Original Article

Prediction of Colorectal Cancer Tumor Location Using Data Mining

Mohammad Mahdi Khakshoor[1], Kazem Pourbadakhshan[2,*], Ladan Goshayeshi[3]

[1] MSc, Control group, Department of Electrical Engineering, Quchan University of Advanced Technologies Engineering, Iran
[2] Assistant professor, Control group, Department of Electrical Engineering, Quchan University of Advanced Technologies Engineering, Iran
[3] Assistant professor, Department of Gastroenterology and Hepatology, School of Medicine, Mashhad Uiversity of Medical Sciences, Mashhad, Iran

ABSTRACT

Background:
Colorectal cancer is one of the most common cancers in terms of morbidity and mortality worldwide. A lot of research have been done in this field in Iran and worldwide, which have positive results. The aims of this study were firstly doing a statistical study on colorectal cancer in Mashhad, Iran, and finally predicting the colorectal location of cancer based on the clinical data by using data mining science and decision tree model.

Materials and Methods:
The data of 316 patients with colorectal cancer (including 14 features) were extracted from the archive of Imam Reza Hospital, Mashhad. The instrument used in this research was RapidMiner data mining software that would try to be extract the details of the relevant data by statistical surveys and then would do initial simulations and the use of classification and decision tree method have predicteion the location of cancer.

Results:
Male to female ratio of 56% to 44%, family history of 37%, more young patients, and relatively more distally located cancers (39%) compared with the proximal (35%), and rectum (26%) were the striking findings of this study. The final and most important stage of research models were presented, which was able to predict the location of the cancerous tumor with 80% accuracy.

Conclusion:
Similarities with global statistics, such as the ratio of men to women and family history were observed. But there were also differences with global statistics including the Iran's younger patients and relatively more patients with distal cancers. The efficiency of data mining techniques to predict the location of cancer as well as cost reduction was among the most important results of this study.

Keywords: Colorectal cancer, Location, Data mining, Decision tree, Cost matrix, Predict

please cite this paper as:
Khakshoor M.M, Pourbadakhshan K, Goshayeshi L. Prediction of Colorectal Cancer Tumor Location Using Data Mining. *Govaresh* 2017;22:154-163.

Corresponding author:
Kazem Pourbadakhshan , Phd of control
Control group, Department of Electrical Engineering,
Quchan University of Advanced Technologies
Engineering, Quchan, Iran
Tel: + 98 51 47344001
Fax: + 98 51 47343001
E-mail: k_pour@yahoo.com

Received: 30 Apr. 2017
Edited: 02 Aug. 2017
Accepted: 03 Aug. 2017

گوارش/ دوره ۲۲، شماره ۳/ پاییز ۱۳۹۶/۱۶۳-۱۵۴

مقاله پژوهشی

پیش بینی ناحیه تومور سرطان کولورکتال با استفاده از داده کاوی

محمدمهدی خاکشورکامه علیا[۱]، کاظم پوربدخشان[۲،*]، لادن گشایشی[۳]

[۱] کارشناسی ارشد، گروه کنترل، دانشکده برق، دانشگاه مهندسی فناوریهای نوین، قوچان، ایران
[۲] استادیار، گروه کنترل، دانشکده برق، دانشگاه مهندسی فناوریهای نوین، قوچان، ایران
[۳] استادیار، بخش گوارش و کبد، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران

چکیده

زمینه و هدف:
سرطان کولورکتال یکی از شایع ترین سرطان ها از نظر ابتلا و مرگ و میر در جهان است. تابحال در این زمینه تحقیقات فراوانی مبتنی بر بحث های آماری و مدلسازی در ایران و جهان صورت گرفته که نتایج مثبتی نیز در پی داشته است. هدف از این پژوهش نیز نخست یک بررسی آماری درمورد سرطان کولورکتال در مشهد- ایران و در نهایت پیش بینی ناحیه بروز سرطان با توجه به داده های بالینی و با استفاده از علم داده کاوی و مدل درخت تصمیم است.

روش بررسی:
داده ها شامل ۳۱۶ بیمار مبتلا به سرطان کولورکتال با ۱۴ ویژگی بوده که از بیمارستان امام رضا (ع) مشهد استخراج شده است. ابزار مورد استفاده این پژوهش نرم افزار داده کاوی RapidMiner است که با استفاده از آن سعی خواهد شد جزئیات داده های مربوطه توسط بررسی های آماری استخراج شود و سپس با انجام شبیه سازی های اولیه و به کارگیری روش دسته بندی و الگوریتم درخت تصمیم بیان دقیق تری از داده ها صورت گرفته و ناحیه بروز سرطان پیش بینی شود.

یافته ها:
نسبت ابتلای مردان به زنان با تناسب ۵۶٪ به ۴۴٪، سابقه فامیلی مبتلایان با نسبت ۳۷٪، جوان بودن سن مبتلایان و آمار نسبتاً زیاد افراد مبتلا به سرطان ناحیه دیستال[۱] (۳۹ درصد) در مقایسه با پروگزیمال[۲] (۳۵ درصد) و رکتوم[۳] (۲۶ درصد) یافته های قابل بحث این پژوهش بودند. در قسمت نهایی و مهم ترین مرحله پژوهش نیز مدل هایی ارائه شد که قادر به پیش بینی محل قرارگیری تومور سرطانی با دقت بالای ۸۰٪ بود.

نتیجه گیری:
شباهت هایی با آمار جهانی همچون نسبت ابتلای مردان به زنان و سابقه فامیلی مشاهده شد. اما تفاوت هایی نیز با آمار جهانی وجود داشت که ازجمله آنها می توان به جوان تر بودن جمعیت مبتلایان ایران و آمار نسبتاً زیاد افراد مبتلا به سرطان ناحیه دیستال اشاره کرد. کارآمدی تکنیک های داده کاوی در پیش بینی ناحیه بروز سرطان و کاهش هزینه ها نیز اساسی ترین نتیجه پژوهش بود.

کلید واژه: سرطان کولورکتال، ناحیه، داده کاوی، درخت تصمیم، ماتریس هزینه، پیش بینی

گوارش/ دوره ۲۲، شماره۳/ پاییز ۱۳۹۶/۱۶۳-۱۵۴

1. Distal
2. Proximal
3. Rectum

***نویسنده مسئول:** کاظم پوربدخشان
گروه کنترل، دانشکده برق- دانشگاه مهندسی فناوریهای نوین قوچان
تلفن: ۰۵۱-۴۷۳۴۴۰۰۱
نمابر: ۰۵۱-۴۷۳۴۳۰۰۱
پست الکترونیک: k_pour@yahoo.com

تاریخ دریافت: ۹۶/۲/۱۰
تاریخ اصلاح نهایی: ۹۶/۵/۱۱
تاریخ پذیرش: ۹۶/۵/۱۲

زمینه و هدف:

آمار جهانی سرطان کولورکتال نشان می دهد که این سرطان از نظر شیوع، سومین سرطان در بین مردان (بعد از سرطان ریه و پروستات) و دومین سرطان در بین زنان (بعد از سرطان پستان) است. همچنین به ترتیب چهارمین و سومین علت مرگ و میر ناشی از سرطان در بین مردان و زنان است. نرخ این سرطان در مردان نسبت به زنان در اکثر نقاط جهان بیشتر و بطور کلی درحدود ۵۵٪ به ۴۵٪ است.(۱) طی جدیدترین مطالعه انجام شده در سال ۲۰۱۶ در آمریکا سرطان کولورکتال با آماری در حدود ۵۰۰۰۰ نفر و نسبت ۸٪، رتبه سوم را هم از نظر شیوع و هم از نظر مرگ و میر در بین زنان و مردان دارد، همچنین آمار سال ۲۰۱۵ نیز

دقیقاً مبین همین تناسب است. جزئیات این آمار بیان می کند که بیشترین نرخ ابتلا برای مردان در سنین ۶۰ تا ۸۰ سال، با تناسب ۵۰٪ و برای زنان در سنین بالاتر از ۸۰ سال، با تناسب ۴۰٪ و کمترین نرخ ابتلا برای مردان و زنان در سنین ۲۰ تا ۴۰ سال، با تناسب ۱.۵٪ است.(۲و۳) بررسی آمار قدیمی تر در سال ۲۰۰۶ نیز تفاوت چندانی را نشان نمی دهد که بیانگر ثبات نرخ این سرطان در سال های اخیر است.(۴) در ایران نیز بررسی آماری در بین سال های ۱۹۹۶ تا ۲۰۰۰ نشان می دهد که مشابه آمار جهانی سرطان کولورکتال به ترتیب سومین و چهارمین سرطان شایع در بین مردان و زنان با نسبت ۵۵٪ به ۴۵٪ بوده است. همچنین میانگین سن افراد بیمار ۵۷ سال بیان شده، اما آمار ابتلای افراد جوان و زیر ۴۰ سال با نسبت ۱۷٪ درمقایسه با آمار جهانی بیشتر بوده که علت آن جوان بودن جمعیت ایران بیان شده است (۵)که البته در (۶) حدود ۴۳٪ افراد مورد مطالعه آن زیر ۵۰ سال سن دارند، جوانی جمعیت مبتلایان در ایران را معلول نقش پررنگ عوامل ژنتیکی می داند. بررسی های جدیدتر نیز نشان می دهد آمار افراد مبتلا بین سال های ۲۰۰۰ تا ۲۰۰۵ در ایران مشابه آمار قبلی است، به طوری که چهارمین و دومین علت سرطان در مردان و زنان با تناسب ۵۶٪ و ۴۴٪ مربوط به سرطان کولورکتال بوده که حدود ۳۰٪ از آنها نیز منجر به مرگ شده است. این آمار نیز به روشنی بیان کننده تعداد نسبتاً زیاد افراد جوان مبتلا در ایران است به طوری که حدود ۴٪ از مبتلایان زیر ۳۰ سال و حدود ۱۳٪ کمتر از ۴۰ سال سن داشته اند.(۷) جدیدترین آمار در بین سال های ۲۰۰۵ تا ۲۰۰۹ هم مطالب گفته شده فوق را تایید می کند به طوری که در آن مطالعه نیز ۲۵٪ مبتلایان ایرانی زیر ۵۰ سال سن داشته اند.(۸) از حیث بررسی مبتلایان با سابقه فامیلی نیز مطالعات انجام شده در ایران مانند (۶و۹) آماری در حدود ۳۵٪ را در این باره نشان می دهد.

از طرفی بروز و شیوع این نوع از سرطان در شمال آمریکا، اروپا و استرالیا بیشترین و در آفریقا و آسیای مرکزی و جنوبی کمترین حد را داشته، اما با توجه به غربالگری مناسب و به موقع در کشورهای توسعه یافته اروپایی و آمریکایی تا حد زیادی این سرطان کنترل شده است که این امر در کشورهای در حال توسعه آفریقایی و آسیایی کمتر به چشم می خورد. قابل ذکر است که غربالگری به موقع می تواند حداقل از ۶۰٪ مرگ و میرهای ناشی از این سرطان جلوگیری کند.(۱و۸)

با توجه به مطالب بالا به نظر می رسد پیش بینی و تشخیص زودهنگام سرطان می تواند به صورت چشمگیری به درمان و بهبود فرد بیمار کمک نماید که یکی از راه های مؤثر آن فرآیند کشف دانش از پایگاه داده (KDD) است. KDD عبارتست از تبدیل داده های سطح پایین به دانش سطح بالا با استفاده از علم داده کاوی. مراحل داده کاوی نیز بطور خلاصه عبارتند از: ۱- استخراج داده و آماده سازی، ۲- مدلسازی و ۳- تفسیر و ارزیابی مدل. داده کاوی روش های گوناگونی دارد، با این حال امروزه اغلب از روش های دسته بندی برای طبقه بندی بیماری ها و سرانجام پیش بینی آنها استفاده می شود که از مهمترین الگوریتم های دسته بندی نیز درخت تصمیم است.(۱۰) در مطالعات پزشکی بسیاری، بخصوص در زمینه سرطان از داده کاوی برای پیش بینی و استخراج مدل استفاده و نتایج بسیار مؤثری هم دریافت کرده اند که از جمله آنها می توان به (۱۱) و (۱۲) اشاره کرد. همچنین (۱۳) یک بررسی جامع بر روی کاربردهای داده کاوی در پزشکی انجام داده است. نرم افزارهای متنوعی نیز برای داده کاوی ساخته شده اند که از جمله پرکاربردترین و ساده ترین آنها نرم افزار داده کاوی RapidMiner، ابزار مورد استفاده این پژوهش است. در بحث سرطان و بخصوص سرطان کولورکتال پژوهش های کمی به چشم می خورد که از این ابزار استفاده کرده باشند، مانند (۱۴و۱۵)که نتایج مطلوبی نیز بدست آوردند. لذا هدف این پژوهش مقایسه و بررسی آمار مبتلایان به سرطان کلورکتال در مشهد و درنهایت ارائه مدل درخت تصمیم برای پیش بینی ناحیه بروز این سرطان با استفاده از علوم داده کاوی و به کارگیری نرم افزار RapidMiner است.

روش بررسی:

داده های این مطالعه از مبتلایان به سرطان کولورکتال مراجعه کننده به بیمارستان امام رضا (ع) مشهد جمع آوری شده که شامل ۳۱۶ بیمار با ۱۴ ویژگی بررسی شده در مورد آن هاست. این ویژگی ها در حوزه های زیر هستند: sex, age, familial history, AmsterdamII, RevisedBethesda, Location که دو ویژگی مهم AmsterdamII و RevisedBethesda نیز با توجه با مطالعات انجام شده در (۱۶) مقدار دهی شده است.

ابزار مورد استفاده این پژوهش نرم افزار داده کاوی RapidMiner است که طبق نظرسنجی سالانه KDnuggest Data Mining / Analytics از افراد حرفه ای کاوش گر داده، در سال ۲۰۱۰ به عنوان محبوب ترین ابزار داده کاوی مورد استفاده ی آن ها شناخته شد. لذا ابتدا سعی خواهد شد جزئیات داده های مربوطه توسط بررسی های آماری استخراج و سپس با انجام شبیه سازی های اولیه و پس از آن به کارگیری روش دسته بندی و الگوریتم درخت تصمیم بیان دقیق تری از داده ها صورت گرفته و امکان بروز سرطان پیش بینی شود. در ادامه توضیحاتی درمورد روش دسته بندی و الگوریتم درخت تصمیم ارائه خواهد شد.

دسته بندی

به الگوریتم های دسته بندی، الگوریتم های «باناظر» گفته می شود، چرا که هر رکورد با برچسبی مشخص شده که هدف یافتن نظم همین برچسب ها بر اساس سایر ویژگی های رکوردها است. در الگوریتم های دسته بندی، مجموعه داده ها به دو مجموعه ی داده، تحت عنوان داده های آموزشی و آزمایشی تقسیم می شوند. توسط مجموعه داده آموزشی، مدل ساخته شده و از مجموعه داده آزمایشی برای تعیین دقت و اعتبارسنجی مدل استفاده می گردد. با این وجود پس از آموزش و در مرحله بخاطرسپاری، دقت داده های آزمایشی کاهش می یابد که برای حل این مشکل مجموعه دیگری تحت عنوان داده ی ارزیابی، با جداکردن قسمتی از داده ی آموزشی، ایجاد می گردد. نتیجه ی بسیار سودمند این کار این است که مدل، مجموعه داده های آموزشی را واقعا یاد می گیرد و دیگر به دنبال حفظ آن نخواهد بود! دسته بندی، الگوریتم های بسیار گوناگون و متفاوتی دارد که ازجمله مهم ترین و پرکاربردترین آن ها درخت تصمیم می باشد که الگوریتم مورد بررسی این پژوهش است.(۱۵و۱۷)

جدول۱: ارزیابی یک درخت با مقادیر فرضی مثبت(P) و منفی(N)

	True P	True N
Pred P	TP	FP
Pred N	FN	TN

درخت تصمیم

یکی از مهمترین خصوصیات این روش، شاخه به شاخه بودن آن بوده که قدرت تفسیر فوق العاده ای را فراهم کرده است. این ویژگی باعث شده که به این الگوریتم ها جعبه سفید گفته شود، برخلاف روش شبکه های عصبی که جعبه سیاه نامیده شده و به هیچ عنوان روند آن قابل تفسیر نیست. کارکرد الگوریتم درخت به این صورت است که پس از هر شکست، خانه ای که دارای دقت ۱۰۰٪ باشد، نمایانگر خاتمه ی شکست در آن شاخه است. اما خانه ای که دارای دقت کمتر از ۱۰۰٪ باشد بیان کننده ی این مطلب است که، رکوردهای رسیده به آن به دسته های مختلفی تعلق دارند، لذا شکست درآن قسمت ادامه می یابد.(۱۷)

ارزیابی درخت تصمیم

ماتریس فرضی ۲*۲ (جدول ۱) جدول ارزیابی دسته بندی برای یک ویژگی دو مقداری با مقادیر فرضی مثبت و منفی را نشان می دهد که تفسیر آن به شرح زیر است: (ستون ها مقادیر حقیقی و سطر ها مقادیر پیش بینی می باشند). همچنین دقت دسته بندی از فرمول ۱ محاسبه می گردد:

$$accuracy = (TN+TP) / (TN+TP+FN+FP)$$

TP: تعداد نمونه های مثبتی که به درستی مثبت تشخیص داده شده اند.
FP: تعداد نمونه های منفی که به اشتباه مثبت تشخیص داده شده اند.
FN: تعداد نمونه های مثبتی که به اشتباه منفی تشخیص داده شده اند.
TN: تعداد نمونه های منفی که به درستی منفی تشخیص داده شده اند.

عملگر MetaCost

در شبیه سازی صورت گرفته برای پیاده سازی مدل درخت تصمیم از عملگر Metacost استفاده شده است. این ترکیبی از ماتریس هزینه و یک چرخه تکرار است. اساس کار آن به این گونه است که ابتدا به مقادیر ماتریس پیش بینی شده و حقیقی داده ها که توسط نرم افزار محاسبه می شود، ارزش هایی به انتخاب اپراتور تعلق می گیرد. در واقع یک ماتریس هزینه تشکیل می شود که مقدار مثبت نشان دهنده کاهش و مقدار منفی نمایانگر افزایش ارزش هر درایه از ماتریس خواهد شد. سپس تا رسیدن به یک آستانه ی مطلوب، چندین بار و در یک چرخه، عمل دسته بندی توسط درخت تصمیم انجام می گیرد تا نهایتاً بهترین درخت تصمیم حاصل شود.(۱۵)

یافته‌ها:

بررسی های آماری

پس از بررسی های آماری و با توجه به جدول ۲، مشخص شد نسبت مبتلایان مرد به زن در مشهد برابر ۵۶٪ به ۴۴٪ بود. همچنین مبتلایان با سابقه فامیلی نیز ۳۷٪ موارد را شامل می شدند. از طرف دیگر درصد افراد جوان مبتلا نسبتاً بالا بود، به طوری که ۱۴٪ افراد در بازه سن ۳۹-۲۰ سال و ۳۷٪ افراد کمتر از ۵۰ سال سن داشتند. در بحث ناحیه تومور سرطانی نیز آمار افراد مبتلا به سرطان ناحیه دیستال (۳۹ درصد) در مقایسه با پروگزیمال (۳۵ درصد) و رکتوم (۲۶ درصد) بیشتر بود. نتیجه بررسی ویژگی مهم AmsterdamII تنها شامل ۳٪ مورد مثبت بود. از طرفی ویژگی Revisedethesda که نسبت به ویژگی AmsterdamII حساسیت پایین تر و شمول بیشتری دارد، ۴۲٪ مورد مثبت را در بر داشت. همچنین بررسی های آماری بصورت جداگانه برای زنان و مردان، تفاوت زیادی بین این دو گروه با آمار کلی را نشان نمی داد، جز موارد اندکی مانند این که در افراد گروه سنی زیر ۵۰ سال، نسبت مردان به زنان برابر ۳۲٪ به ۴۴٪ بود که بیان کننده جوان تر بودن مبتلایان زن نسبت به مردان است.

بررسی ناحیه دقیق تومور

در این مرحله ابتدا با استفاده از نرم افزار، داده ها بارگذاری شدند. پس از آن با استفاده از ویژگی خروجی سه مقداری Location.cat که در واقع خود از ویژگی ۸ مقداری Location و طبق جدول ۳ استخراج شده است، برای دو برچسب پروگزیمال و دیستال الگوریتم درخت تصمیم پیاده سازی شد. (بدیهی است با توجه به این که در جدول ۳ مشخص شد برچسب رکتوم خود یک مقداری است، لذا درخت تصمیم برای آن قابل پیاده سازی نخواهد بود، درنتیجه برای بررسی آن در مراحل بعدی راه کارهای دیگری ارائه خواهد شد.) قبل از پیاده سازی درخت تصمیم از عملیات MetaCost استفاده شد. به این صورت که پس از اختصاص ماتریس هزینه مشخص (Class2: cecum, Class1: transverse و Class3: ascending) که در جدول ۴ نشان داده شده، تعداد تکرار ۱۷ برای عملگر تعیین شد. برای بهبود عملکرد درخت تصمیم و بر اساس تجربه مقادیری خاص برای هر درایه منظور گردید که به عنوان مثال به درایه ۳*۳ ارزشی بیشتر از همه (عدد ۴.۰-) اختصاص یافت، زیرا فقط ۳ عدد ناحیه ascending وجود داشت و لازم بود حتماً همگی به درستی ارزیابی و پیش بینی شوند. پس از تنظیم ماتریس هزینه، با در نظر گرفتن عمق ۸ و confidence ۰/۲۵ برای الگوریتم، مدل درخت تصمیم برای ناحیه پروگزیمال پیاده سازی شد. همانطور که در جدول ۵ مشخص شده درخت تصمیم حاصله، دارای دقت مناسب ۸۸/۲۹٪ است. در مورد تفسیر این جدول و همانطور که پیش تر به آن اشاره شد ستون ها مقادیر حقیقی و سطر ها مقادیر پیش بینی می باشند. لذا به عنوان مثال عدد ۷۴ در درایه ۱*۱ این جدول مبین این موضوع است که ۷۴ داده transverse بوده اند که به درستی پیش بینی شده اند، اما در مقابل آن و عدد ۴ نشان دهنده آن است که ۴ عدد از داده ها در اصل cecum بوده اند که به اشتباه transverse پیش بینی شده است. مدل درخت تصمیم حاصله نیز در شکل ۱ به نمایش درآمده است.

به طور مشابه همین فرآیند، البته با ماتریس هزینه ای متفاوت دقیقاً برای ناحیه دیستال اجرا شد. دقت این مدل نیز برابر با ۸۲/۹۳٪ ارزیابی شد (جدول ۶) که مدل درخت تصمیم آن در شکل ۲ نشان داده شده است.

جدول۲: کمیت و درصد برخی ویژگی های مهم با تفکیک جنسیت

Attribute	Total		Female		Male	
Value	no	٪	No	٪	No	٪
Sex						
Female(۱)	۱۴۰	٪۴۴	-	-	-	-
Male(۲)	۱۷۶	٪۵۶	-	-	-	-
Location.cat						
Proximal(۱)	۱۱۱	٪۳۵	۵۱	٪۳۶	۶۰	٪۳۴
Distal(۲)	۱۲۳	٪۳۹	۵۵	٪۳۹	۶۸	٪۳۹
Rectum(۳)	۸۲	٪۲۶	۳۴	٪۲۵	۴۸	٪۲۷
Family history						
Present(۱)	۱۱۸	٪۳۷	۵۵	٪۳۹	۶۳	٪۳۴
Absent(۲)	۱۹۸	٪۶۳	۸۵	٪۶۱	۱۱۳	٪۶۶
AmsterdamII						
Present(۱)	۹	٪۳	۴	٪۲/۹	۵	٪۲/۸
Absent(۲)	۳۰۵	٪۹۶/۴	۱۳۵	٪۹۶/۴	۱۷۰	٪۹۶/۷
Unknown(۳)	۲	٪۰/۶	۱	٪۰/۷	۱	٪۰/۵
RevisedBethesda						
Present(۱)	۱۳۳	٪۴۲	۶۵	٪۴۶	۶۱	٪۳۹
Absent(۲)	۱۵۸	٪۵۰	۶۴	٪۴۶	۹۴	٪۵۳
Unknown(۳)	۲۵	٪۸	۱۱	٪۸	۱۴	٪۸
Age.cat						
[۲۰,۳۹]	۴۴	٪۱۴	۲۰	٪۱۴	۲۴	٪۱۴
[۴۰,۵۹]	۱۳۸	٪۴۴	۶۴	٪۴۶	۷۴	٪۴۲
[۶۰,۷۹]	۱۱۷	٪۳۷	۴۹	٪۳۵	۶۸	٪۳۹
More than ۸۰	۱۷	٪۵	۷	٪۵	۱۰	٪۵
Age.cat ۵۰						
۰	۱۱۸	٪۳۷	۶۱	٪۴۴	۵۷	٪۳۲
۱	۱۹۸	٪۶۳	۷۹	٪۵۶	۱۱۹	٪۶۸

جدول۳: چگونگی تقسیم بندی ویژگی location به سه ناحیه ویژگی location.cat و کمیت آنها

Value	۱	۲	۳	۴	۵	۶	۷	۸
Count	۲۶	۸۶	۸۲	۳۶	۳	۸۲	۰	۱
Location	cecum	Sigmoid	rectum	rectosigmoid	ascending	transverse	descending	anus
Location.cat	proximal	Distal	rectum	distal	proximal	proximal	distal	distal

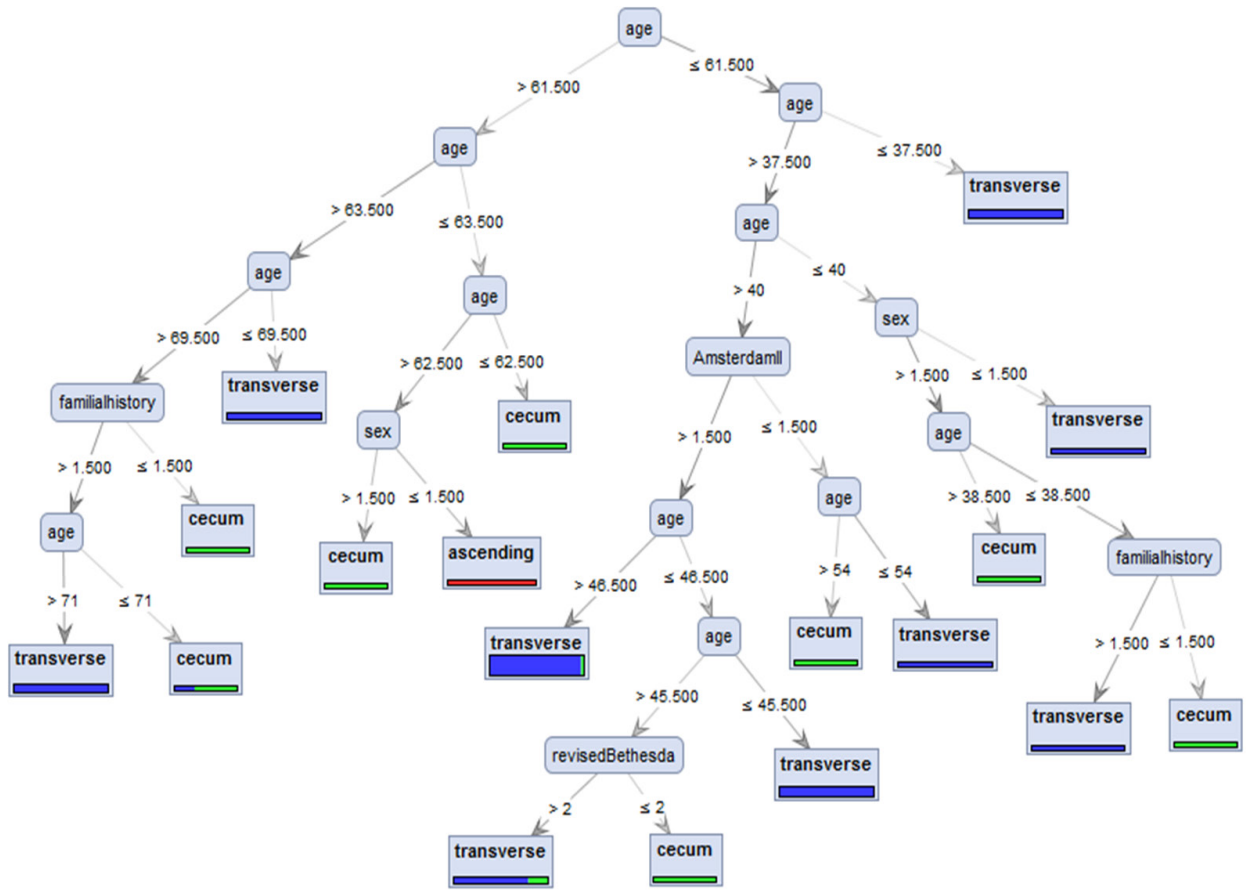
جدول۴: ماتریس هزینه تعیین شده برای درخت تصمیم ناحیه proximal

Cost Matrix	True Class 1	True Class 2	True Class 3
Predicted Class 1	-۱.۰	۲.۰	۱.۰
Predicted Class 2	۱.۵	-۲.۰	۱.۰
Predicted Class 3	۱.۰	۱.۰	-۴.۰

جدول۵: ارزیابی دقت درخت تصمیم گیری ناحیه proximal

Accuracy: 88/29%

	True Transverse	True cecum	True Ascending	Class Precision
Pred. Transverse	۷۴	۴	۰	٪۹۴/۸۷
Pred. Cecum	۶	۲۱	۰	٪۷۷/۷۸
Pred. Ascending	۲	۱	۳	٪۵۰/۰۰
Class Recall	٪۹۰/۲۴	٪۸۰/۷۷	٪۱۰۰/۰۰	



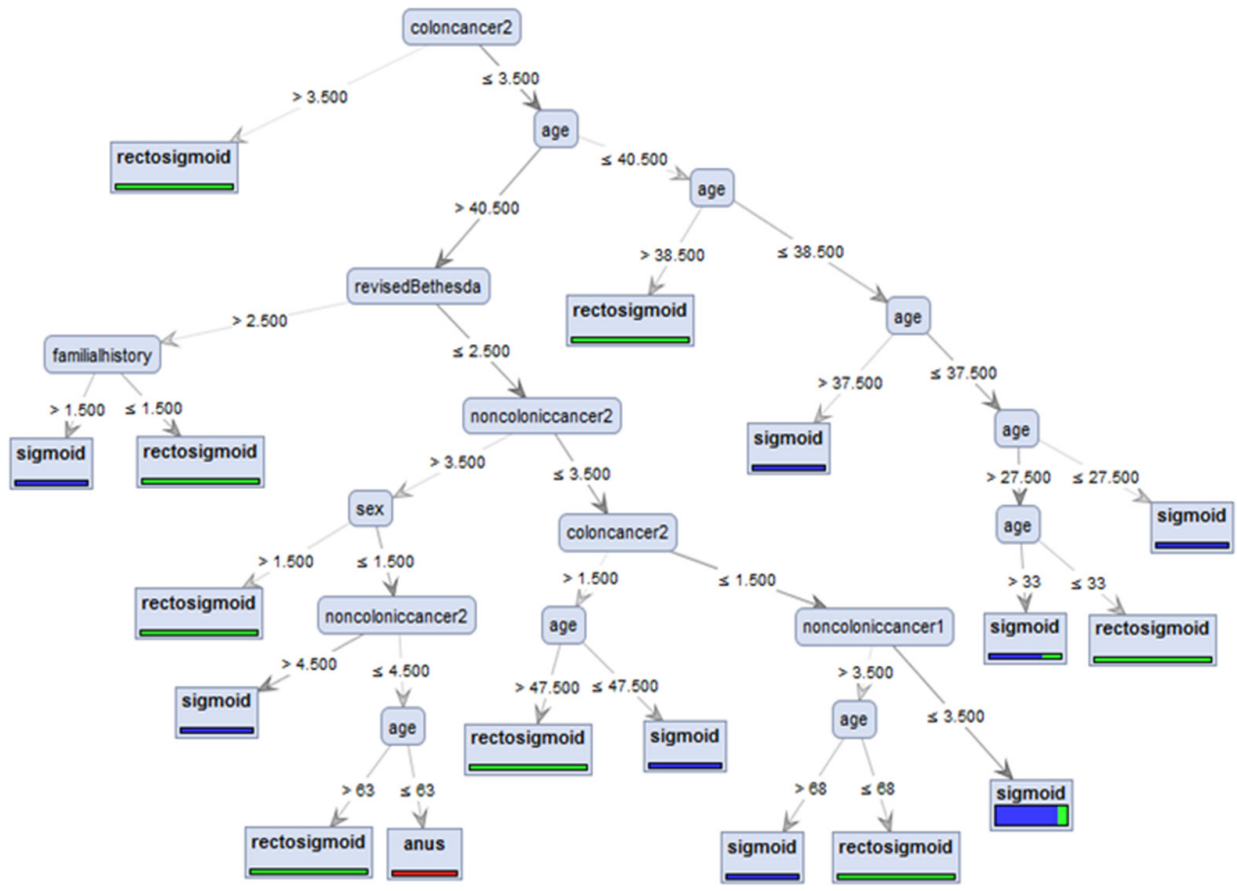
شکل۱: درخت تصمیم گیری ناحیه proximal

در تفسیر مدل های درخت تصمیم بدست آمده به عنوان مثال و با توجه به جدول ۵ می توان پیش بینی کرد که آن دسته از افرادی که تومور آنها در ناحیه پروگزیمال است و سن کمتر از ۶۱/۵ و بیشتر از ۴۰ سال و AmsterdamII کوچکتر از ۱/۵ (value: ۱) دارند، اگر سنشان کمتر یا مساوی ۵۴ سال باشد، به احتمال فراوان تومور آنها در ناحیه transverse و اگر سنشان بیشتر از ۵۴ سال باشد تومور آنها در ناحیه cecum خواهد بود. برای شکل مشابه ۲ نیز می توان به عنوان مثال پیش بینی کرد آن دسته از افرادی که تومور آنها در ناحیه دیستال و با این ویژگی ها باشد که؛ Coloncancer2 کمتر از ۳/۵ (value: ۰,۱,۲,۳)، سن بزرگتر از ۴۰/۵ و RevisedBethesda بیشتر از ۲/۵ (value: ۳) باشد، اگر familialhistory بزرگتر از ۱/۵ (value: ۲) داشته باشند تومور آنها به احتمال فراوان در ناحیه سیگموید و اگر familialhistory کوچکتر از ۱/۵ (value: ۱) داشته باشند تومور آنها در ناحیه رکتوسیگموید خواهد بود. (مقادیر عددی هر ویژگی یا همان value در جداول ۳ و ۵ ذکر شده است. به جز Coloncancer2 که در آن؛ none = ۰، grand father = ۱، grand mother = ۲، aunt = ۳، uncle = ۴، cousin = ۵، niece = ۶ و more than one = ۷ است.)

جدول۶: ارزیابی دقت درخت تصمیم گیری ناحیه distal

Accuracy: 82/93%

	True Sigmoid	True Rectosigmoid	True Anus	Class Precision
Pred. Sigmoid	۸۵	۲۰	۰	٪۸۰/۹۵
Pred. Rectosigmoid	۱	۱۶	۰	٪۹۴/۱۲
Pred. Anus	۰	۰	۱	٪۱۰۰/۰۰
Class Recall	٪۹۸/۸۴	٪۴۴/۴۴	٪۱۰۰/۰۰	



شکل۲: درخت تصمیم گیری ناحیه distal

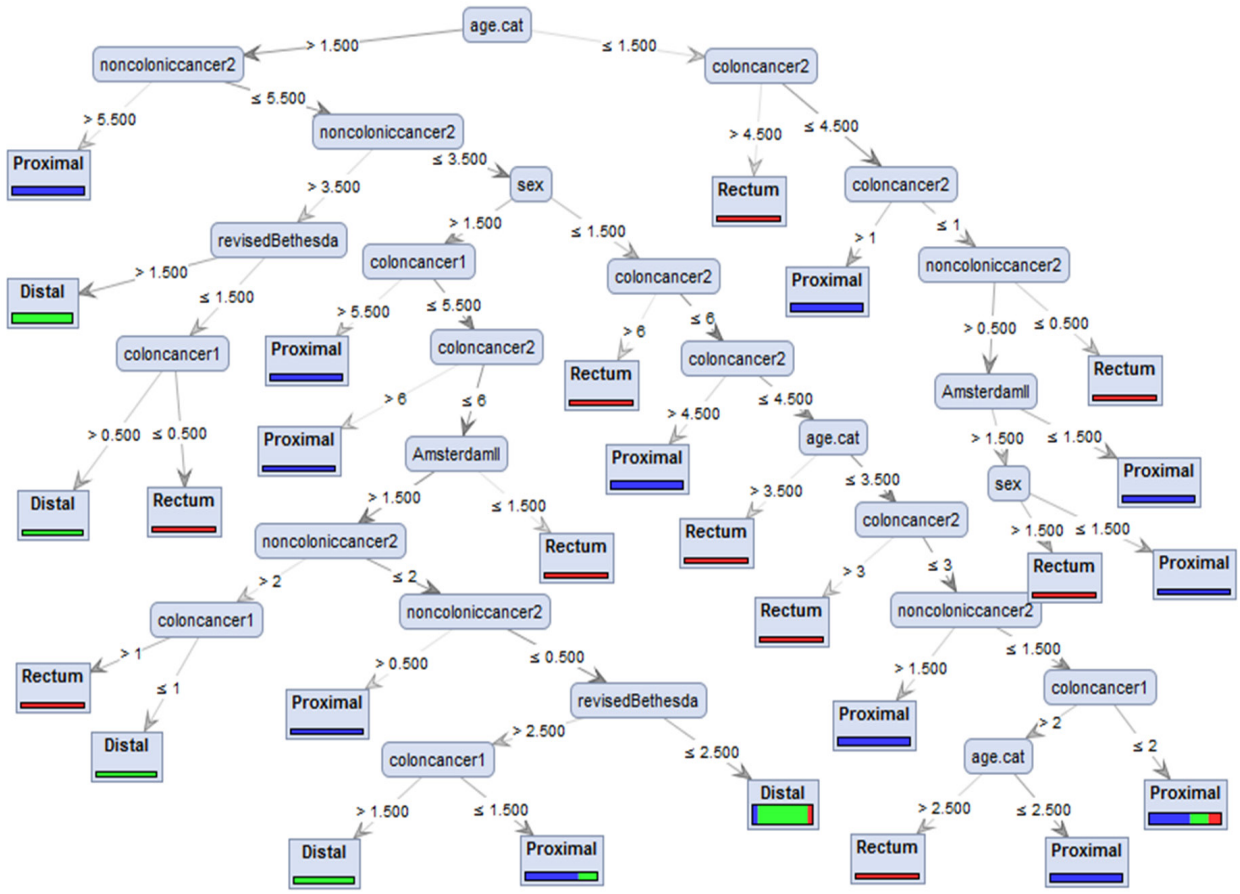
بررسی ناحیه کلی تومور

در این مرحله از شبیه سازی نیز فرآیند کلی پیاده سازی الگوریتم درخت تصمیم مشابه حالت قبل صورت گرفت، با این تفاوت که تنها با اتکا به ویژگی خروجی Location.cat که نسبت به ویژگی Location دارای حالتی کلی تر و فقط شامل سه ناحیه اصلی است، فرآیند انجام شد. در این مرحله از شبیه سازی برای دقت بیشتر و تفسیر ساده تر، ابتدا داده ها به دو قسمت دسته بندی شدند. قسمت یک برای داده ها با سابقه خانوادگی (familialhistory = ۱) و قسمت دوم برای داده ها بدون سابقه خانوادگی (familialhistory = ۲) ابتدا برای داده ها با familialhistory = ۱، مشابه قبل یک ماتریس هزینه درنظر گرفته و تعداد تکرار عملگر برابر ۱۰ منظور شد. عمق ۱۲ نیز برای درخت انتخاب شد. نتیجه عملیات، درخت تصمیمی با دقت مناسب ۸۲.۲ بود که نتایج آن در جدول ۷ و شکل ۳ به تصویر در آمده است. به عنوان مثال در تفسیر یکی از شاخه های این درخت، می توان پیش بینی کرد آن دسته از بیماران مراجعه کننده که دارای سابقه خانوادگی باشند، اگر age.cat آنها بزرگتر از ۱۵/۵ (value: ۲,۳,۴) و noncoloniccancer2 آنها بین ۵/۵ و۳/۵ (value: ۴,۵)

جدول ۷: ارزیابی دقت درخت تصمیم گیری کلی نواحی تومور افراد با سابقه خانوادگی

Accuracy: 82/20%

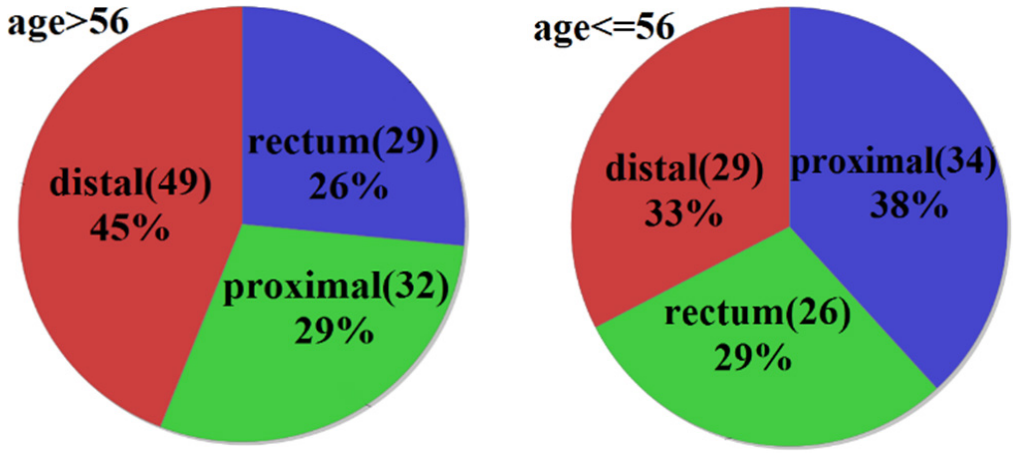
	True Proximal	True Distal	True Rectum	Class Precision
Pred. Proximal	۴۰	۹	۲	٪۷۸/۴۳
Pred. Distal	۳	۳۲	۰	٪۹۱/۴۳
Pred. Rectum	۲	۵	۲۵	٪۷۸/۱۲
Class Recall	٪۸۸/۸۹	٪۶۹/۵۷	٪۹۲/۵۹	



شکل ۳: درخت تصمیم گیری کلی نواحی تومور افراد با سابقه خانوادگی

باشد، اگر RevisedBethesda آنها بزرگتر از ۱۵/۵ (value: ۲,۳) باشد، به احتمال بسیار زیاد تومور آنها در ناحیه دستال و اگر RevisedBethesda آنها کوچکتر از ۱/۵ (value: ۱) و coloncancer1 برابر ۰ داشته باشند تومور آنها در ناحیه رکتوم خواهد بود (مقادیر noncoloniccancer2 مشابه coloncancer2 است که انتها به آن اشاره شد، همچنین مقادیر coloncancer1 عبارتند از: ۰ = none، ۱ = father، ۲ = mother، ۳ = child، ۴ = sister، ۵ = brother و ۶ = more than one).

اما برای داده ها با ۲ = familialhistory و پس از انجام مراحل مشابه فوق، مدل مناسبی حاصل نشد که دلیل آن نبودن نظم نه چندان زیاد در داده های مربوط به این دسته است. البته برای این دسته از داده ها نیز با استفاده از مشاهدات نموداری می توان به نتایجی کلی و تقریبی رسید. به عنوان مثال با بررسی این دسته از داده ها و تقسیم آنها به افراد زیر و بالای سن میانگین (کوچک تر و بزرگتر از ۵۶ سال) این نتیجه حاصل شد که با اختلاف معنا داری اکثر افراد بالای ۵۶ سال این دسته، تومورشان در ناحیه دیستال بود (٪۴۴)، از طرفی افراد زیر ۵۶ سال نیز با اختلاف کمی اکثرا دارای تومور در ناحیه پروگزیمال بودند (٪۳۸) که این نتایج در شکل ۴ قابل مشاهده است.



شکل۴: افراد با familialhistory = 2 و دو دسته بالا و پایین سن میانگین

بحث:

با توجه به یافته ها و نتایج بدست آمده و پس از تحلیل آنها مشاهده شد که تفاوت ها و شباهت هایی با آمارهای جهانی در این زمینه وجود دارد. از جمله می توان به درصد پایین تر مبتلایان با سابقه فامیلی در این بررسی و همچنین نسبت مبتلایان مرد به زن که ۵۶٪ به ۴۴٪ بود اشاره کرد و این مشابه آماری است که در (۱و۵و۶و۷و۹و۱۸) نیز به آن اشاره شده است. اما آمارهایی متفاوت با نرم جهانی نیز وجود داشت؛ نخست آنکه سهم افراد جوان مبتلا به این سرطان نسبتاً زیاد بود، به طوری که درصد بالایی از بیماران، کمتر از ۵۰ سال سن داشتند. البته این مورد مشابه مطالعات بومی دیگری چون (۹-۵) بود که در (۶) علت آن نقش مهم عوامل ژنتیکی بیان شده است، لذا با در نظر گرفتن فرضیه مذکور، بررسی عوامل ژنتیکی و همچنین توجه ویژه به بستگان درجه ۱ و ۲ بیمار در جهت غربالگری مناسب می تواند زمینه ساز مقابله مؤثر با سرطان کلورکتال و کاهش مرگ و میر ناشی از آن باشد. اما نکته قابل توجه، آمار بیشتر افراد مبتلا به ناحیه دیستال (۳۹ درصد) در مقایسه با پروگزیمال (۳۵ درصد) و رکتوم (۲۶ درصد) بود که این مورد حتی بر خلاف برخی آمار بومی از جمله بود.(۶) همچنین در بخش تفکیک جنسیت نیز مشخص شد بیماران مرد با تومور ناحیه دیستال به میزان ۱۰٪ بیشتر از بیماران زن بودند. جالب اینجاست در (۱۹) که مربوط به مطالعه سرطان کولورکتال در کشور سوئد است، بیان شده که مصرف گوشت قرمز به طور معنا داری با سرطان کولورکتال ناحیه دیستال ارتباط مستقیم دارد. از طرفی (۲۰) بیان می کند که آمار بالاتر سرطان ناحیه دیستال نسبت به ناحیه پروگزیمال نشان دهنده پرخطر بودن آن جامعه است. لذا با توجه به نتایج و مطالعات موجود می توان به این فرضیه رسید؛ یکی از دلایل ابتلای زیاد به سرطان کولورکتال ناحیه دیستال، مصرف بالای گوشت قرمز در مشهد و بخصوص در مردان می باشد، که در نتیجه ممکن است شهر مشهد برای سرطان کولورکتال یک جامعه پرخطر محسوب شود. بررسی دو ویژگی مهم AmsterdamII و Revisedethesda نیز با توجه به مطالعاتی که در (۱۶) صورت گرفته بود، برای بیماران صورت پذیرفت که طبق انتظار ویژگی AmsterdamII دارای حساسیت بیشتر و شمول بسیار کمتری نسبت به ویژگی Revisedethesda برای داده های بیماران بود که این آمار و جزئیات مربوط به آن که در بالا به آن اشاره شد می تواند در تحقیقات آتی و تکمیلی تر مورد استفاده محققان قرار گیرد.

در بخش دیگری از مطالب همبستگی ویژگی های داده ها مورد ارزیابی قرار گرفت که مشخص شد سه ویژگی age، sex و familialhistory همبستگی بالاتری با ویژگی خروجی مد نظر این مطالعه که location بود داشتند. با توجه به نتایج حاصله از جمله فواید آشکار این بررسی کمک به پیش بینی ناحیه تومور سرطانی بود چرا که این امر رابطه مستقیمی با میزان همبستگی هر ویژگی با ویژگی location دارد. از طرفی دیگر کمک شایانی به طراحی الگوریتم درخت های تصمیم پیاده شده نمود. زیرا در پیاده سازی مدل درخت تصمیم، استفاده از ویژگی هایی که دارای همبستگی بیشتر با ویژگی خروجی هستند، نتیجه و دقت بیشتری را در مقایسه با استفاده از تمامی ویژگی ها ارائه می کند.

اما در قسمت نهایی پژوهش با توجه به یافته های قبلی و همچنین شاخصه همبستگی که به آن اشاره شد، مدل هایی مبتنی بر علم داده کاوی ارائه شد. پیش تر در مطالعاتی چون (۱۴و۲۱) نیز کارهایی مبتنی بر داده کاوی بر روی سرطان کولورکتال صورت گرفته که بیشتر در حوزه اطلاعات ژنتیکی بوده و با روش های ارائه شده در این مطالعه کمی متفاوت است. در این پژوهش با استفاده از روش های دسته بندی و مدل درخت تصمیم عملیات کاوش بر روی داده های بالینی صورت گرفت. نخست در دید جزئی تری درخت هایی برای نواحی پروگزیمال و دیستال جهت تشخیص دقیق محل تومور پیاده شد که با توجه به یک دسته ای بودن ناحیه رکتوم این فرآیند برای آن قابل اجرا نبود. پس از آن و در یک نگاه کلی تر از درخت تصمیم برای یافتن ناحیه کلی تومور سرطانی استفاده شد. با توجه به مدل های بدست آمده نتایج بسیار مهمی از درخت های ترسیم شده قابل استخراج هستند که از طرفی می تواند به پزشکان در امر تشخیص و حتی پیش بینی ناحیه تومور سرطانی کمک شایانی کند و از طرف دیگر در زمان، هزینه و آزمایش ها صرفه جویی قابل ملاحظه ای را

به ارمغان آورد که برآیند آن سلامتی و بقای بیشتر بیماران خواهد بود.

اما در بخش بررسی ناحیه کلی و برای دسته familialhistory = ۲ روند مطالعه و پیاده سازی مدل درخت تصمیم با چالش مواجه شد، و آن این بود که با وجود بهره گیری از روش های مختلف دسته بندی و حتی خوشه بندی، مدل درخت تصمیم با دقت مناسبی برای این دسته از بیماران قابل پیاده سازی نبود! هرچند که تا به این مرحله و با استفاده از علم داده کاوی و نرم افزار RapidMiner نتایج دقیق و مفیدی حاصل شد که می تواند قابل توجه پژوهشگران و محققان این زمینه باشد اما با توجه به چالش به وجود آمده، این نتیجه حاصل شد که رسیدن به پیش بینی های با دقت مناسب برای مجموعه داده ها بزرگتر و دارای نظم کمتر، داشتن ویژگی های بیشتر و تخصصی تر از بیمار را می طلبد. ویژگی های نظیر خصوصیت های ژنتیکی، جغرافیایی، غذایی و غیره که از جمله محدودیت های این مطالعه بود و پیش تر در مطالعاتی چون (۱۴و۱۹و۲۱) از آنها استفاده شده است.

در واقع هدف اصلی این پژوهش پیش بینی ناحیه تومور سرطانی کولورکتال با استفاده از ویژگی هایی ساده و کم هزینه و بهره گیری از علم داده کاوی و مدل درخت تصمیم بود که با توجه به مدل های ارائه شده و یافته های آماری که در مراحل اولیه ذکر شد، می تواند به پژوهشگران و محققان این زمینه در جهت مطالعات و تحقیقات آتی آنها کمک کند. همچنین بررسی نواحی مختلف بروز سرطان کولورکتال نیز مهم بنظر می رسد. در پایان نیز برای بهبود نتایج پژوهشگران، استفاده از روش های داده کاوی و ابزارهای کاربردی آن مانند RapidMiner نیز پیشنهاد می شود.

REFERENCES:

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. *CA Cancer J Clin* 2015;6:87-108.
2. Siegel R L, Miller K D , Jemal A.Cancer Statistics, 2016. *CA Cancer J Clin* 2016;66:7-30.
3. Siegel R L, Miller K D , Jemal A. Cancer Statistics, 2015. *CA Cancer J Clin* 2015;65:5-29.
4. Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, et al.Cancer Statistics, 2006. *CA Cancer J Clin* 2006;56:106-30.
5. Ansari R, Mahdavinia M, Sadjadi A, Nouraie M, Kamangar F, Bishehsari F, et al. Incidence and age distribution of colorectal cancer in Iran: Results of a population-based cancer registry. *Cancer Lett* 2006;240:143-7.
6. Azadeh S, Moghimi-Dehkordi B, Fatemi SR, Pourhoseingholi M, Ghiasi S, Zali MR. Colorectal Cancer in Iran: an Epidemiological Study. *Asian Pac J Cancer Prev* 2008;9:123-6.
7. Moradi A, Khayamzadeh M, Guya M, Mirzaei HR, Salmanian R, Rakhsha A,et al. Survival of Colorectal Cancer in Iran. *Asian Pac J Cancer Prev* 2009;10:583-6.
8. Azadeh S, Fatemi SR, Sara A, Mohsen V, Bijan M, Zali MR. Four years Incidence Rate of Colorectal Cancer in Iran: A Survey of National Cancer Registry Data - Implications for Screening. *Asian Pac J Cancer Prev* 2012;13:2695-8.
9. Azadeh S, Moghimi-Dehkordi B, Fatemi SR, Maserat E, Pourhoseingholi M, Ghiasi S, Zali MR. Risk of Colorectal Cancer in Relatives: A Case Control Study. *Knowledge & Health* 2009;4:12-5
10. Gupta S, Kumar D, Sharma A. Data mining classificationte chniques applied for breast cancer diagnosis and prognosis. *IJCSE* 2011;2:188-95.
11. Lihua L, Hong T, Zuobao W, Jianli G, Michael G, Jun Z, et al. Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med* 2004;32:71-83.
12. Kang J O, Chung SH, Suh YM. Prediction of Hospital Charges for the Cancer Patients with Data Mining Techniques. *J Kor Soc Med Informatics* 2009;15:13-23.
13. Lisboa PJ, Vellido A, Tagliaferri R, Napolitano F, Ceccarelli M, Martín-Guerrero JD, et al. Data Mining in Cancer Research. *Ieee Computational Intelligence Magazine* 2010:14-8.
14. Cava C, Zoppis I, Gariboldi M, Castiglioni I, Mauri G, Antoniotti M. Combined analysis of chromosomal instabilitiesand gene expression for colon cancer progression inference. *J Clin Bioinforma* 2014;4:2.
15. Khakshoor MM, Poorbadakhshan K, Abbaszadeh H, Khakshoor J, A Study About Investigating of Effective Factors Influencing Breast Cancer and Identifying Their Relevance with Data Mining. in Nastaran Cancer Symposium(NCCP), Mashhad, 2015.
16. Giardiello FM, Allen JI, Axilbund JE, Boland C, Burke CA, Burt RW, et al. Guidelines on Genetic Evaluation and Management of Lynch Syndrome: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. *Am J Gastroenterol* 2014:1-21.
17. Abadeh MS , Mahmudi S. Practical Data mining (In Persian), Tehran: *Niazedanesh* 2014.
18. Siegel R, DeSantis C, Jemal A. Colorectal Cancer Statistics, 2014. *CA Cancer J Clin* 2014;64:104-17.
19. Larsson SC, Rafter J, Holmberg L, Bergkvist L, Wolk A. Red meat consumption and risk of cancers of the proximal colon, distal colon and rectum: The Swedish Mammography Cohort. *Int J Cancer* 2005;113:829-34.
20. Corman ML. Colon and Rectal surgery. 4thed., New York: Lippincott company, 1999.
21. Thomas A, Patterson NH, Marcinkiewicz MM, Lazaris A, Metrakos P, Chaurand P. Histology-Driven Data Mining of Lipid Signatures from Multiple Imaging Mass Spectrometry Analyses: Application to Human Colorectal Cancer Liver Metastasis Biopsies. *American Chemical Society* 2013;85:2860-6.