patient data into clusters or subsets, which share common variable values and survivability. The survivability prediction accuracy of a MLP is improved by using identified patient cohorts as opposed to using raw historical data. Analysis of variable values in each cohort provide better insights into survivability of a particular subgroup of breast cancer patients.

Keywords: breast cancer survivability study; SEER data; machine learning

1. Introduction

Prediction of cancer survivability and treatment have been of interest to everyone worldwide. World Health Organisation indicated that cancer is the second leading cause of death worldwide. There are different treatment modalities for cancer such as surgery; chemotherapy, hormonal adjuvant therapy and radiation therapy [1].
It was found that early diagnosis of cancer, combined with early treatment will improve prognosis for cancer [2]. Prognosis also depends on the spread of cancer, in which cancer can spread to lymph node drainage areas (Node), and at different sites (Metastasis). A cancer staging system called TNM (Tumor, Node, Metastasis) classification is widely used for identifying the status of the cancer. Based on the types and stage of the cancer, treatment modalities outcomes could vary. For example, it was found that breast cancer, limited to first stage only, 96% of patients will be alive in five years after diagnosis [3]. Spread to regional nodes or other nodes and distant metastasis will reduce the survivability. The availability of treatments and the high rate of incidents made survivability become a subject of much interest to health professionals and researchers.

Data-driven predictive models for survivability of cancer can assist in prognosis and management of cancer. This also addresses initiatives such as Learning Healthcare Systems as identified by Institute of Medicine (United States). The proposed model is novel and provides a decision aid for understanding survivability of cancer patients and evidence-based cancer management. Evidence based medicine and evidence based healthcare has been a focus of modern clinical medicine. This study contributes to cancer management as it engages knowledge discovery technologies to cancer patient records. The study will identify breast cancer as an exempler and will use the SEER breast cancer dataset. While the scope of this paper is limited to cases of breast cancer the proposed methodologies are suitable for any other cancer management applications.

Demographics in Breast cancer

Previous studies on breast cancer indicated that survivability notably varies with the variation in different factors. For instance, survival difference based on the age has been noted due to not having a recommended guideline therapy [4, 5]. Ethnic background and the racial difference play a role in the survivability and noted that survival rate is lower in African American than the white women [6]. Madubata et al. identified in their study in Missouri that there is a racial difference in Ductal Carcinoma in situ (DCIS) [7]. Their study also identified that there is a higher risk of ipsilateral breast cancer in black women [7]. In addition, study on DCIS from National Cancer Database reported that there are differences in treatment chosen based on ethnicity and geographical region [8].

Kheirelseid et al. found that there were no differences in survival rate among those with unilateral or bilateral cases. The synchronus bilateral breast cancer has the poorer survival rate than the metachronus bilateral cases [9]. Synchronus breast cancer is seen more often in the elder women than the metachronus breast cancer [10]. The 15 years of follow-up study of 1187 with stage T1-2 N0 breast cancer patients in Sweden identified that ipsilateral breast cancer was reduced in patients having radiation therapy after breast conserving surgery [11].

2. Previous research on breast cancer data mining

Previous studies categorized survivability by using a threshold of 5 years [12, 13]. It is said that a breast cancer patient has survived if alive for five years since the first diagnoses. Previous prediction studies for breast cancer survivability treatment studies are based on 5-year survivability and they lack detailed explanation of survival years. Those survived less than 5 years are considered not survived and those more than 5 years are considered as survived [12-14] The proposed approach allows us to have a flexible method of having varying levels of survival years. Thus, allowing medical professionals to identify high risk subgroup of cancer patients even within 5-years survivability. Although Boughorbel et al. conducted survival prediction for either of 2, 5, 8 and 11 years their target value for prediction is considered as the binary value and 4 separate analysis were conducted [15].
Missing values are dealt differently in different studies. Rathore et al. replaced missing value with mean value in their data preprocessing [16]. In a study from Boughorbel, missing values of related subjects were removed [15]. Lotfnezhad Afshar used the multiple imputation method for missing value. For each complete dataset, average of each is used as a single data [17]. However, in their study, Delen et al. dealt with missing data in dealt by combining four different variables for site specific surgery and mapped them into one variable. However, missing data was still present for 1000 records so they were removed from the study [12].

Umesh and Ramachandra 2016, analysed the SEER breast cancer dataset [18]. They have included an external attribute "Menopause" which is not recorded in the SEER dataset by assuming that a female of an age above 50 years implies menopause.

Civcik et al.[19] proposed the use of multistable cellular neural networks in microcalcification detection in the early diagnosis of breast cancer. They used their method on the MIAS (Mammographic Image Analysis Society) database to provide accuracy, sensitivity and specificity values. Several research studies have been proposed in the literature related to analysis of image based medical datasets for detection and diagnosis of breast cancer. Jalalian et al [20]reviewed extant literature on computer-aided detection and diagnosis for breast cancer using ultrasound and mammography datasets. Sokouti et al [21] provided a framework for MLP-driven cervical cancer diagnosis based on cell image datasets. Similarly, Anousouya Devi et al [22] provided a review of different types of artificial neural networks used for cancer diagnosis and prediction.

Although the SEER dataset was used for most of the studies the variables chosen on different studies varied. Table 1 presents the comparison of variables used in the different studies. Studies included patient demographic (such as age, race, site, marital status), Site and Morphology (such as primary site, laterality, behaviour code, histology), Stage (such as Grade, Tumor size, lymph node, extension, TNM stage), Treatment modality such as (radiation, Surgery). Several different variables were used in the studies regarding staging. For comparison, those were mapped accordingly in Table 1. Attribute selection with "Lymph Node Status" being the only attribute that is commonly used in all of these studies.

Table 1: Variables commonly used in breast cancer dataset

	Boughorbel et al	Rathore et al.	Umesh & Ramachandra	Loftnez et al.	Solthi & Zhai	Delen	BellaAchia	Park et al.
AGE at DIAGNOSIS	✓	✓	✓	✓	✓		✓	✓
RACE		✓	✓	✓	✓	✓	✓	✓
SEX					✓			
MARITAL STATUS		✓	✓	✓		✓	✓	✓
PRIMARY SITE		✓	✓	✓		✓	✓	✓
LATERALITY		✓						
BEHAVIOR CODE			✓	✓		✓	✓	✓
GRADE	✓		✓	✓	✓	✓	✓	✓
HISTOLOGY	✓		✓	✓		✓		
EXTENSION OF TUMOR			✓	✓	✓	✓	✓	
LYMPH NODE STATUS	✓	✓	✓	✓	✓	✓	✓	✓
RADIATION		✓	✓	✓	✓	✓		
SURGERY		✓	✓	✓		✓	✓	
SURVIVAL TIME		✓						
TUMOR SIZE	✓			✓	✓		✓	
NO: OF PRIMARIES			✓	✓			✓	✓
STAGE					✓	✓	✓	
ER STATUS	✓			✓	✓			
PR STATUS	✓			✓	✓			

Different studies indicated factors related to breast cancer prognosis depending on different factors. There are several studies regarding breast cancer data analysis. More recent studies focused on predicting breast cancer through SVM [23], and on survival since the time of first diagnosis [12] [24].

It could be seen that breast cancer data analysis is a challenging task. Although several studies had been conducted to analyse breast cancer survivability through data mining methods, the majority of studies considered target variable as those who lived more than 5 years as survived and less than that is considered not survived. Prediction accuracies are generally lacking although some factors are known to greatly affect the prediction accuracy. Those factors are selected by hand and it is not known whether or not other factors (or combination of factors) also affect the prediction accuracy. Study on characteristics of patients group with similar features have been missing or limited.

3. Methodology

Machine Learning is the science of developing computer algorithms that give computers the ability to learn from observations. Figure 1 conceptually illustrates the overall research methodology employed in this paper and its comparison in relation to state-of-the-art survivability prediction methods. Figure 1 illustrates the novel inclusion of unsupervised learning methods (i) SOM and (ii) DBSCAN for segmenting patient records as the basis for improving the survival prediction performances of MLP classifiers. The approach implements a divide-and-conquer strategy to predicting survivability of cancer patients. The SOM and DBSCAN are engaged to divide patient records into cohorts of patients with similar properties. The approach is entirely data driven and has several advantages over ad-hoc approaches. While ad-hoc methods often divide records on the basis of single attributes such as gender or ethnicity, the proposed approach computes cohorts on the basis of all attributes. The significance of attribute combinations is also easily overlooked and difficult to capture by ad-hoc methods whereas the SOM method can capture similarities of any linear combination of attributes.

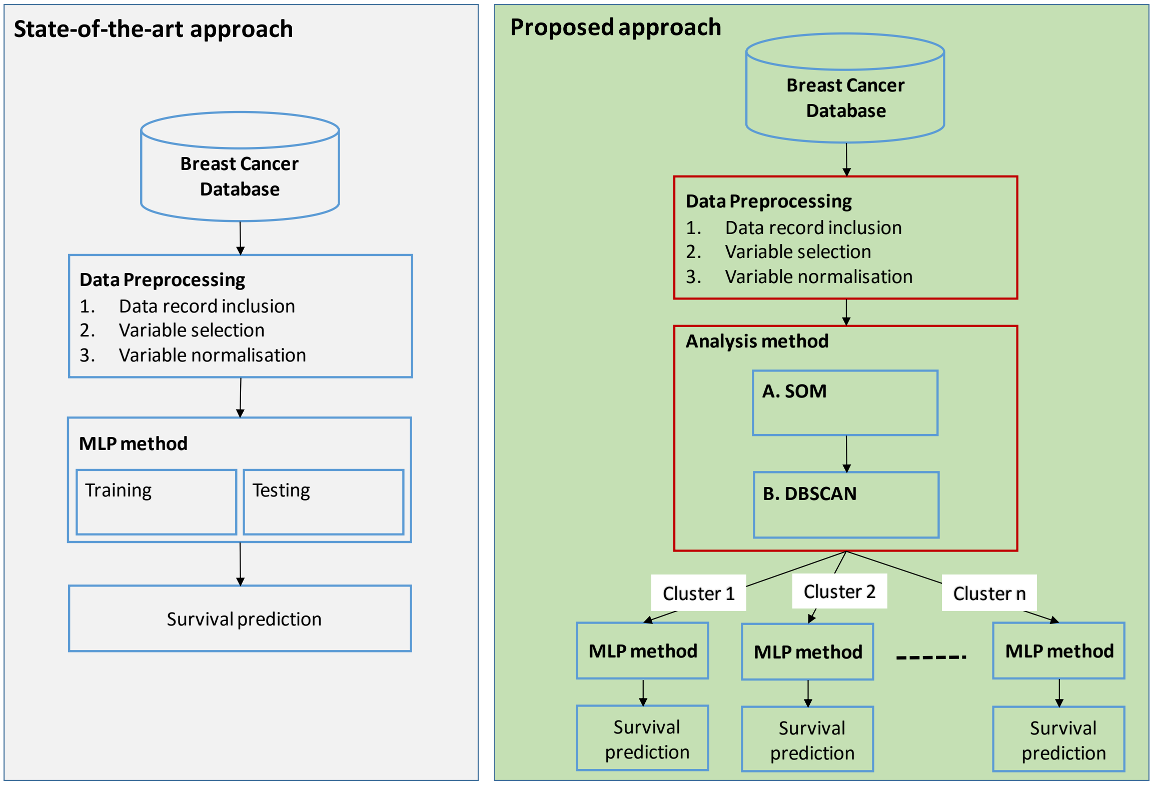


Figure 1: Methodological approach

The proposed methodology starts with a data preprocessing step which involves (i) a data driven approach to selecting patient records and data variables for analysis and (ii) normalizing selected data variables so that all the variables have a unit norm, which is useful in distance computation. More information about the data preprocessing step is discussed in Section 4.

The SOM is deployed after the data-pre-processing step. The SOM maps the data onto a low-dimensional display space and, by doing so, consolidates similar samples into denser groups. The SOM algorithm hence helps to form clusters although the algorithm cannot actually identify or extract the clusters. Clustering methods such as K-means or DBSCAN are commonly used to extract clusters from a SOM. Since we cannot assume that the clusters of patient cohorts are globular (round) in shape and hence we use DBSCAN with our approach. DBSCAN identifies and extracts arbitrary shaped clusters and requires low user-defined parameterization. Once, these clusters are identified, these are then used individually to perform MLP based survival prediction analysis. Here, we have used MLP to test if prediction performances are improved when MLP-classifiers are trained on individual clusters (or patient cohorts). The rationale of the approach is that the creation of models which are expert in predicting survivability of patients that share similarities this should improve on overall prediction accuracy when compared to a single holistic classifier. Also, we have adopted a new strategy of replacing missing values of a patient with the most frequent variable value in a patient cluster identified by SOM-DBSCAN (see Fig 2). This was done for all the clusters identified using SOM-DBSCAN approach. The main advantage of this strategy was to replace missing values of patients in a cluster by the variable values from the same cluster. This was done to accurately impute missing values as each cluster contains sub-group of patients sharing similar properties. Previous studies, while conducting survivability analysis, used overall dataset for missing value imputation which could be inaccurate. Figure 2 illustrates this procedure of imputing missing values. It was revealed that the MLP performances increased by using abovementioned procedure (see Section 6).

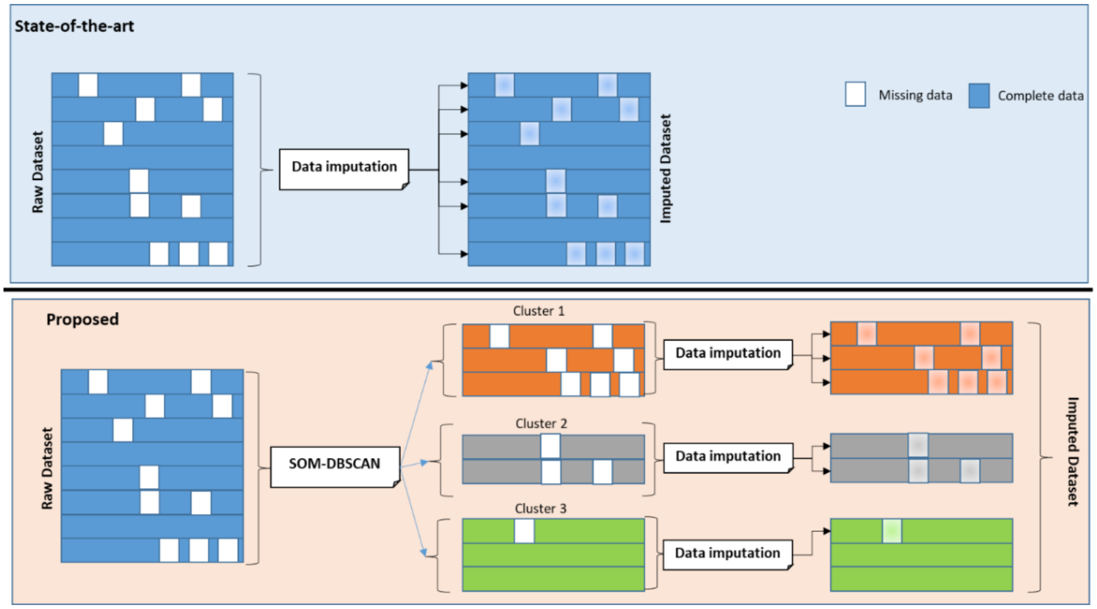


Figure 2: Missing data imputation using SOM-DBSCAN approach

We wish to emphasize that the objective of this paper is not the classification of cancer patient data but rather the development of a methodology that can improve the accuracy of classification systems as well as to provide an explanation facility on what factors or combination of factors influence the prediction accuracy. The proposed method can hence be considered to a precursor to a classification system (such as the MLP). Following section details SOM and DBSCAN method in detail.

3.1 Self-Organizing Maps (SOM)

Artificial Neural Networks (ANNs) simulate learning capabilities of the mammalian biological neural systems. Research on Artificial Neural Networks made great strides in recent years, and are being applied very successfully to a very wide range of hard to solve problems. MLPs are a type of ANNs which are being trained in a supervised fashion. MLPs have been deployed to cancer survivability prediction on a number of occasions [12, 13]. This paper will adopt an unsupervised ANN called the SOM. While MLPs are a black box method for classification problems, the very purpose of SOMs is to provide greater insights into the underlying learning problem. SOMs are also capable of modelling data with missing values.

The SOM is a machine learning algorithm that is popularly applied in Data Mining [25]. SOMs are generally suited for tasks that require clustering, dimensionality reduction, or the visualization of high dimensional data. The SOM algorithm is unsupervised and hence requires little a-priori information about the data. It is thus a great method to help researchers in understanding the dataset under consideration.

The SOM algorithm projects high dimensional data onto a low-dimensional display space called the map. The map consists of a regular grid of dimension N where N is generally chosen to be much smaller than the input dimension of a given learning problem. Associated with each point on the grid is a real valued codebook vector whose dimension matches that of the input vectors. Please refer to the supplementary file for understanding the details of the SOM learning process.

A particular strength of the SOM algorithm is that the SOM preserves the topology when mapping data [25]. Thus, data that are similar to each other in data space remain close when projected onto the map whereas dissimilar data will be mapped distant on the map. By using a map of two dimensions this allows for the visualization of high dimensional data. It is for this reason that two-dimensional SOMs are by far the most common.

Resent research introduced high-resolution SOMs which are able to reveal intricate details about the data and also allow SOMs to be applied to limited supervised learning problems [26]. The paper will adopt the high resolution SOM for the following reasons:

(i) It is a scalable method which allows the visualization of data.
(ii) It is best suited as a tool for developing a better understanding of the data and the underlying learning problem and, in particular,
(iii) It allows an investigation into the existence of correlation of patient descriptive data with survivability of breast cancer.

The application of the SOM to the learning problem of breast cancer survivability will enable us to answer the following questions:

(i) What renders breast cancer survivability such a difficult to achieve problem?
(ii) Are there cohorts of similar cases of patients and, if so, which are these cohorts and do they share a similar outcome in terms of survivability of breast cancer?
(iii) Is breast cancer survivability predictable and, if so, which circumstances contribute to the accuracy of the prediction?

Moreover, the SOM will be deployed to help understand the results of related works, and to provide future direction for methods on breast cancer survivability prediction.

3.2 DBSCAN method

DBSCAN is an algorithm, which identifies clusters among a set of records by detecting areas of high density [27]. There are many advantages of using DBSCAN on the dataset [27]: (i) it can detect the number of clusters and does hence not require a predetermined number of clusters; (ii) can find arbitrary shaped clusters; (iii) robust to noise and outliers; and (iv) requires only two algorithmic parameters (minimum number of records in a cluster) which can be easily set. In this study, we have used the resulting two-dimensional map from SOM for DBSCAN based clustering for identifying clusters of patient records sharing similar characteristics. Supplementary file contains preliminary information on DBSCAN based clustering process.

After applying DBSCAN, a set of clusters of patient cohorts is identified to analyse patient survivability. Each cluster represents cohort of patients, which share similar characteristics and survivability. The patient cohorts in each clusters can also be used for conducting MLP based classification for improved patient survivability analysis. More information about the results of such analysis have been provided in Section 6.

4. Data Description

In this study, the breast cancer incidence dataset (publically available) from the Surveillance, Epidemiology and End Results (SEER) program is used (http://www.seer.cancer.gov). The data files stored in the Cancer Incidence database for the years 1973 -2012 are used. The dataset consists of 740,506 records and 146 variables for the breast cancer cases identified in the US. The variables provide detailed information about the cancer case including tumor related attributes, staging, patient socio-demographics, mortality and multiple edition/recode of certain variables (e.g. AJCC cancer staging).

We have first used a set of inclusions to select the data for further analysis. All of the inclusions were applied one-by-one to the SEER dataset so that only those records are included where the cause of death was due to breast cancer and complete survival history is known. These specific inclusions are listed as following:

(i) Cases were included when the year of birth is known. Records with missing values were included except if the year of birth is not known.

(ii) Cases were included when the site/histology recode or the adapted classification scheme for tumors (AYA Site Recode/WHO 2008) is "Carcinoma of breast".

(iii) Cases were included when the SEER cause-specific death classification equals 1, i.e., cases where included only when a person died of the cancer.

(iv) Cases were included when the complete dates are available and there are more than 0 days of survival (i.e. Survival Months Flag = 1)

(v) Cases were included when the behaviour recodes were Malignant, Only malignant in ICD-O-3, Only malignant 2010+

All of these inclusions were applied to the dataset and a total of 85,189 cases were selected for further analysis. In terms of variable selection, variables such as FIRSTPRM, BEHANAL and SRV_TIME_MON_FLAG were removed as they had only 1 unique value in the resulting dataset. It could also be noted that although HER2 status has been of interest to breast cancer [28, 29], the current dataset includes measurement of HER2 only after 2010, hence, it was not included for this study. Then, the following procedure was adopted to select variables.

Rather than using ad-hoc criteria, which previous studies have adopted, we use a set of informed criteria to select records and variables from the SEER data as follows: To select relevant variables we have computed the information gain based on entropy for each of the 146 attributes in the SEER dataset. We then select the 29 attributes with (i) the highest information gain, or (ii) its relevance for survival analysis for the breast cancer patients.

Table 1 shows that age, race, marital status, primary site, histology, cancer staging such as tumor size, node status, distant spread or extension of tumor are commonly used attributes in previous studies. Laterality also has an impact on breast cancer survivability [9, 30]. AJCC staging system provided more comprehensive assessment of breast cancer staging than the variables, extent of disease and stage [31], therefore AJCC staging is selected as a variable. Finally, we ended up with 26 variables and 85,189 cases for analysis (see Table 2). The variables selected are – Registry ID, Marital status at diagnosis, ethnicity, Spanish/Hispanic Origin, gender, age at diagnosis, year of birth, sequence of all reportable malignant, year of diagnosis, primary site, Laterality, Surgery of Primary Site, scope of Regional Lymph Node surgery, reason no cancer-directed surgery, method of radiation therapy, Radiation sequence with surgery, Number of primaries, First malignant primary indicator, Histology, site/histology recode, tumor marker 1, tumor marker 2, Survival Months Flag, Survival Months, Breast Adjusted AJCC 6th T , Breast Adjusted AJCC 6th N, Breast Adjusted AJCC 6th M, Breast Adjusted AJCC 6th Stage, behaviour recode and SEER cause-specific death classification.

Table 3 lists variables (such as ADJTM_6VALUE, ADJNM_6VALUE, ADJM_6VALUE, ADJAJCCSTG and SURGPRIM) which have the most number of missing values in the SEER dataset.

The variables containing discrete sequences of values such as *AGE_DX*, *YR_BRTH*, *SEQ_NUM*, *DATE_yr*, *NUMPRIMS* are scaled to values between 0 and 10. This is done so that the difference in the scales of each continuous variable is normalised between 0 and 10 and also because most of the variables in the selected dataset have ranges between 0 to 10. We have not considered *SRV_TIME_MON* for the normalisation process as it is a target value and it is not used in the procedure. In case of categorical variables, we utilise a binary encoding procedure where each categorical variable is transformed into a set of binary variables in such a way that each categorical value is associated with one of the binary variable. For e.g. variable LATERAL (having 5 unique categories – 1, 2, 3, 4 and 9) is replaced by 5 binary variables each representing unique categories (1 – present, 0 – not present). After this transformation, all the nominal variables are then treated as numeric variables in the domain of {0,1}.

Table 2: Unique values in SEER variables used for the data preparation

Variable Name	**Description**	**Unique Values**
REG	*Registry ID*	8
MAR_STAT	*Marital status at diagnosis*	6
RACE	*Ethnicity*	29
ORIGIN	*Spanish/Hispanic Origin*	10
SEX	*Gender*	2
AGE_DX	*Age at diagnosis*	88
YR_BRTH	*Year of birth*	111

SEQ_NUM	*Sequence of all reportable malignant*	2
DATE_yr	*Year of diagnosis*	40
SITEO2V	*Primary site*	9
LATERAL	*Laterality*	5
SURGPRIM	*Surgery of Primary Site*	7
NO_SURG	*Reason no cancer-directed surgery*	8
RADIATN	*Method of radiation therapy*	10
RAD_SURG	*Radiation sequence with surgery*	7
NUMPRIMS	*Number of primaries*	6
FIRSTPRM	*First malignant primary indicator*	1
HISTREC	*Histology*	7
ERSTATUS	*Tumor marker 1*	5
PRSTATUS	*Tumor marker 2*	5
SRV_TIME_MON_FLAG	*Survival Months Flag*	1
BEHANAL	*Behavior recode*	1
SRV_TIME_MON	*Survival Months*	447
ADJTM_6VALUE	*Breast Adjusted AJCC 6th T*	16
ADJNM_6VALUE	*Breast Adjusted AJCC 6th N*	7
ADJM_6VALUE	*Breast Adjusted AJCC 6th M*	5
ADJAJCCSTG	*Breast Adjusted AJCC 6th Stage*	12

5. Results

The dataset obtained from the pre-processing step is used for training the self-organising map (SOM). This paper uses a 2D SOM to aid the visualization of results. SOM is useful in present situation as the data variables may be non-linearly related to each other. The mappings obtained by a SOM are useful for visualising complex relationship among data variables. The missing values in the dataset are retained and encoded by the SOM. We have trained SOM by varying parameters (such as size, radius, learning rate and iteration). An initial learning rate of 0.32, a radius of size 750, and an iteration of 10000 with 1500 grid points for each of the two dimensions is used when training. This parameter setting was used as it offered better separation among data vectors within reasonable computational time. The mapping (2D-coordinate values) obtained from the trained SOM is then passed through DBSCAN for clustering analysis.

DBSCAN algorithm was run on the dataset with the minimum records in cluster value is twice the value of ε. The results of DBSCAN is shown in Figure 3. It can be seen that DBSCAN has identified 9 irregular shaped clusters.

Table 3: SEER Variables with missing data

Variable Name	**Frequency of Missing Values**
SURGPRIM	62696
ADJTM_6VALUE	37383
ADJNM_6VALUE	37383
ADJM_6VALUE	37383
ADJAJCCSTG	37383

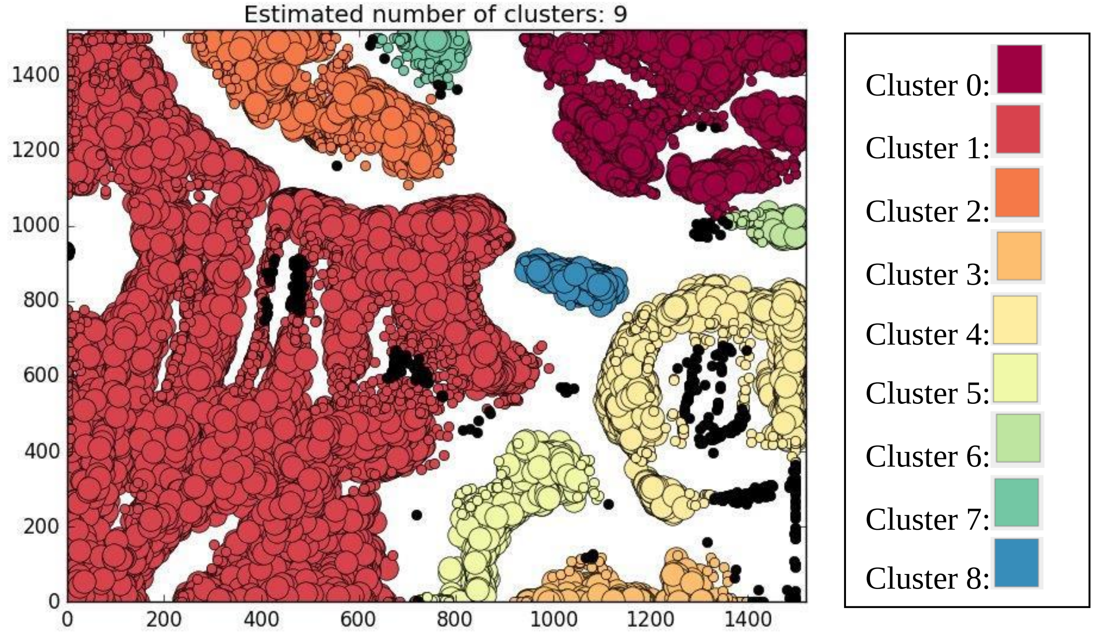


Figure 3: DBSCAN results for SOM with size 1500x1500

Given that a SOM maintains a topology preserving mapping and hence it can be anticipated that data which are mapped to the same cluster share similar properties. For example, when analysing records that share the same cluster on the basis of the property on survival time (in months) we found that such records feature a greater similarity in survival time then records that belong to different clusters. This is illustrated in Figure 4. Figure 4 presents a box plot for the survivability of patients for each of the 9 clusters. It can be seen that the survivability differs considerably among 9 clusters. This suggests that the attribute values that are selected in the clusters have significant effect on the survivability of the breast cancer patients. Table 4 shows the size of clusters obtained by using DBSCAN on the SOM. It is seen that some cohorts (i.e. cluster 1) are much larger than other cohorts (i.e. cluster 6). It will be found that correspond to the frequency of occurrence of patient attributes which differ from cluster to cluster.

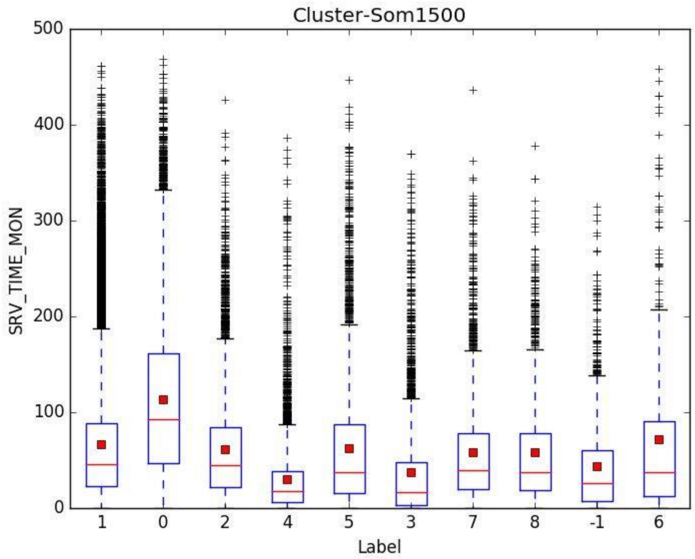


Figure 4: Survival time for cases in 9 clusters in SOM, cluster with label -1 represents outliers which are not attached to any of the 9 clusters

Table 4: Cluster sizes based on the results of DBSCAN for SOM size 1500

Cluster	**Size**	**Mean Survival Time (in months)**	**Standard Deviation (in months)**
0	10896	113.17	55.36
1	48383	66.18	34.23
2	6626	61.69	31.85
3	3741	37.01	17.86
4	5298	29.91	16.14
5	4198	62.77	34.55
6	399	71.37	32.23
7	2756	58.27	28.76
8	1625	58.82	27.24

It can be seen that Cluster 0 has the highest mean survivability, while the patients in cluster 3 and 4 are in the lower end of survivability. When the characteristics of each of these clusters were analysed, Cluster 0 has the lowest number of bilateral involvement of breast cancer, while the Cluster 4 has the highest numbers of bilateral involvement, followed by the cluster 3. Common characteristics from each cluster can be interpreted from the results. The results are analysed using the staging definitions from the AJCC staging [32]. To illustrate the difference in cluster characteristics, Cluster 0 and 4 is selected to be summarized here. Cluster 0 has higher N0 compared to other clusters, which is although patients in Cluster 0 are diagnosed with breast cancer, the cluster involves higher number of patients with no lymph node involvement. The results indicated that Cluster 0 accumulates patients with breast cancer in local stage, i.e Stage I and II of TNM classification. That indicates that there is less distant metastasis or the local spread to the lymph nodes. Cluster 0 has lower number of TIIIS, which indicates the size of the tumor is less than 50 mm in greater diameter. Cluster 0 has low number of M1, while clusters 3 to 6 all have lower M0 and higher M1. However, clusters 3 and 4 have higher M1. This may explain why cluster 0 has higher survivability from breast cancer than other clusters.

Highest number of IIINOS are in cluster 4, and a high number of stage IV are also in cluster 4. Cluster 3, 4 and 5 have lesser numbers of TI and TII. Therefore, it could be seen that these clusters have less number of people with tumor size less than 2 cm or 5 cm. Both Clusters 3 and 4 have low numbers of stage I, IIA and II B and higher number of Stage IIIC and IV. That indicates that the tumor has spread to ipsilateral supraclavicular lymph nodes, infraclavicular lymph nodes, internal mammary lymph nodes or axillary lymph nodes and/or to different organs. Very high number of cases in cluster 4 consists of either the patient died before recommended surgery or unknown reason for no surgery or the patient or the patient's guardian refused to have a surgery.

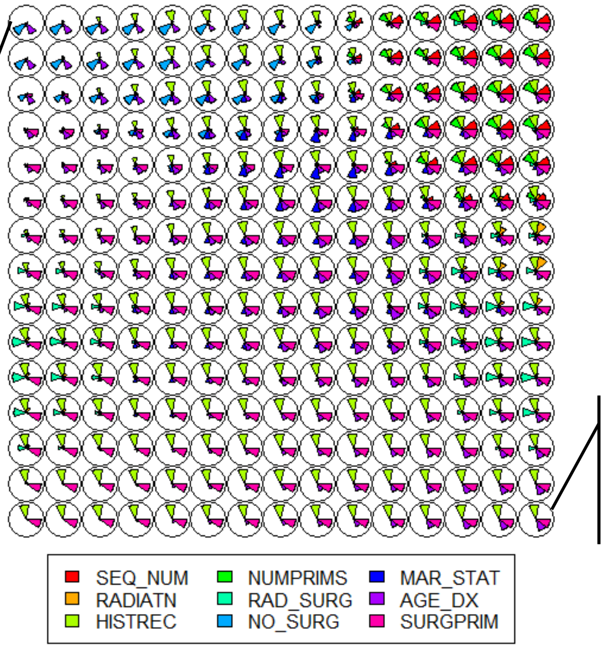
The aforementioned results indicate that SOM-DBSCAN based approach can help explain, through patient attributes, variation among the survivability of breast cancer patients. It can be also seen that there are several clusters of similar cases of patients (see Fig. 3) and these cohorts share a similar outcome in terms of survivability of breast cancer (see Fig. 4 and Table 4).

Following section discusses the results obtained from clustering approach together with the results from obtained by survivability analysis conducted in literature by other researchers on the SEER dataset.

6. Discussion

To better visualize the weight vectors across the SOM, we have selected 9 original variables (*SEQ_NUM, NUMPRIMS, RADIATN, HISTREC, NO_SURG, SURGPRIM, AGE_DX, MAR_STAT*) to visualise their value distribution on the SOM by grouping 15 by 15 codebook vectors. The original SOM (shown in Fig 3) is of size 1500 by 1500 but here we have used 15 by 15 size to illustrate its effectiveness in mapping multidimensional dataset. Figure 5 illustrates the visualization of the node weight vectors (aka *codes*). The node weight vectors are made up of normalised values of the original variables used in generating the SOM. Each node in the map (see Figure 5) constitutes the magnitude of each variable in the weight vector. By visualizing these weight vectors we can identify patterns in the distribution of records and variables. For e.g., top left nodes shows that *NO_SURG* and *AGE_DX* are similar for the subset of records which are mapped on these nodes. This analysis confirms that the SOM is effective in grouping patients according to similarities in feature space and, paired with DBSCAN, in obtaining clusters of patients that exhibit similar properties.

Illustrates the magnitude of NO_SURG and AGE_DX in the weight vector
Also, patient cases with similar NO_SURG and AGE_DX values are mapped here



HISTREC, SURGPRIM and AGE_DX are similar for the cases mapped here

Figure 5: Visualising the weight vectors across the SOM (only selected variables are considered for 15 by 15 node map)

Previous studies have shown that the accuracy of cancer survivability prediction systems can vary significantly among different cohorts of patients. Existing work use either a holistic approach in which a classifier is designed to predict survivability of any patient or use classifiers that are designed on a limited number of cohorts (i.e. limited to a particular race, or only those who went through the menopause). Systems that specialize on cohorts perform generally much better then holistic systems but are limited to predicting outcomes only for patients that fit a given cohort. Moreover, the specialized systems group patients on the basis of a single attribute. Multiple attributes or combination of attributes are not used due to the combinatorial difficulty in studying possible attribute combinations.

The SOM-DBSCAN methodology presented in this paper allows the development of prediction systems that can be deployed to cohorts that share similarities in a combination of attributes. This data driven approach detects cohorts without human intervention. Based on the work presented in this paper it becomes possible to develop specialized classification systems by, for example, training a classifier on each of the detected cohorts (the samples that share the same cluster). For any new patient, we would identify the cluster to which this new sample would map to and then consult the classifier which corresponds to that cluster.

We verify the effectiveness of such an approach. Multiplayer perceptron (MLP), a feedforward artificial neural network, is commonly considered for predicting cancer survivability tasks. For simplicity, we train the MLP to map the data into cases with survivability of less than 5-year (class 1) and the rest of the cases are to be mapped to class 0 to represent cases that survive for five or more years. We first used a holistic approach to obtain a baseline result by training the MLP on the whole set of data. In the baseline MLP training, we have replaced missing values of a variable with the most frequent variable value. Table 5 illustrates the performance of baseline MLP on train, test and validation data when all 85,189 patients records are used. It can be seen that MLP can correctly classify ~ 63% of the patient records. The result is similar to results reported in literature.

Table 5: Classification performance of MLP on all the data records without clustering

	MLP		
Data	**Training (%)**	**Test (%)**	**Validation (%)**
85,189	68.68	63.27	63.16

We then trained one MLP for each of the 9 clusters (as was obtained by the proposed SOM-DBSCAN based clustering method). It should be noted that missing value of patient in a cluster is replaced with the most frequent variable value in the same cluster. In other words, we made this replacement for each of the nine MLPs so that patient attributes from the same cohort are used for replacement. Table 6 shows the results obtained after applying these MLPs to cases in the validation and test set. It can be seen that for most of the clusters (test and validation data), the MLP performance is well above the baseline of ~ 63% (see Table 5). For instance, the MLP performance, when considering records identified as cluster 0, is 72.98% on train, 73.18% on test and 72.43% on validation data. The 10-fold cross validation performances for MLP on cluster 1 (with high mean survivability) and cluster 4 (with low mean survivability) is shown in Figure 6. It shows that the proposed approach is robust in terms of consistently predicting survivability at an appropriate performance level.

Table 6: Classification performance of MLP on the data points mapped in each cluster

						MLP		
Cluster ID	**Cluster Size**	**Mean Survivability (months)**	**Patients surviving <5yrs**	**Patients surviving >5yrs**	**Ratio (<5yrs:Total)**	**Training (%)**	**Test (%)**	**Validation (%)**
0	10896	113.17	7245	3651	0.66	72.98	**73.18**	**72.43**
1	48383	66.18	29337	19046	0.61	64.55	62.73	**63.51**
2	6626	61.69	4137	2489	0.62	62.82	62.69	62.86
3	3741	37.01	3008	733	0.80	80.41	**80.43**	**80.43**
4	5298	29.91	4602	696	0.87	86.84	**86.96**	**86.96**
5	4198	62.77	2714	1484	0.65	65.01	**64.99**	**65.61**
6	399	71.37	262	137	0.66	65.45	**72.05**	**71.54**
7	2756	58.27	1831	925	0.66	66.45	**66.55**	**66.58**
8	1625	58.82	1074	551	0.66	66.91	**67.59**	**69.57**

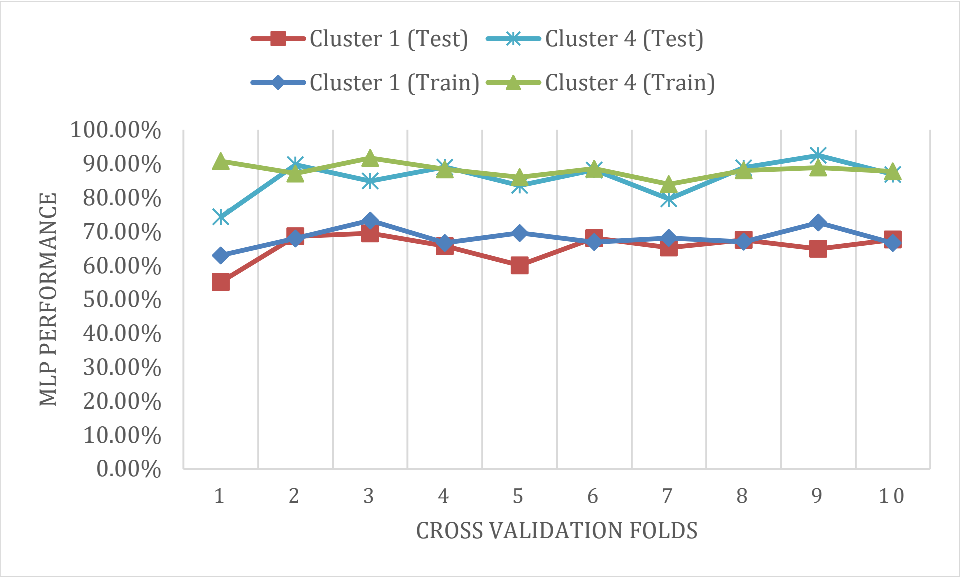


Figure 6: Ten-fold cross validation performance of MLP on clusters 1 and 4

The results confirm that the overall classification rate is significantly improved by using the proposed SOM-DBSCAN method and missing value replacement strategy. The prediction accuracy can be as high as 86.96% for patients who share the properties of cluster 4. To the best of our knowledge, the quality of the overall results is better than any other method which is capable of predicting breast cancer survivability of any patient and including those with missing values. To this end, the proposed approach based on SOM-DBSCAN can identify or segment raw dataset into several meaningful subgroups, which share common attributes and survivability, and these subgroups can also be used to improve baseline MLP performances. Thus, it can be construed that the survivability of patients mapped to these clusters can be predicted. It should be also noted that we have conducted 5-year survivability analysis only to compare our results with the state-of-the-art survivability prediction algorithms. The proposed methodology can also be applied to identify high risk subgroup of cancer patients even within or more than 5-years survivability.

We conducted another set of experiments to analyse situations when the survivability cut-off years is considered to be 3 years, 5 years (as discussed above) and 7 years. MLP performances on each of the 9 clusters identified by SOM-DBSCAN method is shown in Table 7. It can be broadly seen that MLP performances on each of the clusters varies considerably when cut-off year for survivability period is changed to 3, 5 and 7 years. Table 7 shown that MLP performance for 5 years survivability is lower for all the 9 clusters when compared against 3 and 7 years survivability. For instance, MLP performance on cluster 1 (with high mean survivability of ~66 months) is improved in case of 3 and 7 years survivability (~70% and ~73%) when compared against 5 years survivability (~65%). However, in case of cluster 4 (with low mean survivability of 29.9 months) MLP performance increased to ~90% when 7 years survivability is considered. It can be also seen that the MLP performances is better even when the class distribution (for e.g. <3 yrs or >3yrs) is imbalanced caused due to the changes in cut-off period (3, 5, 7 yrs). In other

Table 7: MLP performances in case of 3, 5 and 7 years survivability period

Cluster ID	Cluster Size	Mean Survivability (months)	Ratio (<3yrs:Total)	Ratio (<5yrs:Total)	Ratio (<7yrs:Total)	MLP (survival = 3yrs)		MLP (survival = 5yrs)		MLP (survival = 7yrs)	
						Training (%)	Test (%)	Training (%)	Test (%)	Training (%)	Test (%)
0	10896	113.17	0.19	0.66	0.46	83.24%	85.13%	73.40%	75.56%	69.00%	70.13%
1	48383	66.18	0.41	0.61	0.73	70.24%	70.04%	68.14%	65.22%	73.88%	72.90%
2	6626	61.69	0.42	0.62	0.75	66.05%	66.42%	69.03%	66.29%	76.64%	74.38%
3	3741	37.01	0.68	0.80	0.87	76.29%	75.56%	82.82%	80.90%	87.51%	87.13%
4	5298	29.91	0.73	0.87	0.93	76.33%	75.23%	88.10%	85.74%	93.53%	90.48%
5	4198	62.77	0.49	0.65	0.74	70.00%	70.30%	68.80%	68.31%	76.02%	74.93%
6	399	71.37	0.48	0.66	0.72	70.88%	69.87%	75.92%	71.95%	82.06%	77.93%
7	2756	58.27	0.47	0.66	0.77	69.22%	64.98%	72.91%	71.25%	78.49%	77.52%
8	1625	58.82	0.48	0.66	0.77	69.93%	72.19%	76.19%	72.86%	82.42%	74.76%

Words, the MLP performances are better than the random classifier (built by assigning classes based on class ratios shown in Table 7). Table 7 highlights these class imbalances as well.

This analysis can also aid decision maker to decide best survivability period to get better survival prediction accuracies. Thus, we can conclude that the proposed approach can also be used to identify best cut-off survivability years to further improve survivability prediction performances (in comparison to the standard 5 years survivability period).

In comparison with previous studies, the use of SOM and DBSCAN for survivability studies is novel. The SOM is trained unsupervised whereas all other studies consider various types of supervised methods. We are only aware of one study which considers an unsupervised component: A semi-supervised learning (SSL) approach in [13]. They compute a smoothed similarity measure of k-nearest neighbour for each of the input samples. A key difference to the SOM is that the SSL does not consider the relationship of samples that are outside the *k*-nearest neighbourhood. It is reported in [13] that their method can achieve a prediction accuracy of 73%, and when combined with a supervised methods the results improve to about 78%. One may argue that a random classifier would achieve the same accuracy because the dataset is unbalanced (79% of cases are in the positive class). However, they use random sampling to obtain a class-balanced dataset and hence their method boasts a 28% improvement over a random classifier.

Nevertheless, the study in [13] as well as all other studies on this research question use a holistic approach in that they study the dataset as a whole rather than segmenting (clustering) the samples then addressing the simpler sub-problems.

7. Conclusions

Understanding cancer survivability for patients based on historical medical records is of significant interest among researchers. Challenges associated with missing data, large number of data variables, selection of cut-off survivability period and holistically applying one prediction model over all dataset renders the survivability analysis difficult to resolve. This paper introduced a novel approach to addressing breast cancer survivability by (i) using the information gain or data-driven method for variable selection instead of manual variable selection methods used in prior studies, (ii) using a unsupervised learning method, SOM, which can handle missing values and map raw patient records onto lower dimensions and (iii) finally using DBSCAN to segment the lower dimensional mapped data into subsets of patient cohorts (or clusters) which share commonalities in value as well as in survivability. Hence, the proposed method created multiple patient subgroups with different properties which helps in improving the baseline accuracy of MLP classifiers after missing data imputation. Also, MLP analysis showed that selection of cut-off survivability period significantly impacts MLP performances. This means that decision makers can select appropriate MLP classifier and survivability years for new patients, mapped onto one of these clusters, to have better survivability prediction accuracies. It is hence interesting to study and refine these methods in future work to improve the prediction capabilities of classifiers for cancer survivability analysis.

In terms of limitations, the SOM presented in this paper is only two dimensional which causes a very significant compression of the data and hence carries the risk of information loss. The 2D SOM is used to help with the visualization of results whereas in practical systems it should be advisable to consider higher dimensional SOMS (i.e. 3D or 4D) for further improvements in results. Moreover, DBSCAN is a partitional cluster analysis algorithm, which cannot detect sub clusters. Hierarchical cluster analysis algorithms can be considered to obtain more finely granulated cohorts of patient for improving the accuracy of prediction systems even further. In fact, hierarchical clustering can create sub clusters at any granularity and hence it would be possible to develop very personalized classification systems. Nevertheless, this paper has shown that the general principle of projection and clustering introduces a data driven approach to obtaining patient cohorts for the development of accurate classifications systems. The future work can focus on using the proposed approach for classification other databases (such as MIAS) for classification.

Acknowledgements
Authors also acknowledge the funding received from the EIS seed grant (University of Wollongong) in 2016 to carry out this research study. Authors thank the National Cancer Institute, Surveillance, Epidemiology, and End Results (SEER) Program for providing the access to the research data sets used in this study.

Conflicts of Interest
None declared.

Abbreviations

AJCC: American Joint Committee on Cancer
ALL-AML: Acute Lymphocytic leukemia and Acute Myeloid Leukemia
ANN: artificial neural networks
DBSCAN: density-based clustering Algorithm
DLBCL: diffuse large B cell Lymphoma
DT: decision trees
KRBM: Kent Ridge Bio-Medical
LYMLLEUK: Lymphoma of all sites and leukemia
M: Metastsis (distant spread)
MAR: multiple association rules
MLP: Multi Layered Perception
N: local lymph node involvement
NB: Naïve Bayes
RBF: radial basis function
RNN: recurrent neural network
SEER: Surveillance, Epidemiology, and End Results program
SOM: self-organising Map
SSL Semi-supervised Learning
SVM: support vector machine
T1-4: Size of tumor in greatest dimension
TNM: Tumor, Node, Metastasis

References

1. Tjandra J, Collins JP: **Breast surgery.** In *Textbook of Surgery*

3rd edition edition. Edited by Clunie GJA, Tjandra J, Smith JA, Kaye AH. Singapore: Blackwell; 2008

2. Lippman ME: **Breast Cancer.** In *Harrison's Principles of Internal Medicine, 19e.* Edited by Kasper D, Fauci A, Hauser S, Longo D, Jameson JL, Loscalzo J. New York, NY: McGraw-Hill Education; 2015
3. Grau JJ, Zanon G, Caso C, Gonzalez X, Rodriguez A, Caballero M, Biete A: **Prognosis in women with breast cancer and private extra insurance coverage.** *Annals of Surgical Oncology* 2013, **20:**2822-2827.
4. Owusu C, Lash TL, Silliman RA: **Effect of undertreatment on the disparity in age-related breast cancer-specific survival among older women.** *Breast Cancer Research and Treatment* 2007, **102:**227-236.
5. Pollock YYG, Blackford AL, Jeter SC, Wright J, Cimino-Mathews A, Camp M, Harvey S, Asrari F, Schoenborn NL, Stearns V: **Adjuvant radiation use in older women with early-stage breast cancer at Johns Hopkins.** *Breast Cancer Research and Treatment* 2016, **160:**291-296.
6. Curtis E, Quale C, Haggstrom D, Smith-Bindman R: **Racial and ethnic differences in breast cancer survival: how much is explained by screening, tumor severity, biology, treatment, comorbidities, and demographics?** *Cancer* 2008, **112:**171-180.

7. Madubata CC, Liu Y, Goodman MS, Yun S, Yu J, Lian M, Colditz GA: **Comparing treatment and outcomes of ductal carcinoma in situ among women in Missouri by race.** *Breast Cancer Research and Treatment* 2016**:**1-10.

8. Shiyanbola OO, Sprague BL, Hampton JM, Dittus K, James TA, Herschorn S, Gangnon RE, Weaver DL, Trentham-Dietz A: **Emerging trends in surgical and adjuvant radiation therapies among women diagnosed with ductal carcinoma in situ.** *Cancer* 2016, **122:**2810-2818.

9. Kheirelseid EAH, Jumustafa H, Miller N, Curran C, Sweeney K, Malone C, McLaughlin R, Newell J, Kerin MJ: **Bilateral breast cancer: Analysis of incidence, outcome, survival and disease characteristics.** *Breast Cancer Research and Treatment* 2011, **126:**131-140.

10. Jobsen JJ, van der Palen J, Ong F, Riemersma S, Struikmans H: **Bilateral breast cancer, synchronous and metachronous; differences and outcome.** *Breast Cancer Research and Treatment* 2015, **153:**277-283.

11. Killander F, Karlsson P, Anderson H, Mattsson J, Holmberg E, Lundstedt D, Holmberg L, Malmström P: **No breast cancer subgroup can be spared postoperative radiotherapy after breast-conserving surgery. Fifteen-year results from the Swedish Breast Cancer Group randomised trial, SweBCG 91 RT.** *European Journal of Cancer* 2016, **67:**57-65.

12. Delen D, Walker G, Kadam A: **Predicting breast cancer survivability: a comparison of three data mining methods.** *Artificial Intelligence in Medicine* 2005, **34:**113-127.

13. Park K, Ali A, Kim D, An Y, Kim M, Shin H: **Robust predictive model for evaluating breast cancer survivability.** *Engineering Applications of Artificial Intelligence* 2013, **26:**2194-2205.

14. Bellaachia A, Guven E: **Predicting breast cancer survivability using data mining techniques.** *Age* 2006, **58:**10-110.

15. Boughorbel S, Al-Ali R, Elkum N: **Model Comparison for Breast Cancer Prognosis Based on Clinical Data.** *PLoS ONE* 2016, **11:**e0146413.

16. Rathore N, Divya, Agarwal S: **Predicting the survivability of breast cancer patients using ensemble approach.** In *2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014; Ghaziabad*. IEEE Computer Society; 2014: 459-464.

17. Lotfnezhad Afshar H, Ahmadi M, Roudbari M, Sadoughi F: **Prediction of breast cancer survival through knowledge discovery in databases.** *Global journal of health science* 2015, **7:**392-398.

18. Umesh DR, Ramachandra B: **Association rule mining based predicting breast cancer recurrence on SEER breast cancer data.** In *2015 International Conference on Emerging Research in Electronics, Computer Science and Technology, ICERECT 2015*. 2016: 376-380.

19. Civcik L, Yilmaz B, Özbay Y, Emlik GD: **Detection of microcalcification in digitized mammograms with multistable cellular neural networks using a new image enhancement method: Automated lesion intensity enhancer (ALIE).** *Turkish Journal of Electrical Engineering and Computer Sciences* 2015, **23:**853-872.

20. Jalalian A, Mashohor SBT, Mahmud HR, Saripan MIB, Ramli ARB, Karasfi B: **Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: A review.** *Clinical Imaging* 2013, **37:**420-426.
21. Sokouti B, Haghipour S, Tabrizi AD: **A framework for diagnosing cervical cancer disease based on feedforward MLP neural network and ThinPrep histopathological cell image features.** *Neural Computing and Applications* 2014, **24:**221-232.
22. Devi MA, Ravi S, Vaishnavi J, Punitha S: **Classification of Cervical Cancer Using Artificial Neural Networks.** In *Procedia Computer Science*. 2016: 465-472.
23. Zheng B, Yoon SW, Lam SS: **Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms.** *Expert Systems with Applications* 2014, **41:**1476-1482.
24. García-Laencina PJ, Abreu PH, Abreu MH, Afonoso N: **Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values.** *Computers in Biology and Medicine* 2015, **59:**125-133.
25. *Self-Organizing Maps.* Springer-Verlag New York, Inc.; 2001.
26. Nguyen VT, Hagenbuchner M, Tsoi AC: **High Resolution Self-organizing Maps.** In *AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings.* Edited by Kang BH, Bai Q. Cham: Springer International Publishing; 2016: 441-454
27. Ester M, Kriegel H-P, Sander J, Xu X: **A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise.** In *Book A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise* (Editor ed.^eds.). pp. 226-231. City: AAAI Press; 1996:226-231.
28. Gyawali B, Niraula S: **Duration of adjuvant trastuzumab in HER2 positive breast cancer: Overall and disease free survival results from meta-analyses of randomized controlled trials.** *Cancer Treatment Reviews* 2017, **60:**18-23.
29. Wilson FR, Varu A, Mitra D, Cameron C, Iyer S: **Systematic review and network meta-analysis comparing palbociclib with chemotherapy agents for the treatment of postmenopausal women with HR-positive and HER2-negative advanced/metastatic breast cancer.** *Breast Cancer Research and Treatment* 2017, **166:**167-177.
30. Killander F, Karlsson P, Anderson H, Mattsson J, Holmberg E, Lundstedt D, Holmberg L, Malmstr√∂m P: **No breast cancer subgroup can be spared postoperative radiotherapy after breast-conserving surgery. Fifteen-year results from the Swedish Breast Cancer Group randomised trial, SweBCG 91 RT.** *European Journal of Cancer* 2016, **67:**57-65.
31. Institute NC: **SEER Program, Comparative Staging Guide for Cancer.** In *Book SEER Program, Comparative Staging Guide for Cancer* (Editor ed.^eds.). City: National Institute of Health; 1993.

32. Frederick L Greene DLP, et al. : *AJCC Cancer Staging Manual* Sixth Edition edn. USA: Springer; 2002.