Knowledge-Based Systems 242 (2022) 108407

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A probabilistic data analytics methodology based on Bayesian Belief network for predicting and understanding breast cancer survival

Asli Z. Dag [a], Zumrut Akcam [b], Eyyub Kibis [c], Serhat Simsek [c], Dursun Delen [d,e,*]

[a] Creighton University, Heider College of Business, Omaha, NE, United States of America
[b] Stevens Institute of Technology, Department of Computer Science, Hoboken, NJ, United States of America
[c] Montclair State University, Feliciano School of Business, Montclair, NJ, United States of America
[d] Oklahoma State University, Spears School of Business, Stillwater, OK, United States of America
[e] Ibn Haldun University, School of Business, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 8 August 2021
Received in revised form 18 January 2022
Accepted 8 February 2022
Available online 15 February 2022

Keywords:
Breast cancer
Data mining
Genetic Algorithm
Machine learning
Sensitivity Analysis

ABSTRACT

Understanding breast cancer survival has proven to be a challenging problem for practitioners and researchers. Identifying the factors affecting cancer progression, their interrelationships, and their influence on patients' long-term survival helps make timely treatment decisions. The current study addresses this problem by proposing a Tree-Augmented Bayesian Belief Network (TAN)-based data analytics methodology comprising of four steps: data acquisition and preprocessing, variable selection via Genetic Algorithm (GA), data balancing with synthetic minority over-sampling and random under-sampling methods, and finally the development of the TAN model to determine the probabilistic inter-conditional dependency structure among breast cancer-related variables along with the posterior survival probabilities The proposed model is compared to well-known machine learning models. A *what-if* analysis has also been conducted to verify the associations among the variables in the TAN model. The relative importance of each variable has been investigated via sensitivity analysis. Finally, a decision support tool is developed to further explore the conditional dependency structure among the cancer-related factors. The results produced by the proposed methodology, namely the patient-specific posterior survival probabilities and the conditional relationships among the variables, can be used by healthcare professionals and physicians to improve the decision-making process in planning and managing breast cancer treatments. Our generic methodology can also accommodate other types of cancer and be applied to manage various medical procedures.

1. Introduction

After skin cancer, breast cancer is the most commonly occurring cancer type affecting women in the United States [1]. Although the number of incidences has been relatively stable over the past decade, in the United States alone, more than 40,000 women died from breast cancer in 2020 [1]. The treatment options are generally determined by the stage of cancer, which also influences the long-term prognosis. Advances in medical technology along with the higher rates of breast cancer screening have increased the number of early breast cancer diagnoses, allowing treatment to be initiated in a timely manner. As a result, the 5-year survival rate has increased from 75.2% in 1975 to 90% in 2020 [1].

While numerous factors determine cancer progression and prognosis, these can be broadly grouped into (1) chronological and (2) biological factors [2]. Chronological factors include those that are primarily impacted by time, such as the status of the lymph nodes, size of the tumors, and historical stage. Lymph nodes filter the harmful substances and, if untreated, allow cancer cells to spread through the breast, armpit, and chest wall [2,3]. Tumor size and historical stage similarly influence the likelihood of the spread to the healthy tissues. On the other hand, biological factors describe the behavioral status of the tumor. For example, the histological grade indicates its aggressiveness [2,3], whereas estrogen receptor (ER) and progesterone receptor (PR) are the indicators of the hormonal structure of the tumor and have implications for relative mortality regardless of tumor histology [4]. These and other factors are crucial for a more comprehensive approach to treating cancer [5].

Survival rate prediction (after the first diagnosis) is vital for helping both doctors and patients explore treatment options. In addition, it helps patients to make important possible lifestyle

* Corresponding author.
E-mail addresses: aslidag@creighton.edu (A.Z. Dag), zakcamki@stevens.edu (Z. Akcam), kibise@montclair.edu (E. Kibis), simseks@montclair.edu (S. Simsek), dursun.delen@okstate.edu, dursun.delen@ihu.edu.tr (D. Delen).
URL: http://spears.okstate.edu/delen (D. Delen).

https://doi.org/10.1016/j.knosys.2022.108407
0950-7051/

changes as well as financial planning. For clinicians, the prediction of a 5-year timeframe is sufficient to explore different treatment paths, analyze their respective outcomes, and utilize this data to advise patients.

Machine learning (ML) algorithms are widely used in medical diagnostics, including breast cancer, mammogram classification, mammography anomalies, heart problems [6–10] and have been proven successful in predicting the survival outputs and variable selection analysis [11–15], owing to their ability to discern hidden patterns in complex datasets including liver and kidney disease diagnosis, comorbidity of cancers, and breast cancer survival prediction [16].

Extant research has been dedicated to (a) detecting the critical genes that are associated with breast cancer via using clustering(unsupervised) techniques, and (b) identifying the important demographic and clinical predictors of survival life after the time of breast cancer diagnosis by using various supervised models. These studies are summarized in the following subsections to clearly identify the gaps that exist in the existing body of related literature.

1.1. Genetic data-oriented studies

A significant number of researchers of extant studies in this field tended to examine genetic markers and predictors and those that relied on unsupervised machine learning algorithms, mostly applied clustering methods, to identify the most relevant genes for each cancer type. For example, the unsupervised algorithm developed by Li et al. [17] is capable of detecting groups of genes that have not been previously recognized as risk factors for glioma. This was a significant advantage relative to the standard approaches based on already identified genes that predispose an individual to this type of brain cancer, which suffers from classification issues [18–20]. Instead, the unsupervised model proposed by Li et al. [17] selects two groups of gliomas before clustering these into six subgroups. This allows the identification of different sets of classifiers for each of these subgroups that can be applied to different datasets for validation purposes. The main drawback of this approach stems from the fact that it cannot be used for predicting patient survival. To address this shortcoming, Lapointe et al. [20] proposed a semi-supervised algorithm which they applied to clinical data as a preprocessing step to identify genes for use in the subsequent unsupervised clustering. For this purpose, 7399 genes from 240 breast cancer patients were ranked according to their Cox scores and 160 training observations to identify the 25 most relevant genes for predicting survival times and corresponding probabilities. Although these models are very useful for analyzing data, they cannot be applied in care decision-making and survival prediction, as genetic factors do not provide information on the degree of cancer spread in a particular patient, which would determine the type of treatment to be used.

1.2. Clinical data oriented studies

Researchers that relied on clinical data when developing their survival prediction models mostly utilized supervised Machine learning (ML) algorithms. In their work, Lundin et al. [21] used clinical data obtained from 951 breast cancer patients to predict 5-, 10-, and 15-year survival rates based on Artificial Neural Networks (ANNs). Their model was also capable of identifying the tumor characteristics that are most influential on cancer survival. Delen et al. [22] similarly used ANNs, along with Decision Trees (DTs) and Logistic Regression (LR), to develop a hybrid model for predicting 5-year cancer survival rates. These authors utilized the 433,272 patient records gathered over nearly three decades to identify the key tumor characteristics affecting survivability.

On the other hand, several authors have attempted to improve the predictive power of existing models. Thongkam et al. [23], for example, enhanced the Support Vector Machine (SVM) model by augmenting it with outlier filtering and oversampling methods. Similarly, Khan et al. [24] proposed a hybrid data mining method based on interference techniques and fuzzy decision trees to enhance the prediction success of an existing crisp classification model. In an earlier study, Pendharkar et al. [25] selected variables for their models via association analysis before rating the importance of several clinical factors as cancer predictors. Using ANN, data envelopment analysis, and discriminant analysis, the authors demonstrated that the prediction accuracy of the latter two components could be improved by utilizing a larger training sample. Zupan et al. [26], on the other hand, developed a prostate cancer survival model based on classification methods. Churilov et al. [27] subsequently improved upon this approach by clustering patients into risk groups based on demographics (age and race), size of the tumors, test results (e.g., prostate cancer-specific antigen concentration in the blood), and pathology scores. Kate and Nadig [28] proposed survival prediction models for each breast cancer stage. They employed several machine learning algorithms such as LR, DT, and Naïve Bayes to compare the prediction powers of the algorithms used for each stage with that obtained for the entire dataset. As expected, survival rates were much lower for patients whose cancer has progressed to more advanced stages. The main advantage of this work is that the model variables were based on the cancer stage rather than survival time. More recently, Simsek et al. [29] employed LR and ANN to identify the critical survival predictors that lose or gain importance over time. The overarching goal of their study was; to guide the medical practitioners as to how much attention should be paid to which demographic/clinical factor and when.

1.3. The contribution of our study

The studies in the literature discussed in the preceding sections focused on predicting cancer survival rates or determining the most critical factors for survival while overlooking the conditional and probabilistic interrelationships among these factors. These vital shortcomings have motivated the present study to adopt a comprehensive data analytics methodology that employs; (a) a wrapper based variable selection method to cherry-pick the most important features/predictors and eliminate the ones that do not contribute to the predictive power of the model, (b) probabilistic-based supervised machine learning model, Tree-augmented Bayesian belief network (TAN), to uncover the hidden, conditional inter-relations among these features as well as to calculate the survival posterior probability of a given patient, and (c) well-known data balancing algorithms to fix the imbalance issue. Specifically, to select the most relevant factors for cancer survival, the purely data-driven variable selection method (GA) is utilized to eliminate the noisy variables. Moreover, by adopting two sampling approaches, SMOTE and RUS, the aim is to increase the sensitivity of the TAN models (i.e., increase the likelihood of detecting patients that will survive less than five years). Lastly, the variable set identified through GA is incorporated into TAN models, allowing the hidden conditional, probabilistic dependencies among the cancer factors to be explored. Finally, the contribution of each variable to the model outcomes is studied via What-if Sensitivity Analysis (SA). As the proposed methodology is not specific to breast cancer, it can be tailored to establish complex interdependent conditional relations among risk factors for other cancer types, thus assisting with timely treatment decisions.

The remainder of the manuscript is structured as follows. Section 2 describes the dataset, data cleaning methodology, variable

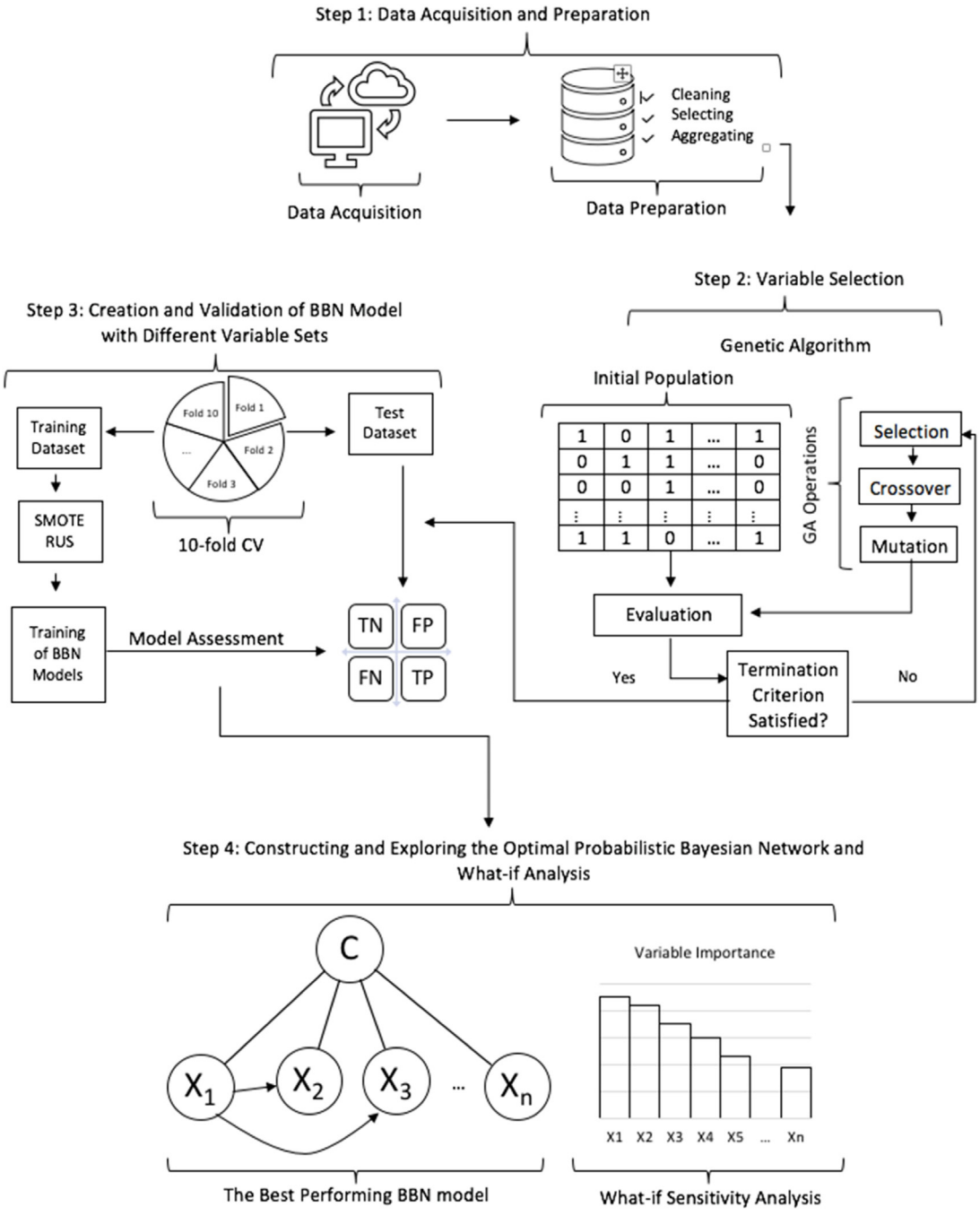

Fig. 1. Proposed Bayesian belief network (BNN)-based data analytics methodology.

selection process, sampling methods, and the predictive models utilized in this work. The results and insights obtained are explained in Section 3. Finally, in Section 4, the main conclusions and suggestions for future research are presented.

2. Research methodology

The study proposes a four-step framework for exploring the significant variables affecting the 5-year survival by uncovering the probabilistic relations among all the features. As shown in Fig. 1, The steps include (1) preprocessing, (2) deploying a feature selection algorithm, (3) generating TAN to uncover probabilistic, conditional interrelationship among variables, (4) conducting sensitivity analysis by using the TAN model. In Step 1, the data has been prepared for further analysis with the methods described in the Cross-Industry Standard Process for Data Mining (CRISP-DM). In Step 2, GA is used to select the significant features. In Step 3, the predictor variables chosen in Step 2 are deployed into the TAN model. The predictive performance of the model is evaluated using 10-fold cross-validation, where the train datasets are balanced using two sampling techniques: SMOTE and RUS. In Step 4, the performance of the TAN models is compared, and the best performing TAN model is used to obtain the conditional probabilistic dependency among the predictor variables. These relations are further investigated by conducting a what-if analysis. Lastly, a decision support tool that can be adopted by practitioners without having any background in machine learning, statistics, or optimization is developed as a proof of concept. Each step is explained in detail in the following sections.

Table 1
Explanation and data structure of the variables.

Variable code	Variable	Description	Variable type
MAR_STAT	Marital Status	Marital status at the time of diagnosis	Nominal
SEX	Sex	Sex of the patient	Nominal
AGE_DX	Age	Age of the patient at the time of diagnosis	Numeric
SEQ_NUM	Sequence number	Number and sequence of all reportable tumors over the lifetime of the patient	Numeric
PRIMSITE	Primary site	The site in which the primary tumor originated	Nominal
HISTO3V	Histologic type	Microscopic composition of the tumor	Nominal
BEHO3V	Behavior code	Malignancy level of the tumor	Nominal
GRADE	Grade	The factor that represents how fast the cancer may grow and spread.	Ordinal
EOD10_SZ	Tumor size	Largest dimension of the tumor in millimeters	Numeric
EOD10_EX	Extension	Farthest documented spread of the tumor away from the originated site	Nominal
EOD10_ND	Lymph node involvement	Chain of lymph nodes involved with tumor	Numeric
EOD10_PN	Regional nodes positive	Number of lymph nodes that contains tumor cells	Numeric
TUMOR_1V	Tumor marker	Prognostic indicators for breast cancer	Nominal
SURGPRIF	Surgical Procedure	Surgical procedure at the primary site	Nominal
RAC_RECA	Race	Race of the patient	Nominal
HST_STGA	Historic stage	Stage of the cancer	Ordinal
ERSTATUS	Estrogen-receptor-positive	The factor that represents whether tumor cells receive signals from estrogen that could promote their growth	Nominal
PRSTATUS	Progesterone-receptor-positive	The factor that represents whether tumor cells receive signals from progesterone that could promote their growth	Nominal

2.1. Data and data preprocessing

The feature-rich dataset utilized in this project was obtained from the Surveillance, Epidemiology, and End Results (SEER) program of the US National Cancer Institute. For the purpose of advanced research, SEER collects detailed data on cancer patients, such as demographics, primary tumor site, morphology and stage at diagnosis, the first course of treatment, and the follow-ups for vital status.

Particularly, the breast cancer dataset that the present study utilizes comes from 1975–2015 in the SEER Program, and it includes 133 variables and over 100,000 observations. In order to clean the dataset, all missing and irrelevant variables, such as patient ID, registry ID, and other non-cancer-related variables, have been removed. Moreover, we excluded the patients who died of any causes but breast cancer, leaving us a dataset with over 50,000 observations and 18 variables. The description of these variables is given in Table 1.

The purpose of this research is to model the conditional impact of cancer-related factors on 5-year breast cancer survivability. Thus, a binary target variable is created using the existing two variables, namely the survival month and the vital status (patient's current status after the follow-up date, either alive or dead). A patient with a survival month of 60 or more is considered as alive and dead otherwise. Besides, alive patients with a survival month of less than 60 have been censored as their exact survival time is unknown.

The creation of the binary variable causes an uneven number of survivors (accounting for ∼80% of the entire dataset) and deaths (accounting for the remaining ∼20%), thereby being an imbalanced dataset. When a dataset is imbalanced, machine learning algorithms can produce a high predictive accuracy for the majority class (survivors in our case) while producing relatively low accuracy for the minority class (deaths in our case) because the overall accuracy contribution of the minority class is negligible. SMOTE and RUS are utilized to balance the accuracy for both the minority and majority classes and achieve improved AUPR results.

2.2. Variable selection

Variable selection is an essential part of the model-building process that allows researchers to model the underlying relationship between the dependent and independent variables with fewer variables. The variable selection process brings certain advantages. For instance, it renders the model to a simpler one, thereby enhancing its interpretability [30]. It also eliminates unnecessary variables from the dataset, which regularizes the model, thus improving its performance [31]. Moreover, performing variable selection leads to dimensionality reduction and decreases the computation expense—the time required to train the model. In the preliminary analysis stage of our study, we have employed several both filter- and wrapper-based variable selection techniques such as *Simulated Annealing* (SA), *Genetic Algorithm* (GA), and *Relaxed Lasso (*RL), to cherry-pick, potentially, the most important variables. Our findings show that the set of variables selected by GA (wrapper-based variable selection method) yields the best performing TAN model with a desirable level of parsimony compared to the other variable selection algorithms employed. Therefore, we chose not to include the RL, SA for the sake of the readability of our paper. The detailed description of GA, which outperformed all the other competing ones, is provided in Section 2.2.1

2.2.1. Genetic algorithm

GA is a metaheuristic optimization technique inspired by the natural selection process and creates a set of sufficiently high-quality solutions to continuous and discrete functions [32]. GAs have been used in many research fields for various reasons [33,34], including variable selection [16,35,36] and prediction [37].

The operationalization of the algorithm starts with an initial set of potential solutions, called a population. The initial population goes through "selection", "crossover", and "mutation" operations to select the fittest individual solutions. A population is composed of chromosomes, and chromosomes are bit strings that are consisted of genes. Each gene is encoded as binary values "0" and "1", "0", indicating the absence of a predictor, and "1", indicating otherwise. A fitness score indicates the strength of an

individual for survival. Until the convergence of fitness scores, selection, crossover, and mutation processes repeat.

In this paper, GA has been used as a variable selection tool, and the pseudocode of the proposed/customized genetic algorithm is given in Algorithm 1. For the algorithm, we set the initial population size to 50, use a random forest model as the fitness function, and tune the parameters with k-fold cross-validation. The parameter γ in Algorithm 1 represents the rate of mutation, which is set to 0.3, allowing the algorithm to make diversification on individuals. Choosing a high mutation probability, such as 0.05 and above, might increase the genetic algorithm's convergence time to an optimal solution. However, the reason we set the mutation probability high was for the algorithm to escape the local optima, even though it increases the execution time.

The rate of elitism (η) indicates the number of fittest chromosomes that move from the current generation to the next generation. After experimenting with several levels, we set the rate of elitism to 3. The elites are pushed to the next generation without going through the crossover process. K indicates the number of generations, which is initialized as 150, and τ is used to define the multiplication of population size and the rate of elitism. We set the crossover between the pairs of chromosomes to 0.8, such that the algorithm can move to the next generations with the fittest chromosomes.

Algorithm 1 Proposed Genetic Algorithm

Generate: Initial population $\Theta = \{\theta_1, \theta_2, \theta_3, ..., \theta_N\}$
Initialize: γ, η, K and $\tau = \eta * N$
for $k \leftarrow 0$ to K **do**
  Tune the parameters of $f(\theta_i^{(k)})$ with k-fold cross validation
  for $i \leftarrow 1$ to 3 **do**
    $trainDataset = originalDataset[sample == i,]$
    Select m variables, $m \in \{2, 3, 4, 5\}$
    $Q \leftarrow 500$
    for $w \leftarrow 1$ to Q **do**
      Bootstrap a new dataset, b, with size n, $b \in trainDataset$
      Grow Random Forest Tree T_b
      Choose best split point in m variables, $m \in \{2, 3, 4, 5\}$
      Split the node into two daughter nodes
    end for
    for all specified parameters **do**
      Produce a set of trees $\{T_b\}_1^B$
      Predict x with $\widehat{C}_{rf}^B(x) = majorityVote\{\widehat{C}_b(x)\}_1^B$
    end for
  end for
  $\Theta_1 \leftarrow$ best τ chromosomes $\in \Theta$
  $\Theta_2 = \{\theta \in \Theta : \theta \notin \Theta_1\}$, $\delta = \frac{|\Theta_2|}{2}$
  for $j \leftarrow 1$ to δ **do**
    $\{\theta_x, \theta_y \in \Theta : f(\theta_x) > f(\theta_i)$ and $f(\theta_y) > f(\theta_i), \forall \theta_i \in \Theta\}$
    Select $loci$ randomly
    $\theta_a, \theta_b \leftarrow$ Crossover θ_x and θ_y
    $\Theta_2 \leftarrow \theta_a, \theta_b$
  end for
  for $j \leftarrow 1$ to δ **do**
    Mutate $\theta_j \in \Theta_2$ with rate γ
  end for
  Update $\Theta \leftarrow \Theta_1 + \Theta_2$
end for
  return $\theta^* = argmax_{\theta_i^{(k)}} f(\theta_i^{(k)})$

2.3. Data balancing

When one or more response classes are represented less than other classes, then that dataset is imbalanced. This issue is innate in many real-world datasets, causing complexities for machine learning algorithms such as bias towards majority classes and overfitting the training model [38]. To deal with these complexities, several different techniques have been proposed to address these problems [39–41]. We would like to note that we have employed several different balancing algorithms such as SMOTE ADASYN, and RUS [6,16,22,42]. However, we only presented the results from SMOTE and RUS, due to the poor performance obtained through ADASYN.

Among these balancing techniques, RUS is an under-sampling technique, which simply drops some of the instances belonging to the majority class at random to equalize the number of observations for both classes. On the other hand, the SMOTE algorithm over samples the minority class by creating synthetic instances, which forces the machine learning algorithm to expand its decision boundary of the minority class into the majority class region. The creation of the synthetic instances is based on the following algorithm [43]:

1. Select a random sample x_i from the dataset
2. Find its k-nearest neighbors in the feature space and randomly select one of them, e.g., x_j
3. Calculate the Euclidian difference between x_i and x_j and multiply it by a random number drawn from the continuous [0, 1] range.
4. Add this difference to x_i to create a new synthetic instance, x, along the straight line connecting x_i and x_j.

The prediction results of the TAN models obtained using the SMOTE and RUS algorithms are provided in Section 3.1. Also, it should be noted that SMOTE and RUS are applied only on the train data.

2.4. Tree-augmented Bayesian Belief Network

The Tree-augmented Bayesian Belief Networks (TAN) are gaining popularity in data science in recent years. Typically, a Bayesian network is used to encode a joint distribution over a random vector $\boldsymbol{X} = \{X_1, \ldots, X_m\}$ [44]. The network comprises of nodes and arcs, whereby nodes represent the random variables in $\boldsymbol{X}$ and the arcs denote the conditional relationships (dependencies) between them. The equation given below shows the joint probability distribution defined by the TAN over $\boldsymbol{X}$:

$$P(\boldsymbol{X}) = \prod_{j=1}^{m} P\left(X_i | P_\gamma(X_i)\right) \tag{1}$$

where m indicates the number of variables and the set of nodes in the joint probability distribution that is connected to X_i is denoted by $P_\gamma(X_i)$.

The Naïve Bayes (NB) is the simplest form of a Bayesian network as it assumes independence between all nodes, due to which all NB nodes are disconnected. However, as this is rarely the case in practice, Friedman et al. [45] developed a TAN that relaxes this assumption that each node can be connected to one other node while being connected to the target variable. Thus, the TAN can be considered a TAN network if the following conditions are met:

$$P_\gamma(X_i) = \begin{cases} \{C, X_{\delta(i)}\}, & if\, \delta(i) > 0. \\ \{C\}, & if\, \delta(i) = 0. \end{cases} \tag{2}$$

where the output variable is C, δ is the tree function, and $P_\gamma(C) = \varnothing$. An optimal tree structure is obtained by maximizing the log-likelihood of δ via Chow-Liu's algorithm for a one-dependence estimator [46]. In the present study, the parameters of the algorithm, such as Laplace's correction, are tuned/optimized using the k-fold cross-validation technique, which would, in turn, optimize the TAN network, as discussed in Section 3.3.

2.5. Performance evaluation

Performance of the machine learning algorithms depends on the prediction power of the model, which can be calculated using a confusion matrix that represents four populated cells: True

Table 2
Selected features by Genetic Algorithm.

MAR_STAT	SEX	AGE_DX	SEQ_NUM	PRIMSITE	HISTO3V	BEHO3V	GRADE	EOD10_SZ	EOD10_EX	EOD10_ND	EOD10_PN	TUMOR_1V	SURGPRIF	RAC_RECA	HST_STGA	ERSTATUS	PRSTATUS
✓		✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	

Positives (TP)-number of correctly predicted positive observations; True Negatives (TN)-number of correctly predicted negative observations; False Positives (FP)-number of incorrectly predicted negative observations; and False Negatives (FN)-number of incorrectly predicted positive observations.

The most prominent performance evaluation metrics in the literature derived from the confusion matrix are overall accuracy and per class accuracies (i.e., sensitivity and specificity).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

Additionally, an AUPR (area under the precision–recall curve) is an assessment measurement representing the performance of a classification model at varying classification thresholds with respect to the precision and the recall. Specifically, AUPR measures the two-dimensional area under the precision–recall curve and provides a measure of performance across all possible thresholds [47]. In this study, we use the AUPR metric to evaluate the performance of the TAN models, as the other metrics can be manipulated by changing the decision probability threshold. Yet, they are still provided to illustrate the accuracy of the models.

3. Experimental results and sensitivity analysis

3.1. Feature selection and classification results

As mentioned in the earlier section of the manuscript, the customized GA along with several other well-known feature selection models such as Relaxed Lasso (RL) and Simulated Annealing (SA) are employed to perform variable selection, in the preliminary analysis stage. However, for the sake of simplicity and conciseness, here we present the model that outperformed the other variable selection models, which in this case was GA. The variable set of the best performing gene (i.e., the variable set that leads to the highest accuracy) is illustrated in Table 2. It can be observed that the variable set that is selected by GA does not include the *SEX*, *PRIMSITE*, *HISTO3V*, and *PRSTATUS* variables.

The cross-validated results of GA-TAN models with SMOTE and RUS balancing techniques are compared to the imbalanced benchmark model in Table 3. Each cell contains the average and standard deviation of the corresponding predictive performance measure, respectively. Results show that no single model outperforms the others in all measures. While accuracy and specificity decline with SMOTE and RUS balancing methods, sensitivity improves significantly for all models compared to the imbalanced dataset. Since there is a 20:80 ratio between positive (diseased patients) and negative (alive patients) classes, the result shows the adverse impact of imbalanced datasets on accuracy and specificity while predicting the positive class.

Moreover, AUPR results with RUS (0.608 (0.017)) are slightly better compared to AUPR results with SMOTE (0.589 (0.018)) and imbalanced dataset (0.601 (0.019)). Therefore, we decide to further analyze the TAN model where the RUS sampling technique was used, as it provided the highest AUPR value.

In addition to the proposed TAN model, we employed several well-known machine learning models to compare our model's prediction power against well-known benchmark models. Here, it should be noted that we are not competing the proposed TAN model against other models that we have used, as the central premise of our study is to uncover the hidden, conditional relations among the potential predictors of cancer. These additional (benchmark) models include Artificial Neural Networks (ANN), Random Forests (RF), and Gradient Boosting (XGB). As these models are popular, well-known machine learning algorithms, here we believe that providing a detailed description of these models would be unnecessary for the sake of conciseness and it would not serve our overarching goal in the proposed study. The results that were obtained from these models are presented in Tables 4, 5, and 6. Even though it does not carry much practical meaning, it should be noted that the proposed TAN model outperformed these well-known models in terms of AUPR, which is the primary evaluation metric that we employ in the current study. This justifies that we are not sacrificing from the model's performance for the sake of uncovering conditional relations among the predictors.

3.2. Tree-augmented Bayesian Belief Network Model

Fig. 2 illustrates the developed TAN graphical model. The figure exhibits the relationships in terms of conditional probabilities/dependencies among the fourteen significant predictors selected by the GA methodology. The constructed TAN model provides a holistic view and insights about the interdependencies among the predictor variables and the outcome variable. For practical reasons, one of the ten-folds had to be picked as an exemplary model among the six models. Therefore, the TAN is generated using the fifth fold from the GA-TAN model with RUS since the AUPR is the highest in this model. The TAN structure can assist practitioners in making better decisions.

Recall that when interpreting the TAN structure, the direction of the arrows provides information about dependencies and indirect relations among the predictors and the outcome variable. An arrow from a predictor (parent) to another (child) indicates that the relation between the child node and the outcome variable is dependent on the value of the parent node [42]. For example, *AGE_DX* (age of the patient) and *EOD10_SZ* (tumor size) are the two variables that have only the outcome variable as a parent node. This means that the contribution of the two variables predicting the outcome variable does not change with the values of other predictor variables. In addition, there are five predictors, namely *EOD10_EX* (extension), *EOD10_ND* (node involvement), *EOD10_PN* (nodes positive), *SURGPRIF* (surgical procedure), and *GRADE* that are not parents of any other predictors, yet children of other variables. This means that while their contribution to predicting the outcome variable depends on another variable, they do not impact another predictor.

Furthermore, the TAN structure depicts the effects of *EOD10_EX*, *EOD10_ND*, *EOD10_PN*, and *SURGPRIF* are dependent on *HST_STGA* (stage of breast cancer) that is dependent on *BEHO3V* (malignancy level). This interrelation is consistent with the literature, which demonstrates that the malignancy level of the tumor, in fact, impacts the stage of breast cancer [48], which in turn affects the prognostic factors such as extension, node involvement, lymph nodes positive as well as surgery type that is required for the annihilation of cancer cells [49]. Our results also reveal the relation between *GRADE*, *ERSTATUS* (estrogen receptor-positive), *TUMOR_1V* (tumor marker), *BEHO3V*, *RAC_RECA* (race),

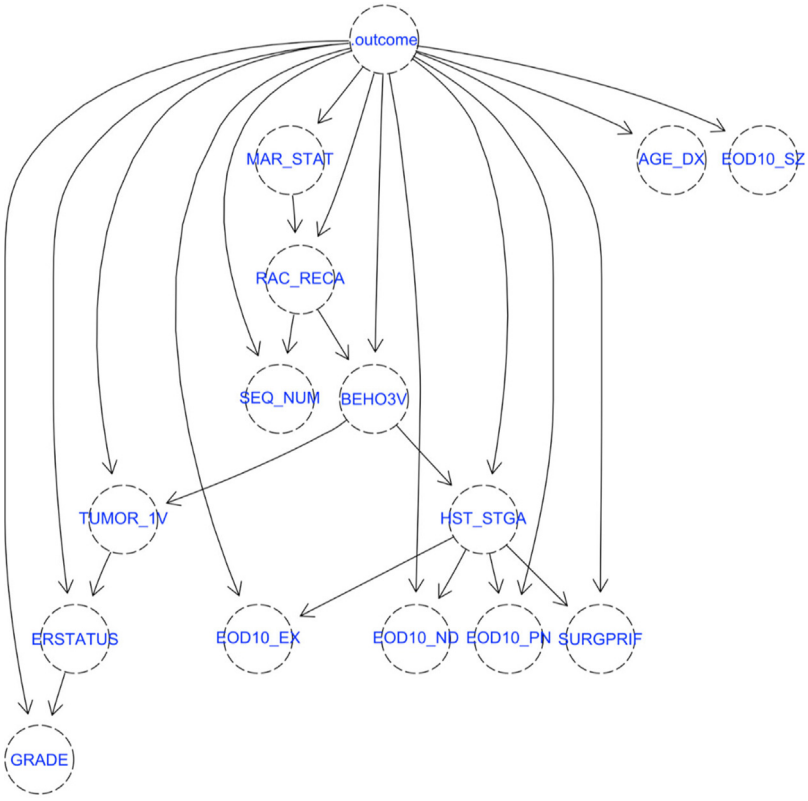

Fig. 2. TAN structure of breast cancer patient 5-year survival.

Table 3
Ten-fold cross-validation performance results for the proposed TAN model.

Model	Number of variables	Balancing technique	Accuracy	Sensitivity	Specificity	AUPR
GA-TAN	14	SMOTE	0.780 (0.004)	0.785 (0.019)	0.779 (0.006)	0.589 (0.018)
TAN	18	SMOTE	0.777 (0.005)	0.789 (0.015)	0.776 (0.008)	0.582 (0.015)
GA-TAN	**14**	**RUS**	**0.813 (0.002)**	**0.751 (0.015)**	**0.822 (0.004)**	**0.608 (0.017)**
TAN	18	RUS	0.813 (0.005)	0.750 (0.018)	0.823 (0.006)	0.601 (0.017)
GA-TAN	14	NONE	0.897 (0.004)	0.479 (0.021)	0.960 (0.003)	0.601 (0.019)
TAN	18	NONE	0.896 (0.005)	0.483 (0.019)	0.956 (0.003)	0.601 (0.018)

Table 4
Ten-fold cross-validation performance results for ANN model (Benchmark Model 1).

Model	Number of variables	Balancing technique	AUPR
GA-ANN	14	SMOTE	0.556 (0.018)
ANN	18	SMOTE	0.519 (0.029)
GA-ANN	**14**	**RUS**	**0.569 (0.023)**
ANN	18	RUS	0.532 (0.035)
GA-ANN	14	NONE	0.561 (0.019)
ANN	18	NONE	0.539 (0.025)

Table 5
Ten-fold cross-validation performance results for RF model (Benchmark Model 2).

Model	Number of variables	Balancing technique	AUPR
GA-RF	14	SMOTE	0.566 (0.035)
RF	18	SMOTE	0.571 (0.034)
GA-RF	**14**	**RUS**	**0.606 (0.03)**
RF	18	RUS	0.601 (0.049)
GA-RF	14	NONE	0.590 (0.031)
RF	18	NONE	0.560 (0.059)

Table 6
Ten-fold cross-validation performance results for XGB model (Benchmark Model 3).

Model	Number of variables	Balancing technique	AUPR
GA-XGB	14	SMOTE	0.542 (0.05)
XGB	18	SMOTE	0.554 (0.043)
GA-XGB	14	RUS	0.572 (0.034)
XGB	**18**	**RUS**	**0.574 (0.036)**
GA-XGB	14	NONE	0.566 (0.027)
XGB	18	NONE	0.570 (0.033)

and *MAR_STAT* (marital status). Tumor markers are proteins that are produced by normal and cancer cells but higher amounts by cancer cells [48]. On the other hand, a cancer cell is estrogen receptor-positive if the cancer cells have receptors that promote the growth of cancer cells [50]. Finally, the grade of cancer indicates how slow or fast cancer cells are proliferating: In grade I, cancer cells are similar to normal cells and not growing rapidly; in grade II, cancer cells are not like normal cells and grow faster than normal cells; and in grade III, grade cells are abnormal and grow very aggressively [51].

The TAN structure in Fig. 2 demonstrates that the contribution of the grade of breast cancer in predicting the survival of the patient depends on estrogen receptor, tumor marker, malignancy level, as well as race and marital status of the patient. This means that the malignancy of cancer cells impacts the over-production of proteins and cancer-related substances, which are considered

Table 7
Cancer survival probability for race according to marital status.

Race	Marriage status	Posterior survival probability
White	Married	48%
Black		41%
White	Not Married	31%
Black		30%

as tumor markers [52]. Moreover, although the relation between tumor marker and ER-status is not clearly defined in the cancer literature, Dunnwald et al. (2007) expressed that resulting tumor gene mutations due to the malignancy level of cancer can also affect the estrogen receptor status of the cancer cells, which is consistent with the TAN structure. Besides, consistent with [4,50], Fig. 2 proves that ER-status significantly impacts the grade of breast cancer. Finally, according to The TAN structure, the impact of race on breast cancer survival needs to be studied per marital status, which we analyze in Section 3.3 below.

3.3. Scenario analysis of conditional dependency structure with DSS tool

As discussed in Section 3.2, the GA-TAN model reveals the conditional dependency structure among the variables. This suggests that the individual impact of each variable on 5-year breast cancer survivability can change depending on the status of its parent variable and thus needs to be interpreted accordingly. In order to enable practitioners to conduct scenario analysis, a decision support (DS) tool that incorporates the GA-TAN model is developed. Such a tool can be used by practitioners to not only quantify the strength and the impact of each variable on breast cancer survival but also analyze their conditional impacts.

For example, according to the Bayesian network created by the TAN model, the race of patients plays an important role in 5-year survivability. However, its impact depends on the marital status of the patients. In order to explore this relationship, we conduct scenario analysis via the DS tool. Table 7 shows the posterior survival probabilities obtained by using the tool. One can observe from the table that the chances of 5 or more years of survival for unmarried black and white patients, with all else being the same, are 30% and 31%, respectively.

However, even though being married positively affects the survival change according to the provided posterior survival probabilities, its impact on the white race survivability is significantly higher than the black race. In other words, the survival chance of a married white patient is 17% higher than an unmarried white patient, while this rate is only 11% higher for the black counterparts.

This interesting interrelation between race and marital status has also been pronounced in the existing literature. For example, Martinez et al. (2016) indicated that the mortality rate for unmarried non-Hispanic patients is up to 24% higher than married non-Hispanic patients. However, this change between unmarried and married patients drops down to 6% for Hispanic patients. This signifies the dependency between race and marital status on breast cancer survival. In another study, Zhai et al. (2019) investigated the effect of marital status based on race. They concluded that married patients with breast cancer had a better prognosis than divorced or widow counterparts while noting that race affects the correlation between marital status and breast cancer survival.

The lower mortality rate for married patients, compared to unmarried patients, might exist because married patients are able to make easier decisions on receiving breast-conserving surgeries (BCS) – that damages the body cosmetic – which enables physicians to remove cancerous cells, thus improving the prognosis [53]. However, its correlation with race can stem from the fact that black cancer patients have a greater degree of difficulty with the treatment decision compared to white patients [54]. Thus they are less likely to make a surgical decision, although there is professional consensus that BCS is a good treatment option for early-stage breast cancer patients [55].

Similar to the race and marital status discussion above, different scenario analyses can be conducted by physicians to better understand the complex nature of breast cancer as well as the conditional dependencies among the cancer-related factors. With that said, the DS tool can also be utilized by patients to learn their 5+ year survival chance and can make their decisions accordingly. A screenshot of the proposed tool is given in Fig. 3, and the tool can be accessed via the following link: https://rsearch.shinyapps.io/breast_cancer/.

Lastly, to verify the associations and relations between variables, we performed a what-if analysis by removing each predictor from the TAN model while keeping the other thirteen predictors and analyzing the network structure. With this analysis, we tried to answer the question of what would have happened if the GA feature selection algorithm has not selected a given feature. Also, the *what-if* analysis has been performed to verify the sparsity of the TAN structure (See appendix, Fig. A1). Even though TAN is sensitive to variable addition and removal [56], the direction of the relations and associations do not change in the acyclic graph unless *BEHOV3*, *RAC_RECA*, or *HST_STGA* is removed from the TAN model (Figs. A1-e, m, n, respectively). The most likely reason for the change in the TAN structure when one of the three variables is removed from the model is that these variables are in the center of the TAN structure.

3.4. Sensitivity analysis

The most important advantage of employing the TAN model is to investigate parameter uncertainty and sensitivity. Sensitivity analysis provides evidence about the amount of information each variable suggests in establishing the model, namely, their importance. This can be calculated using entropy function [57] as follows: $I = H(Q) - H(Q|F)$ where $H(Q)$ and $H(Q|F)$ represent the entropy of Q before and after findings, respectively. Results in Fig. 4 show that the historic stage is the most significant variable for predicting the survival of breast cancer patients. As the stage of breast cancer increases, the likelihood of survival decreases. This is consistent with Simsek et al. (2020), which states that 5-year survival of stage-0 and -1 breast cancer are 100%, while the 5-year survival is 21% for stage-4 breast cancer patients. It is shown that the sequence number is the least important predictor in predicting the 5-year survival.

4. Summary, conclusion, and future research directions

The main objectives of this research were (1) to reveal the complex interrelations among the breast cancer-related factors, (2) to find the most important variables contributing to cancer survival. To that end, in this study, we proposed a holistic analytical framework consisting of multiple steps. The framework has been developed using the SEER dataset that spans the period of 1975–2015. As the number of patients who survived over five years is excessively more than the patients who died in 5 years, RUS and SMOTE have been employed to balance the dataset. In order to select the most important variables that relate to breast cancer survival, a genetic algorithm is employed. Afterward, the selected variables are deployed into Tree-augmented Bayesian Belief Network to find complex inter-relation among the cancer-related factors. In order to validate the associations and relations

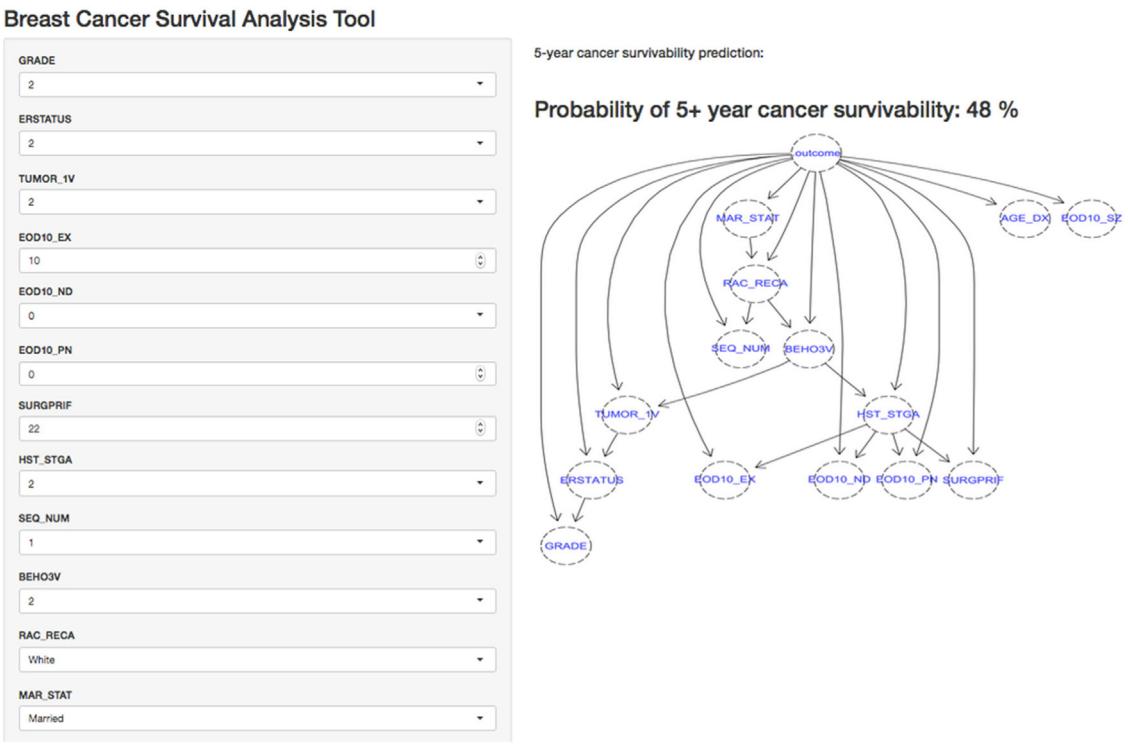

Fig. 3. A Screenshot of the proposed DSS tool.

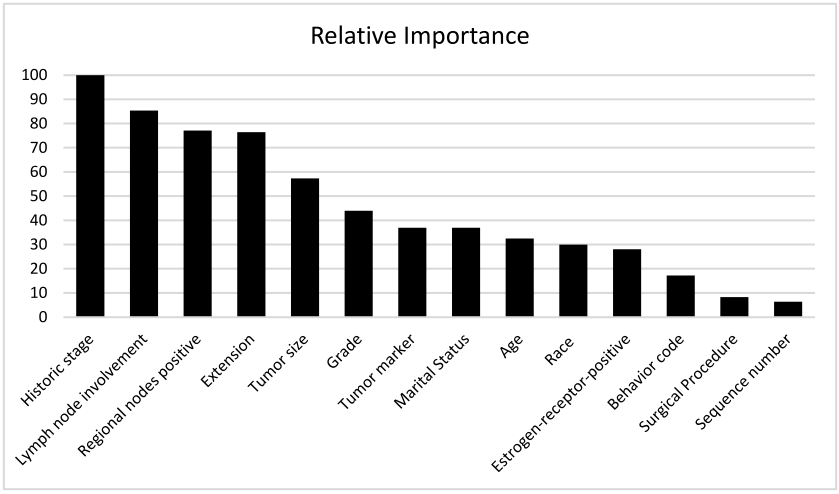

Fig. 4. Variable importance via sensitivity analysis.

between variables, we performed a *what-if* analysis by removing one predictor at a time from the TAN model to analyze the impacts of variable elimination on the TAN tree structure. Finally, we conducted a sensitivity analysis to find the relative importance of each variable in cancer survival.

The findings of the study indicate that the GA-TAN model was able to predict cancer survival with 0.608, 0.813, 0.751, and 0.822 for AUPR, Accuracy, Sensitivity, and Specificity values, respectively, when the data is balanced with RUS. The employment of the balancing algorithms increases the sensitivity of the model in differentiating patients who are likely to die in 5 years. Moreover, an interesting dependency structure among the cancer-related factors (i.e., 14 variables) are revealed, which can help medical practitioners to have a better understanding and estimation of the course of the disease, and consequently, can make better, more efficient treatment plans. The study findings reveal that stage of cancer is relatively the most important factor affecting breast cancer survival while the sequence number variable (i.e., count of the reportable tumors) has the least importance.

Some of the perceived limitations of the study can be listed as follows. First of all, although the dataset covers a long-time window, years from 1975–2015, there are some important variables that have been recorded in the SEER dataset after 2010, and have not been included in this study due to their absence in the cases recorded prior to 2010. These variables include HER2, which represents the presence of a gene type that can play a role in the development of breast cancer; Dx-Bone/Brain/Liver/Lung, which represents where the metastasis happens in stage 4; and AJCC −7 T/N/M, which gives detailed information about the tumor, node, and metastasis of the cancer tumor. Second, other than the TAN method, Markov blanket [58] could have been used as a

structural learning algorithm, for which, however, more detailed patient-level data would be required.

In summary, this paper proposes a probabilistic framework that provides the interrelations among the important variables in the tree-augmented network along with the conditional survival probabilities. The proposed methodology not only provides the prediction for breast cancer survival but also reveals the hidden nonlinear interrelations among the variables in different scenarios. The methodology can also be integrated into a decision support tool to help medical practitioners by augmenting their knowledge to make better treatment decisions.

CRediT authorship contribution statement

Asli Z. Dag: Conceptualization, Design and implementation of the underlying analytics study, Writing of the current manuscript. **Zumrut Akcam:** Conceptualization, Design and implementation of the underlying analytics study, Writing of the current manuscript. **Eyyub Kibis:** Conceptualization, Design and implementation of the underlying analytics study, Writing of the current manuscript. **Serhat Simsek:** Conceptualization, Design and implementation of the underlying analytics study, Writing of the current manuscript. **Dursun Delen:** Conceptualization, Design and implementation of the underlying analytics study, Writing of the current manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] SEER, Colorectal cancer - cancer stat facts, SEER cancer stat facts color, Cancer (2020) https://seer.cancer.gov/statfacts/html/breast.html.

[2] N.J. Bundred, Prognostic and predictive factors in breast cancer, Cancer Treat. Rev. 27 (2001) 137–142, http://dx.doi.org/10.1053/ctrv.2000.0207.

[3] R.S. Rampaul, S.E. Pinder, C.W. Elston, I.O. Ellis, Prognostic and predictive factors in primary breast cancer and their role in patient management: The nottingham breast team, Eur. J. Surg. Oncol. 27 (2001) 229–238, http://dx.doi.org/10.1053/EJSO.2001.1114.

[4] L.K. Dunnwald, M.A. Rossing, C.I. Li, Hormone receptor status, tumor characteristics, and prognosis: A prospective cohort of breast cancer patients, Breast Cancer Res. 9 (2007) R6, http://dx.doi.org/10.1186/bcr1639.

[5] J.F. Desforges, W.L. McGuire, G.M. Clark, Prognostic factors and treatment decisions in axillary-node-negative breast cancer, N. Engl. J. Med. 326 (1992) 1756–1761, http://dx.doi.org/10.1056/NEJM199206253262607.

[6] E. Kibis, E. Buyuktahtakin, A. Dag, Data analytics approaches for breast cancer survivability: comparison of data mining methods, in: Proc. 2017 Ind. Syst. Eng. Conf, 2017.

[7] M. Nasir, C. South-Winter, S. Ragothaman, A. Dag, A comparative data analytic approach to construct a risk trade-off for cardiac patients' readmissions, Ind. Manag. Data Syst. 119 (2019) 189–209, http://dx.doi.org/10.1108/IMDS-12-2017-0579.

[8] N. Arya, S. Saha, Multi-modal advanced deep learning architectures for breast cancer survival prediction[Formula presented], Knowl.-Based Syst. 221 (2021) http://dx.doi.org/10.1016/J.KNOSYS.2021.106965.

[9] G. Magna, P. Casti, S.V. Jayaraman, M. Salmeri, A. Mencattini, E. Martinelli, C. Di Natale, Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system, Knowl.-Based Syst. 101 (2016) 60–70, http://dx.doi.org/10.1016/J.KNOSYS.2016.02.019.

[10] L. Xie, L. Zhang, T. Hu, H. Huang, Z. Yi, Neural networks model based on an automated multi-scale method for mammogram classification, Knowl.-Based Syst. 208 (2020) http://dx.doi.org/10.1016/J.KNOSYS.2020.106465.

[11] S. Gunasundari, S. Janakiraman, S. Meenambal, Velocity bounded boolean particle swarm optimization for improved feature selection in liver and kidney disease diagnosis, Expert Syst. Appl. 56 (2016) 28–47, http://dx.doi.org/10.1016/J.ESWA.2016.02.042.

[12] S. Gupta, A. Sharma, Data mining classification techniques applied for breast cancer diagnosis and prognosis, Indian J. Comput. Sci. Eng. 2 (2011) 188–195.

[13] Y.U. Ryu, R. Chandrasekaran, V.S. Jacob, Breast cancer prediction using the isotonic separation technique, European J. Oper. Res. 181 (2007) 842–854, http://dx.doi.org/10.1016/J.EJOR.2006.06.031.

[14] D. West, P. Mangiameli, R. Rampal, V. West, Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application, European J. Oper. Res. 162 (2005) 532–551, http://dx.doi.org/10.1016/J.EJOR.2003.10.013.

[15] H.M. Zolbanin, D. Delen, A. Hassan Zadeh, Predicting overall survivability in comorbidity of cancers: A data mining approach, Decis. Support Syst. 74 (2015) 150–161, http://dx.doi.org/10.1016/J.DSS.2015.04.003.

[16] S. Simsek, U. Kursuncu, E. Kibis, M. AnisAbdellatif, A. Dag, A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival, Expert Syst. Appl. 139 (2020) http://dx.doi.org/10.1016/j.eswa.2019.112863.

[17] A. Li, J. Walling, S. Ahn, Y. Kotliarov, Q. Su, M. Quezado, J.C. Oberholtzer, J. Park, J.C. Zenklusen, H.A. Fine, Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes, Cancer Res. 69 (2009) 2091–2099, http://dx.doi.org/10.1158/0008-5472.CAN-08-2100.

[18] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, L.M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, Nature 403 (2000) 503–511, http://dx.doi.org/10.1038/35000501.

[19] D.G. Beer, S.L.R. Kardia, C.-C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M.G. Taylor, M.D. Iannettoni, M.B. Orringer, S. Hanash, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (2002) 816–824, http://dx.doi.org/10.1038/nm733.

[20] J. Lapointe, C. Li, J.P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A.M. DeMarzo, R. Tibshirani, D. Botstein, P.O. Brown, J.D. Brooks, J.R. Pollack, Gene expression profiling identifies clinically relevant subtypes of prostate cancer, Proc. Natl. Acad. Sci. 101 (2004) 811–816, http://dx.doi.org/10.1073/pnas.0304146101.

[21] M. Lundin, J. Lundin, H.B. Burke, S. Toikkanen, L. Pylkkänen, H. Joensuu, Artificial neural networks applied to survival prediction in breast cancer, Oncology 57 (1999) 281–286, http://dx.doi.org/10.1159/000012061.

[22] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2005) 113–127, http://dx.doi.org/10.1016/J.ARTMED.2004.07.002.

[23] J. Thongkam, G. Xu, Y. Zhang, F. Huang, Toward breast cancer survivability prediction models through improving training space, Expert Syst. Appl. 36 (2009) 12200–12209, http://dx.doi.org/10.1016/J.ESWA.2009.04.067.

[24] M.U. Muhammad Umer Khan, J.P. Jong Pill Choi, H. Hyunjung Shin, M. Minkoo Kim, Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare, in: 2008 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc, IEEE, 2008, pp. 5148–5151, http://dx.doi.org/10.1109/IEMBS.2008.4650373.

[25] P.C. Pendharkar, J.A. Rodger, G.J. Yaverbaum, N. Herman, M. Benner, Association, statistical, mathematical and neural approaches for mining breast cancer patterns, Expert Syst. Appl. 17 (1999) 223–232, http://dx.doi.org/10.1016/S0957-4174(99)00036-6.

[26] B. Zupan, J. Demsar, M.W. Kattan, J.R. Beck, I. Bratko, Machine learning for survival analysis: a case study on recurrence of prostate cancer, Artif. Intell. Med. 20 (2000) 59–75.

[27] L. Churilov, A.M. Bagirov, D. Schwartz, K. Smith, M. Dally, Improving risk grouping rules for prostate cancer patients with optimization, in: 37th Annu. Hawaii Int. Conf. Syst. Sci. 2004. Proc, IEEE, 2004, p. 9.

[28] R.J. Kate, R. Nadig, Stage-specific predictive models for breast cancer survivability, Int. J. Med. Inform. 97 (2017) 304–311, http://dx.doi.org/10.1016/j.ijmedinf.2016.11.001.

[29] S. Simsek, U. Kursuncu, E. Kibis, M. AnisAbdellatif, A. Dag, A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival, Expert Syst. Appl. 139 (2020) http://dx.doi.org/10.1016/j.eswa.2019.112863.

[30] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer New York, New York, NY, 2013, http://dx.doi.org/10.1007/978-1-4614-7138-7.

[31] J. Eskelinen, Comparison of variable selection techniques for data envelopment analysis in a retail bank, European J. Oper. Res. 259 (2017) 778–788, http://dx.doi.org/10.1016/j.ejor.2016.11.009.

[32] D.E. Goldberg, J.H. Holland, Genetic algorithms and machine learning, Mach. Learn. 3 (1988) 95–99, http://dx.doi.org/10.1023/A:1022602019183.

[33] O. Alp, E. Erkut, Z. Drezner, An efficient genetic algorithm for the p-median problem, Ann. Oper. Res. (2003) 21–42, http://dx.doi.org/10.1023/A:1026130003508.

[34] F. Pezzella, G. Morganti, G. Ciaschetti, A genetic algorithm for the flexible job-shop scheduling problem, Comput. Oper. Res. 35 (2008) 3202–3212, http://dx.doi.org/10.1016/j.cor.2007.02.014.

[35] S. Simsek, T. Tiahrt, A. Dag, Stratifying no-show patients into multiple risk groups via a holistic data analytics-based framework, Decis. Support Syst. (2020) 113269, http://dx.doi.org/10.1016/j.dss.2020.113269.
[36] S. Simsek, A. Dag, T. Tiahrt, A. Oztekin, A Bayesian belief network-based probabilistic mechanism to determine patient no-show risk categories, Omega (U. K.) (2020) http://dx.doi.org/10.1016/j.omega.2020.102296.
[37] G. Sermpinis, C. Stasinakis, K. Theofilatos, A. Karathanasopoulos, Modeling forecasting and trading the EUR exchange rates with hybrid rolling genetic algorithms - support vector regression forecast combinations, European J. Oper. Res. 247 (2015) 831–846, http://dx.doi.org/10.1016/j.ejor.2015.06.052.
[38] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Prog. Artif. Intell. 5 (2016) 221–232, http://dx.doi.org/10.1007/s13748-016-0094-0.
[39] C.X. Ling, V.S. Sheng, Cost-sensitive learning and the class imbalance problem, 2009, https://www.semanticscholar.org/paper/Cost-Sensitive-Learning-and-the-Class-Imbalance-Ling-Sheng/9c4a953ed2cfc999eef0901d43097f9d2933005c (Accessed August 6 2018).
[40] N.V. Chawla, Data mining for imbalanced datasets: An overview, Springer-Verlag, New York, 2005, pp. 853–867, http://dx.doi.org/10.1007/0-387-25465-X_40.
[41] X. Guo, Y. Yin, C. Dong, G. Yang, G. Zhou, On the class imbalance problem, in: 2008 Fourth Int. Conf. Nat. Comput., IEEE, 2008, pp. 192–201, http://dx.doi.org/10.1109/ICNC.2008.871.
[42] A. Dag, K. Topuz, A. Oztekin, S. Bulur, F.M. Megahed, A probabilistic data-driven framework for scoring the preoperative recipient-donor heart transplant survival, Decis. Support Syst. 86 (2016) 1–12, http://dx.doi.org/10.1016/j.dss.2016.02.007.
[43] N.V. Chawla, A. Lazarevic, L.O. Hall, K.W. Bowyer, Smoteboost: improving prediction of the minority class in boosting, Springer, Berlin, Heidelberg, 2003, pp. 107–119, http://dx.doi.org/10.1007/978-3-540-39804-2_12.
[44] H.E. Kyburg, Probabilistic reasoning in intelligent systems: Networks of plausible inference by judea pearl, J. Philos. 88 (1991) 434–437, http://dx.doi.org/10.5840/jphil199188844.
[45] N. Friedman, D. Geiger, G. Provan, P. Langley, P. Smyth, Bayesian Network Classifiers *, Kluwer Academic Publishers, 1997.
[46] C. Chow, C. Liu, Approximating discrete probability distributions with dependence trees, IEEE Trans. Inf. Theory. 14 (1968) 462–467, http://dx.doi.org/10.1109/TIT.1968.1054142.
[47] J. Davis, M. Goadrich, The relationship between precision–recall and ROC curves, ACM Int. Conf. Proceeding Ser. (2006) 233–240, http://dx.doi.org/10.1145/1143844.1143874.
[48] M. Cianfrocca, L.J. Goldstein, Prognostic and predictive factors in early-stage breast cancer, 2004.
[49] K. Inoue, M. Makuuchi, T. Takayama, G. Torzilli, J. Yamamoto, K. Shimada, T. Kosuge, S. Yamasaki, M. Konishi, T. Kinoshita, S. Miyagawa, S. Kawasaki, Long-term survival and prognostic factors in the surgical treatment of mass-forming type cholangiocarcinoma, Surgery 127 (2000) 498–505, http://dx.doi.org/10.1067/msy.2000.104673.
[50] A. Pourzand, M.B.A. Fakhree, S. Hashemzadeh, M. Halimi, A. Daryani, Hormone receptor status in breast cancer and its relation to age and other prognostic factors, Breast Cancer Basic Clin. Res. 5 (2011) 87–92, http://dx.doi.org/10.4137/BCBCR.S7199.
[51] D.E. Henson, L. Ries, L.S. Freedman, M. Carriaga, Relationship among outcome, stage of disease, and histologic grade for 22, 616 cases of breast cancer. the basis for a prognostic index, Cancer 68 (1991) 2142–2149, http://dx.doi.org/10.1002/1097-0142(19911115)68:10{\T1\textless}2142::AID-CNCR2820681010{\T1\textgreater}3.0.CO;2-D.
[52] M.J. Duffy, Predictive markers in breast and other cancers: A review, Clin. Chem. 51 (2005) 494–503, http://dx.doi.org/10.1373/clinchem.2004.046227.
[53] Z. Zhai, F. Zhang, Y. Zheng, L. Zhou, T. Tian, S. Lin, Y. Deng, P. Xu, Q. Hao, N. Li, P. Yang, H. Li, Z. Dai, Effects of marital status on breast cancer survival by age, race, and hormone receptor status: A population-based study, Cancer Med. (2019) http://dx.doi.org/10.1002/cam4.2352.
[54] A.B. Nattinger, Variation in the choice of breast-conserving surgery or mastectomy: Patient or physician decision making? J. Clin. Oncol. (2005) http://dx.doi.org/10.1200/JCO.2005.04.913.
[55] S.J. Katz, P.M. Lantz, N.K. Janz, A. Fagerlin, R. Schwanz, L. Liu, D. Deapen, B. Salem, I. Lakhani, M. Morrow, Patient involvement in surgery treatment decisions for breast cancer, J. Clin. Oncol. (2005) http://dx.doi.org/10.1200/JCO.2005.06.217.
[56] M. Velikova, P.J.F. Lucas, M. Samulski, N. Karssemeijer, On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks, Artif. Intell. Med. 57 (2013) 73–86, http://dx.doi.org/10.1016/j.artmed.2012.12.004.
[57] E. Archer, I.M. Park, J.W. Pillow, BayesIan and quasi-Bayesian estimators for mutual information from discrete data, Entropy (2013) http://dx.doi.org/10.3390/e15051738.
[58] Y. Che, S. Hong, D. Zhang, L. Zhang, Learning markov blanket Bayesian network for big data in mapreduce, in: Proc. - 2016 IEEE 28th Int. Conf. Tools with Artif. Intell., ICTAI 2016, 2017, http://dx.doi.org/10.1109/ICTAI.2016.0138.