Iranian Journal of Medical Informatics (2015) Vol 4, Issue 3

Modeling of Breast Cancer Risk Using Bayesian Networks

مدلسازی احتمال ابتلا به سرطان پستان با استفاده از شبکه بیزین

Marjan Ghazi Saeidi, Mostafa Langarizadeh, Mehrshad Mokhtaran, Zeynab Hasani

Abstract — *Breast cancer is the most common cancer in women at different stages of life affects about 10 percent of them. This Cancer is the second cause of death in women and the most common cause of death among women 45-55 years old. So the find a model to predict the likelihood of breast cancer based on patient past history and other risk factors is helpful. The purpose of this study was Building a Bayesian network model to calculate the risk of breast cancer. This research is developmental. Model and the conditional probability table that obtained from the Clementine 12.0. Breast cancer detection accuracy evaluates and results showed that the accuracy of the model was 96.22 percent*[1].

Keywords — breast cancer, Bayesian networks, modeling.

۱. چکیده

یکی از شایعترین انواع سرطان در زنان سرطان پستان است که در مراحل مختلف زندگی حدود ۱۰ درصد از آنان را تحت تاثیر قرار میدهد. این سرطان دومین علت مرگ در جمعیت زنان میباشد و شایعترین علت مرگ در بین زنان ۵۵-۴۵ سال است. بنابراین یافتن مدلی برای پیش بینی احتمال ابتلا به سرطان پستان بر مبنای سوابق گذشته بیماران و سایر عوامل خطرساز مفید است. هدف این مطالعه ساخت یک مدل شبکه های بیزین برای محاسبه احتمال ابتلا به سرطان پستان است، این پژوهش توسعه ای کاربردی است. مدل و جدول احتمال شرطی که با استفاده از نرم افزار Clementine 12.0 بدست آمده و از نظر صحت تشخیص سرطان پستان بررسی شد، نتایج نشان داد که صحت تشخیص مدل ۹۶/۲۲ درصد بود.

کلمات کلیدی

سرطان پستان، شبکه های بیزین، مدلسازی.

۲. مقدمه

سرطانها بیماریهای مزمنی هستند که علتهای متعددی دارند و در دههای اخیر در بسیاری از جوامع میزان زیادی از مرگها را به خود اختصاص دادهاند.این بیماری یکی از مسائل مهم و اصلی بهداشت و درمان در تمام دنیا محسوب می شود(۱). در ایران بیماری سرطان بعد از بیماریهای قلبی عروقی و حوادث، سومین علت مرگ و میر محسوب میشود (۲). سرطان پستان یکی از شایعترین انواع سرطان در زنان است(۳) و علت پنجم مرگ ناشی از سرطان کشورهای کمتر توسعه یافته و بیشتر توسعه یافتهی جهان است(4). از عوامل خطر دخیل در پیدایش این سرطان میتوان به سن، جنسیت، نژاد، دین، بیماری خوشخیم قبلی در پستان، سابقه سرطان قبلی در فرد، عوامل مربوط به بارداری و هورمونها، سابقه فامیلی سرطانهای پستان و تخمدان، برخورد با اشعه یونیزان و عوامل محیطی (۵, ۶)، تعداد زایمان، مجموع ماه های شیردهی، مدت استفاده از قرص پیشگیری از بارداری، قد و وزن اشاره کرد(۳). میزان مرگ در اثر سرطان پستان با افزایش سن افزایش مییابد؛ به گونهای که ۵۴ درصد مرگهای سرطان پستان در زنان ۶۵ ساله و بیشتر رخ می-دهد(۷, ۸). در کشورهای کمتر توسعه یافته، متوسط سن زنان مبتلا به سرطان پستان حدود ۱۰ سال کمتر از کشورهای بیشتر توسعه یافته تشخیص داده شده است(۴). این سرطان شایعترین بدخیمی در میان زنان ایرانی و کانون اصلی توجهات در کشور ایران میباشد و در سالهای اخیر، میزان شیوع بیماری روند رو به رشدی داشته است(۳). برخی از محققان نتیجه گرفته اند که سرطان پستان حدود یک دهه پیش از آن در چند جمعیت از جمله ایران رخ می دهد(۴). با توجه به پایین بودن سن ابتلا به سرطان پستان در ایران، تشخیص زود هنگام آن یکی از چالشهای اساسی در جهت سلامت جامعه است. از اینرو تشخیص زود هنگام بیماری موجبات درمان بموقع این بیماری را فراهم خواهد کرد(۸).

[1] M. Ghazi Saeidi is with Faculty of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran- Iran (email: ghazimar@tums.ac.ir).
M. Langarizadeh is with School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran (email: langarizadeh.m@iums.ac.ir).
M. Mokhtaran is a PhD student of Medical Informatics, Faculty of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran (email: mehrshad.mokhtaran@gmail.com).
Z. Hasani is MSc in Medical Informatics, Faculty of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran (Corresponding Author; email: z-hasani@razi.tums.ac.ir).

تاکنون مطالعات متعددی در زمینه تشخیص سرطان پستان انجام شده است، که از جمله میتوان به مطالعاتی مانند سیستم خبره برای تشخیص سرطان پستان با استفاده از دادههای پیش پردازش و شبکههای بیزین (۹)، مروری بر ۷ الگوریتم دادهکاوی در پیشبینی بقا، تشخیص و عود بیماران مبتلا به سرطان پستان (۱۰)، پیشبینی پیشآگهی سرطان پستان با ادغام دادههای بالینی و میکروواریها با شبکه های بیزین (۱۱)، پیادهسازی سیستمهای خودکار تشخیص سرطان پستان (۱۲)، تشخیص سرطان پستان با استفاده از شبکههای بیزین (۱۳)، شبکههای بیزین مشتق شده از آسیب شناسی پستان همرخدادی (۱۴) اشاره کرد.

هدف اصلی از این مطالعه ساختن مدل شبکه بیزین برای محاسبه احتمال ابتلا به سرطان پستان با توجه به عوامل خطر است.که باید ابتدا عوامل خطر مهم را شناسایی کرد و سپس مدل پیشنهادی احتمال ابتلا به سرطان پستان را طراحی کرد.

۳.مواد و روش ها

جهت بدست آوردن اینکه فردی با چه احتمالی مبتلا به سرطان پستان میشود از شبکه بیزین استفاده میشود که شامل ۷ گام است: ۱) پیش پردازش، ۲) انتخاب خروجی و ورودی شبکه، ۳) ساخت مجموعه آموزش و آزمون، ۴) یادگیری شبکه بیزین، ۵) تعیین اهمیت نسبی هر متغیر، ۶) ارزیابی مدل تشخیص سرطان پستان، ۷) اعلام نتایج. در گام اول روی دادههای بدست آمده از پرونده بیماران پیش پردازشهایی انجام میگیرد و پروندههایی که تعداد زیادی فیلد خالی دارند حذف میگردد، در گام دوم انتخاب فیلد هدف به عنوان خروجی است، که در اینجا فیلد Diagnosis انتخاب میشود و شامل دو حالت بیمار یا سالم(۰ و ۱) است، و فیلدهای دیگر را به عنوان ورودی شبکه انتخاب میکنیم. در گام سوم دادههای آموزش و آزمایش تعیین میشوند. در گام چهارم شبکه بیزین آموزش داده میشود مدل شبکه که در آن ارتباط بین هر گره و جدول احتمالات شرطی هر گره مشخص میشود. در گام پنجم اهمیت نسبی هرمتغیر نسبت به متغیر هدف بدست میآید. در گام شش و هفت مدل ارزیابی شده و نتایج حاصل اعلام میگردد.

أ. مجموعه داده

دادههای مورد استفاده در این پژوهش از انستیتو کانسر ایران در قالب فایل اکسل دریافت شد، و از نحوه نمونهگیری اطلاعاتی دریافت نشد. فهرست فیلدهای فایل اکسل سال تولد، وضعیت تاهل، شغل، تحصیلات، سن اولین حاملگی، تعداد زایمان، مجموع ماههای شیردهی، سابقه استفاده از قرص پیشگیری از بارداری، مدت استفاده از قرصهای پیشگیری از بارداری، وجود تومور پستان، وجود تومور تخمدان، وجود تومور رحم، وجود تومور کولن، جراحی پستان، رادیولوژی، سابقه خانوادگی سرطان، نوع سرطان، سابقه خانوادگی سرطان پستان، قد، وزن، سن و شاخص جرم بدن بود. رکوردهایی که بیشتر از ۲ فیلد اطلاعاتی خالی داشتند حذف شدند و در نهایت از ۱۹۶۰ رکورد برای محاسبه احتمال ابتلا به سرطان پستان استفاده گردید. فیلدهای پایگاه داده در جدول ۱ مشخص شده است.

جدول ۱: فیلدهای پایگاه داده

ردیف	معانی متغیرهای انتخابی برای مدل شبکه بیزین	متغیرهای انتخابی برای مدل شبکه بیزین
۱	سال تولد	Year.Birth
۲	وضعیت تاهل	Marriage
۳	شغل	Job
۴	تحصیلات	Education
۵	سن اولین حاملگی	First. Pregnant
۶	تعداد زایمان	Parity
۷	مجموع ماه های شیردهی	BreastFeeding
۸	سابقه استفاده از قرص پیشگیری از بارداری	History.Pill
۹	مدت استفاده از قرص پیشگیری از بارداری	Term.Use.Pill
۱۰	وجود تومور پستان	Breast.tumor
۱۱	وجود تومور تخمدان	Ovary.Tumor
۱۲	وجود تومور رحم	Uterus.Tumor
۱۳	وجود تومور کولون	Colon.Tumor
۱۴	جراحی پستان	Breast.Surgery
۱۵	رادیوتراپی	Radiotherapy
۱۶	سابقه خانوادگی سرطان	Family.History.Cancer
۱۷	نوع سرطان	Kind.Cancer.Family
۱۸	سابقه سرطان پستان درخانواده	Family.Histoey.Breast.Cancer
۱۹	قد	Length
۲۰	وزن	Weight
۲۱	سن	Age
۲۲	شاخص جرم بدن	BMI
۲۳	سن اولین ازدواج	First.Marriage

ب. پیش پردازش

در این گام منظور از پیش پردازش حذف رکوردهایی است که فیلد خالی آنها بیشتر از ۲ فیلد است. برای حذف این رکوردها با پزشک مربوطه مشورت صورت پذیرفته است.

ج.انتخاب ورودی و خروجی شبکه

در گام دوم در نرمافزار Clementine 12.0 برای ایجاد شبکه بعد از فراخواندن مجموعه داده باید خروجی و ورودی شبکه را تعیین کرد، که در این نرمافزار تنها میتوان یک خروجی داشت. در این پژوهش خروجی متغیر Diagnosis است که دارای دو مقدار صفر و یک است، و مابقی متغیرها به عنوان ورودی انتخاب میشود

[2] Conditional probability table (CPT)
[3] Excel
[4] Null

د. ساخت مجموعه آموزش و آزمون

در این گام مجموعه داده آموزش و آزمون تعیین میگردد. برای تعیین این دو مجموعه داده، ۷۰ درصد دادهها به عنوان مجموعه آموزش و ۳۰ درصد به عنوان مجموعه آزمون بصورت تصادفی انتخاب میگردد. در این پژوهش نیز مجموعههای آموزش و آزمون به همین صورت در نظر گرفته شدهاند.

ه. یادگیری شبکه بیزین

شبکه با ۷۰ درصد از دادهها شروع به یادگیری روابط میکند، و باتوجه به فیلد خروجی Diagnosis مدل شبکه بیزین برای محاسبه احتمال ابتلا به سرطان پستان را بدست میآورد. در این مدل روابط بین فیلدهای داده که همان عوامل خطر ابتلا به سرطان پستان هستند به صورت تصویری نمایش داده میشود. هر عامل دارای جدول احتمال شرطی است.

و. تعیین اهمیت نسبی هر متغیر

در مدل بدست آمده ممکن هر متغیر با متغیر هدف در ارتباط باشد و یا با واسطه با متغیر هدف در ارتباط باشد و یا با متغیر هدف ارتباطی نداشته باشد.

ز. معیارهای ارزیابی مدل

در این بخش از سه معیار ارزیابی زیر جهت بررسی اثربخشی مدل پیشنهادی محاسبه احتمال ابتلا به سرطان پستان استفاده میکنیم. این معیارها با استفاده از نتایج اعمال این روش بر روی مجموعه آزمون و آزمایش محاسبه میشود. در رابطه های ۱ تا ۳ این سه معیار معرفی شده است. در این روابط TP[5] به معنای تعداد نمونهای است که سالم بودهاند و به درستی تشخیص داده شدهاند. FP[6] نشانگر تعداد نمونههای است که بیمار بودهاند و به اشتباه سالم تشخیص داده شدهاند. FN[7] نشانگر تعداد نمونههایی است که سالم بودهاند ولی به اشتباه بیمار تشخیص داده شدهاند. TN[8] حاوی تعداد نمونههایی است که بیمار بودهاند و به درستی بیمار تشخیص داده شدهاند.

$$\text{حساسیت} = \frac{TP}{TP+FN} \quad (۱)$$

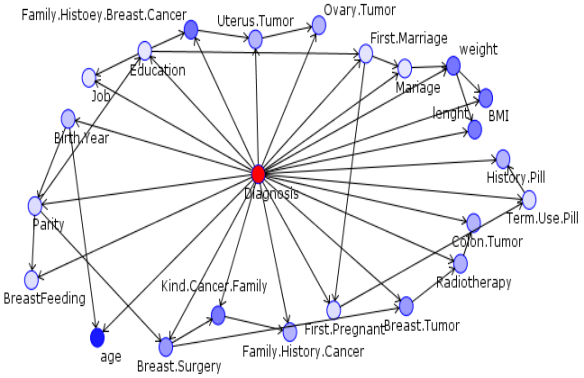
$$\text{ویژگی} = \frac{TN}{FP+TN} \quad (۲)$$

$$\text{صحت} = \frac{TP+TN}{TP+FN+FP+TN} \quad (۳)$$

و همچنین از نمودار Gains برای اندازه گیری اثر بخشی یک مدل طبقه بندی و محاسبه نسبت بین نتایج به دست آمده با مدل استفاده میشود. این نمودار کمک بصری برای ارزیابی عملکرد مدل طبقه بندی است.

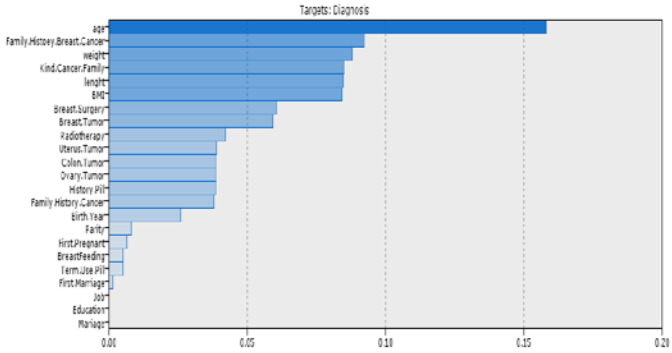
۴. یافته ها

شبکه با ۷۰ درصد از دادهها شروع به یادگیری روابط میکند، و باتوجه به فیلد خروجی Diagnosis مدل شبکه بیزین برای محاسبه احتمال ابتلا به سرطان پستان را بدست میآورد. در این مدل روابط بین فیلدهای داده که همان عوامل خطر ابتلا به سرطان پستان هستند به صورت تصویری نمایش داده میشود. هر عامل دارای جدول احتمال شرطی است. شکل ۱ مدل شبکه بیزین بدست آمده برای محاسبه احتمال ابتلا به سرطان پستان است.



شکل ۱: مدل شبکه بیزین برای محاسبه احتمال ابتلا به سرطان پستان

در این پژوهش و براساس دادههایی که از انستیتو کانسر ایران دریافت شد، مهمترین عامل در ابتلا به سرطان پستان با اهمیت نسبی ۰/۱۵۸۲ متغیر سن است، و عوامل شغل، تحصیلات و وضعیت تاهل دارای اهمیت نسبی صفر هستند. شکل ۲ اهمیت نسبی هر متغیر را نمایش میدهد.



شکل ۲: اهمیت نسبی عوامل خطر نسبت به متغیرDiagnosis

برای بدست آوردن صحت، ویژگی و حساسیت از خروجی مدل محاسبه احتمال ابتلا به سرطان پستان در شکل ۳ استفاده میشود. نتایج محاسبه سه معیار یاد شده برای آموزش مدل در جدول ۲ و برای آزمایش مدل در جدول ۳ است.

Comparing $B-Diagnosis with Diagnosis

'Partition'	1_Training		2_Testing	
Correct	1,278	94.67%	586	96.22%
Wrong	72	5.33%	23	3.78%
Total	1,350		609	

Coincidence Matrix for $B-Diagnosis (rows show actuals)

'Partition' = 1_Training	0	1
0	670	21
1	51	608
'Partition' = 2_Testing	**0**	**1**
0	309	5
1	18	277

Performance Evaluation

'Partition' = 1_Training	
0	0.596
1	0.683
'Partition' = 2_Testing	
0	0.606
1	0.707

شکل ۳- خروجی مدل شبکه بیزین محاسبه احتمال ابتلا به سرطان پستان

[5] True positive
[6] False positive
[7] False negative
[8] True negative

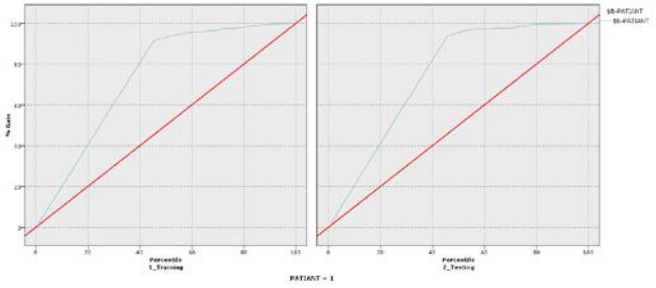
جدول ۲: محاسبه حساسیت، ویژگی و صحت برای آموزش مدل شبکه بیزین محاسبه احتمال ابتلا به سرطان پستان

حساسیت محاسبه شده برای آموزش مدل = $\frac{608}{608+51} * ۱۰۰ =$ ۹۶/۲۶
ویژگی محاسبه شده برای آموزش مدل = $\frac{670}{670+21} * ۱۰۰ =$ ۹۶/۹۶
صحت محاسبه شده برای آموزش مدل = $\frac{608+670}{608+670+51+21} * ۱۰۰ =$ ۹۴/۶۶

جدول ۳: محاسبه حساسیت، ویژگی و صحت برای آزمایش مدل شبکه بیزین محاسبه احتمال ابتلا به سرطان پستان

حساسیت محاسبه شده برای آزمایش مدل = $\frac{277}{277+18} * ۱۰۰ =$ ۹۳/۸۹
ویژگی محاسبه شده برای آزمایش مدل = $\frac{309}{309+5} * ۱۰۰ =$ ۹۸/۴۰
صحت محاسبه شده برای آزمایش مدل = $\frac{277+309}{277\ 309\ 18\ 5} * ۱۰۰ =$ ۹۶/۲۲

نمودار Gains مربوط به مدل شبکه بیزین محاسبه احتمال ابتلا به سرطان پستان در این پژوهش در دو حالت آموزش و آزمایش در شکل ۴ آمده است که مقادیر حساسیت، ویژگی و صحت بدست آمده را تایید میکند.



شکل ۴: نمودار gains مدل شبکه بیزین تشخیص سرطان پستان

۵. بحث

در مطالعه فلاحی و همکاران از پایگاه داده ویسکانسین استفاده شده است، که حاوی ۶۹۹ رکورد است، که ۲۴۱ رکور مربوط به سرطان پستان بدخیم و ۴۵۸ رکورد مربوط به سرطان پستان خوشخیم است و هر رکود ۹ ویژگی وجود دارد، که شبکه بیزینی با استفاده از این اطلاعات بدست آوردند و دارای کارایی ۹۷/۲۸ درصد بود و شبکه باور بیزی مبتنی بر این اطلاعات دارای کارایی ۹۷/۴۲ درصد بود. بهترین نتیجه شبکه برای دادههایی بود که دادهها متعادل تر بودند و دارای کارایی ۹۸/۱۵ درصد شد.(۹). در مطالعه گورت و همکاران از یک پایگاه داده استفاده شده بود که از داده های DNA بیماران برای تشخیص استفاده میکنند، این مطالعه نشان داد که ترکیب مؤثر متغیرهای باینری و میکروآری افزایش عملکرد را در پی دارد . همچنین شاخص بهتر آگهی ترکیبی از متغیرهای باینری و تعداد کمی از ژن شکل میگیرد (۱۱). در مطالعه کروز و همکاران از پایگاه داده پاتولوژی و اطلاعات سن استفاده شده است که ۷ شبکه بیزین طراحی و ارزیابی شدند (۱۳). در مطالعه ماسکری و همکاران به این نتیجه رسیدند که فعل و انفعالات پاتولوژی مشتق شده از مجموعه داده های سازگار با عمل آسیب شناسی می باشد(۱۴).

در این مطالعه از اطلاعات پرونده مراجعین به استیتو کانسر ایران استفاده شد که شامل اطلاعات دموگرافیک، هورمونی و سابقه بیماریهای دیگر در فرد و خانواده او بود. در مطالعه حاضر ماهیت پارامترهای پایگاه داده با مطالعات بررسی شده متفاوت بود با استفاده از این اطلاعات شبکه بیزینی طراحی شد که اهمیت نسبی هر متغیر نسبت به متغیر هدف بدست آمد. متغیر سن با بیشترین اهمیت ۰/۱۵۸۲ بود. صحت، ویژگی و حساسیت مدل پیشنهادی که با دادههای آزمایش ساخته شده به ترتیب ۹۶/۲۲ درصد ، ۹۸/۴۰ درصد و ۹۳/۸۹ درصد بود.

۶. نتیجه گیری

در مطالعات بررسی شده مطالعه مشابهی دیده نشد، زیرا دادهها بومی است بنابراین نتایج متفاوت است. شبکه بیزین یکی از ابزار مهم و کاربردی هوش مصنوعی می باشد. شبکههای بیزین، گرافی با تعدادی گره هستند که با یالهایی در ارتباط هستند. شبکههای بیزین جهت تشخیص و پیشبینی بیماریها استفاده میشوند. در این مطالعه شبکههای بیزین احتمال ابتلا به سرطان پستان را پیشبینی میکند. در مطالعه حاضر شبکه بیزین با استفاده از عوامل خطر تعیین شده احتمال ابتلا به سرطان پستان به صورت قابل قبول محاسبه میکند. با در نظر گرفتن حساسیت و ویژگی در زمان مناسب و .با تهیه داده های کاملتر و با نمونه گیری مطابقت بیشتر با جامعه می توان امید داشت از این نوع سامانه ها در حوزه سلامت استفاده شود. در صورت مناسب بودن داده می توان از این سیستم استفاده کرد و پیشنهاد میشود در مطالعات بعدی خروجی این شبکه جهت اطمینان با داده های بیماران ارزیابی مجدد گردد.

[9] Gevaert
[10] Microarray
[11] Cruz
[12] Maskery

۷. تقدیر و تشکر

با تشکر از دکتر عباس شیخ طاهری، دکتر مجتبی رجب پور و انستیتو کانسر ایران که در هرچه بهترشدن این مطالعه یاریام نمودهاند.

REFERENCES

[1] Yavari P, Abadi A, Mehrabi Y. Mortality and changing epidemiological trends in Iran during 1979-2001. HAKIM. 2003;6(3):7-14.

[2] Kolahdoozan S, Sadjadi A, Radmehr A, Khademi H. Five Common Cancers in Iran. Arch Iran. Med 2010;13(2):143-6.

[3] American Cancer 2014; Available from: http://www.cancer.org.

[4] Ghiasvand R, Adami H-O, Harirchi I, Akrami R, Zendehdel K. Higher incidence of premenopausal breast cancer in less developed countries; myth or truth? BMC cancer. 2014;14(1):343.

[5] Blancas I, Garcia-Puche J, Bermejo B, Hanrahan E, Monteagudo C, Martínez-Agulló A, et al. Low number of examined lymph nodes in node-negative breast cancer patients is an adverse prognostic factor. Annals of oncology. 2006;17(11):1644-9.

[6] Costanza M, Chen W. Factors that modify breast cancer risk in women. Up to Date. 2012; Available from: : http://www.uptodate.com/contents/factors-thatmodify- breast-cancer-risk-in-women.

[7] Kelsey JL, Gammon MD. The epidemiology of breast cancer. CA: a cancer journal for clinicians. 1991; 41(3): 146-165.

[8] Lodish H. Molecular cell biology: Macmillan; 2008.

[9] Fallahi A, Jafari S. An expert system for detection of breast cancer using data preprocessing and Bayesian network. Int J Adv Sci Technol. 2011;34:65-70.

[10] Ghasem Ahmad L. . Review 7 top data mining algorithms to predict survival and recurrence in patients with breast cancer diagnosis. Iranian Journal of Breast Diseases. 20012;6(1):52-61.

[11] Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics. 2006;22(14):e184-e90.

[12] Übeyli ED. Implementing automated diagnostic systems for breast cancer detection. Expert Systems with Applications. 2007;33(4):1054-62.

[13] Cruz-Ramírez N, Acosta-Mesa HG, Carrillo-Calvet H, Alonso Nava-Fernández L, Barrientos-Martínez RE. Diagnosis of breast cancer using Bayesian networks: A case study. Computers in Biology and Medicine. 2007;37(11):1553-64.

[14] Maskery SM, Hu H, Hooke J, Shriver CD, Liebman MN. A Bayesian derived network of breast pathology co-occurrence. Journal of biomedical informatics. 2008;41(2):242-50.