دانشگاه الزهرا - دانشکده فنی و مهندسی

پایان نامه جهت اخذ درجه کارشناسی ارشد

رشته مهندسی کامپیوتر- گرایش هوش مصنوعی

عنوان

کشف تقلب در سیستم‌های مراقبت سلامت با رویکرد تحلیل گراف

استاد راهنما

دکتر محمدرضا کیوان پور

دانشجو

روناک نمکی

تابستان ۱۳۹۹

بسمه تعالی

نصیحت تکمیلی	صورت جلسه دفاع از پایان نامه دوره کارشناسی ارشد ورودی ۹۴ به بعد	کد: EM-FR-۰۷-۰۱ صفحه ۱ از ۱
	بازنگری: ۱ تاریخ بازنگری: ۱۳۹۶/۱۲/۲۱	شماره: ..... تاریخ: .....

نسخه پرونده دانشجو☐ نسخه تحصیلات تکمیلی☐ نسخه مالی (به تعداد مورد نیاز)☐

جلسه دفاع از پایان نامه تحصیلی خانم ..... به شماره دانشجویی .....

دانشجوی کارشناسی ارشد رشته /گرایش ..... با کد درس ..... به ارزش ..... واحد با عنوان:

" ..... "

در تاریخ ...../...../..... از ساعت ..... تا..... با حضور هیأت داوران تشکیل شد و پس از ارزیابی، اعتبار پایان نامه برای اخذ مدرک کارشناسی ارشد

مورد تایید قرارگرفت☐ مورد تایید قرار نگرفت ☐

اعتبار پایان‌نامه برای اخذ مدرک کارشناسی ارشد به شرح زیر مورد تأیید است:

الف) قبول با درجه ب) ☐مردود (نمره کمتر از ۱۴)

☐عالی (نمره ۱۹ تا ۲۰)

☐خیلی خوب (نمره ۱۸ تا ۱۸/۹۹)

☐خوب (نمره ۱۶ تا ۱۷/۹۹)

☐متوسط (نمره ۱۴ تا ۱۵/۹۹)

هیات داوران	نام و نام خانوادگی	امضا	تاریخ
استاد راهنمای اول			
استادراهنمای دوم			
استاد مشاور اول			
استاد مشاور دوم			
داور داخلی **			
داور خارجی**			

نام و نام خانوادگی استاد ناظر تحصیلات تکمیلی: امضا: تاریخ:

نام و نام خانوادگی مدیر گروه: امضا: تاریخ:

نام و نام خانوادگی معاون آموزشی/ رییس دانشکده: امضا: تاریخ:

* این صورت جلسه باید در حضور هیات داوران توسط ناظر در سه نسخه تنظیم و سپس امضا شود.

** مطابق مصوبه شورای آموزشی و تحصیلات تکمیلی ۹۶/۱۲/۲۲ حضور داور داخلی و داور خارجی کارشناسی ارشد تمامی رشته ها الزامی است.

***دفاع از پایان نامه نمی تواند زودتر از چهار ماه پس از تاریخ تصویب دانشکده باشد.

سپاس بی کران پروردگار یکتا را که هستی مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه چینی از علم و معرفت را روزیمان ساخت. جان ما را صفای خود ده و دل ما را هوای خود ده، و چشم ما را ضیای خود ده، و ما را از فضل و کرم خود آن ده که آن به.خداوندا به ما توفیق تلاش در شکست، صبر در نومیدی، رفتن بی همراه، جهاد بی سلاح، کار بی پاداش، فداکاری در سکوت، دین بی دنیا، مذهب بی عوام، عظمت بی نام، خدمت بی نان، ایمان بی ریا، خوبی بی نمود، گستاخی بی خامی، مناعت بی غرور، عشق بی هوس، تنهایی در انبوه جمعیت و دوست داشتن بی آنکه دوستت بدارند، را عنایت فرما.

یارب دل ما را تو به رحمت جان ده درد همه را به صابری درمان ده

این بنده چه داند که چه می باید جست داننده تویی هر آنچه دانی آن ده

به نام خدا

منشور اخلاق پژوهش

با یاری از خداوند سبحان و اعتقاد به این که عالم محضر خداوند است و همواره ناظر به اعمال انسان و به منظور پاس داشت مقام بلند دانش و پژوهش و نظر به اهمیت جایگاه دانشگاه در اعتلای فرهنگ و تمدن بشری ما دانشجویان دانشکده‌های دانشگاه **الزهرا** متعهد می گردیم اصول زیر را در انجام فعالیت‌های پژوهشی مد نظر قرار داده و از آن تخطی نکنیم:

۱- اصل حقیقت جوئی: تلاش در راستای پی جویی حقیقت و وفاداری به آن و دوری از هرگونه پنهان سازی حقیقت،

۲- اصل رعایت حقوق: التزام به رعایت کامل حقوق پژوهشگران و پژوهیدگان (انسان، حیوان و نبات) و سایر صاحبان حق،

۳- اصل مالکیت مادی و معنوی: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه همکاران پژوهش،

٤- اصل منافع ملی: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش،

٥- اصل رعایت انصاف و امانت: تعهد به اجتناب از هرگونه جانب داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار،

٦- اصل رازداری: تعهد به صیانت از اسرار و اطلاعات محرمانه افراد، سازمان‌هاو کشور و کلیه افراد و نهادهای مرتبط با تحقیق،

۷- اصل احترام: تعهد به رعایت حریم‌ها و حرمت‌ها در انجام تحقیقات و رعایت جانب نقد و خودداری از هرگونه حرمت شکنی،

۸- اصل ترویج: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به همکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد،

۹- اصل برائت: التزام به برائت جوئی از هرگونه رفتار غیر حرفه ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به شائبه‌های غیر علمی می آلایند.

نام و نام خانوادگی :

تاریخ و امضاء:

تعهدنامه‌ی اصالت پایان نامه

اینجانب **روناک نمکی** دانش آموخته مقطع تحصیلی کارشناسی ارشد در رشته‌ی **هوش مصنوعی** که در تاریخ ..... از پایان نامه ی خود تحت عنوان **کشف تقلب در سیستم‌های مراقبت سلامت با رویکرد تحلیل گراف** با کسب نمره / درجه..... دفاع نموده ام، متعهد می شوم:

۱- این پایان نامه / رساله حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و درمواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از مقاله، کتاب، پایان نامه و غیره) استفاده نموده ام، مطابق ضوابط ورویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوط ذکر و درج کرده ام.

۲- این پایان نامه / رساله قبلاً برای دریافت هیچ مدرک تحصیلی (هم سطح، پایین تر یا بالاتر) در سایر دانشگاه‌ها و موسسات آموزش عالی ارائه نشده‌است.

۳- چنانچه بعد از فراغت از تحصیل، قصد استفاده از هرگونه بهره برداری اعم از چاپ کتاب، ثبت اختراع و ازین دست موارد از این پایان نامه / رساله را داشته‌باشم، از حوزه معاونت پژوهشی دانشگاه **الزهرا** مجوزهای مربوطه را اخذ نمایم.

٤- چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می پذیرم و دانشگاهی مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و درصورت ابطال مدرک تحصیلی ام هیچ گونه ادعائی نخواهم داشت.

نام و نام خانوادگی : **روناک نمکی**

تاریخ و امضاء:

چکیده:

بیمه مراقبت‌های بهداشتی یک مشکل مبرم است و موجب افزایش هزینه‌های قابل‌توجهی در برنامه‌های بیمه درمانی می‌شود؛ بطوریکه کلاهبرداری در حوزه بهداشت و درمان (HCF) یک کلاهبرداری چند میلیارد دلاری است. وسیع بودن حوزه سلامت و حجم زیاد مالی باعث شده تا این حوزه برای کلاهبرداری مورد هدف قرار بگیرد. بنابراین حوزه سلامت به یک منبع هزینه‌ای قابل توجه در بسیاری از کشورها تبدیل شده است. یکی از منابع هزینه‌های قابل توجه سازمان بهداشت و سلامت، پرداخت سهم بیمه داروهای تجویز شده برای بیماران تحت پوشش است. بطور کلی، هدف از تشخیص تقلب، به حداکثر رساندن پیشبینی‌های درست و حفظ پیشبینی‌های نادرست در یک سطح قابل قبول از هزینه میباشد. با وجود تغییرات پیوسته رفتار متقلبین، مدل‌هایی که براساس تحلیل داده‌های گذشته ساخته میشوند ممکن است نتوانند شکل‌های جدید تقلب را شناسایی کنند. همچنین، هیچ یک از سیستم‌های شناسایی تقلب به تنهایی نمیتواند تمام شکل‌های تقلب را شناسایی و پوشش دهد. در این پایان‌نامه یک رویکرد نوین برای تخمین احتمال تقلب در اسناد درمانی با روش تحلیل گراف مورد بررسی قرار گرفته است. یک گروه از الگوریتم‌ها، شباهت‌های رفتاری را در دو دسته‌ی ارائه‌دهندگان مراقبت‌های بهداشتی کلاهبردار و غیرکلاهبردار با توجه به معیارهای قابل اندازه‌گیری فعالیت‌های مراقبت‌های بهداشتی مانند روش‌های پزشکی و تجویز داروها، محاسبه می‌کنند. مجموعه دیگری از الگوریتم‌ها، میزان انتشار خطر ناشی از تقلب ارائه‌دهندگان مراقبت‌های بهداشتی را از طریق موقعیت جغرافیایی، تکرار مکان‌های مشترک یا آدرس‌های دیگر، تخمین می‌زنند. این الگوریتم‌ها با توجه به توانایی آن‌ها در پیشبینی حضور یک ارائه‌دهنده در لیست ارائه‌دهندگان دفتر بازرسی کل (محرومیت از مشارکت در بیمه پزشکی سالمندان و سایر برنامه های مراقبت های بهداشتی فدرال)، ارزیابی شده‌اند.

کلیدواژه‌ها: نسخه دارویی، کشف ناهنجاری، سیستم‌های مراقبت سلامت، تحلیل گراف

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

فصل ۱

۱ مقدمه و کلیات

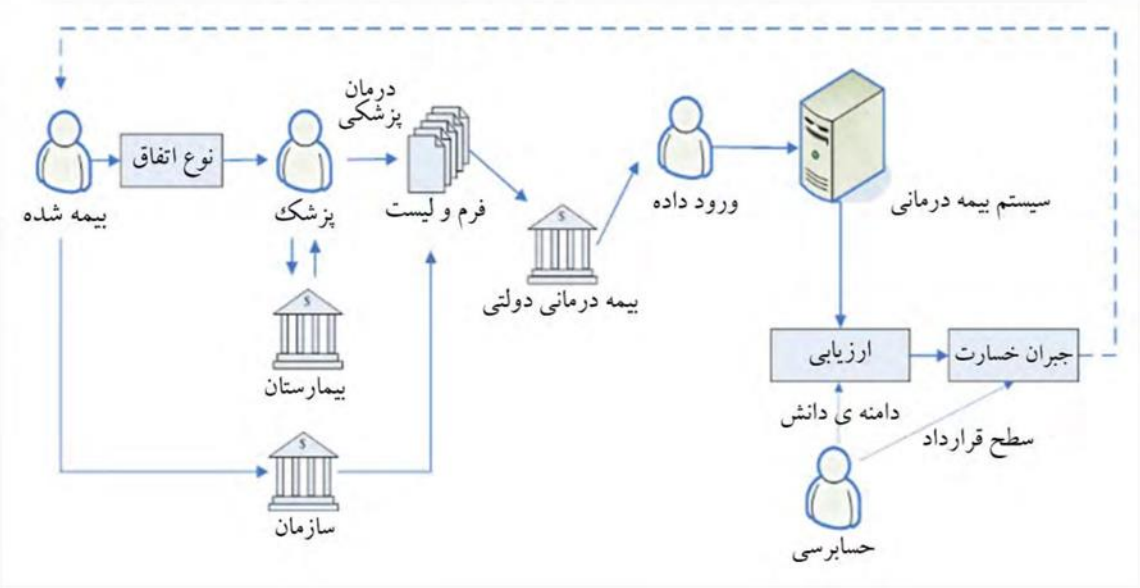
۱-۱ مقدمه

در این فصل از تحقیق ابتدا به تشریح مسئله پرداخته خواهد شد. کلیات و پیش نیازها بیان می شود و در ادامه، توضیحی در مورد تقلب در سیستم‌های پزشکی، اهمیت بررسی آنها و نکات مهم در مورد این مساله و همچنین روش‌های ممکن برای انجام اینکار و بررسی نقاط ضعف و قوت آنها و.. می پردازیم. سپس صورت مسئله تعریف شده و راه پیشنهادی این پایان نامه برای آن بیان خواهد شد.

در شرح مسئله چالش‌های موجود بررسی خواهد شد و پس از آن انگیزه های تحقیق بیان می شود. پس از بیان انگیزه های تحقق اهداف تحقیق بیان می شود. در بخش اهداف به نحوی دست آوردهای تحقیق که در

پایان بدست خواهد آمد نیز بیان می شود و در نهایت در بخش پایانی این فصل به تشریح ساختار پایان نامه پرداخته خواهد شد.

انجمن بیمه سلامت آمریکا، بیمه سلامت را به عنوان پوششی علیه ریسک هزینه‌های درمانی به علت بیماری یا آسیب دیدگی تعریف می‌کند. این پوشش می‌تواند توسط بعضی سازمان‌های مرکزی، برای مثال شرکت‌های خصوصی یا دولتی، ارائه شود. منبع این پوشش در بسیاری از کشورها صرف نظر از سیستم‌های بهداشت و درمانشان، متفاوت است. بررسی سالیانه انجام شده توسط صندوق مشترک المنافع، سیستم‌های بهداشت و درمان استرالیا، نیوزیلند، بریتانیا، آلمان، کانادا و ایالات متحده را مقایسه می‌کند. این بررسی تاکید می‌کند که ایالات متحده تنها کشور بدون پوشش بیمه سلامت سراسری است. اداره آمار ایالات متحده بیان می‌کند که ۳۱ درصد از آمریکایی‌ها طرح بیمه سلامت عمومی دارند، در حالی که ۵۵ درصد از آنها پوشش خود را از طریق کارفرمایانشان می‌گیرند. اگرچه، تحت پوشش بودن تضمین نمی‌کند که شخص بیمه شده هیچ هزینه پزشکی



شکل ۱ نمودار کلی گردش اطلاعات در سامانه‌های بیمه[۴]

پرداخت نکند. میزانی که بیمه شده باید بپردازد، قبل از اینکه بیمه‌گر برای یک ویزیت یا خدمت خاص بپردازد، پرداخت مشترک نامیده می‌شود. جدای از پرداخت مشترک، ممکن است خدماتی باشد که بیمه‌گر بر اساس حق بیمه‌ای که بیمه شونده می‌خرد، بازپرداخت می‌کند. مانند خدماتی که به عنوان بیمه تکمیلی شناخته می‌شوند که در آن درصد بیشتری از هزینه‌ها در قبال دریافت حق بیمه بیشتر پرداخت می‌شود. روند گردش اسناد در سیستم بیمه در شکل(۱) نشان داده شده است[۴].

تقریبا در هر سیستم بیمه سلامت، بیماران با پرداخت حق بیمه، پوشش سلامت می‌خرند و هنگام مراجعه به ارائه‌دهندگان خدمات بهداشتی و درمانی، پرداخت مشترکشان یا همان فرانشیز را انجام می‌دهند و خدمات دریافت می‌کنند. ارائه‌دهندگان، خدماتی را که به بیمار ارائه داده اند ثبت کرده و برای شرکت بیمه می‌فرستند. شرکت‌های بیمه فرم‌های صورتحساب را تحلیل می‌کنند و در خصوص مبلغی که باید به ارائه‌کنندگان بپردازند تصمیم می‌گیرند. این مبلغ به موارد عدم پوشش بیمه‌ای، الزامات پزشکی خدمات و دقت فرم صورتحساب بستگی دارد. شرکت‌های بیمه دستورالعمل هایی به مراکز درمانی ارسال می‌کنند که اعلام می‌کند کدام یک از خدمات پزشکی تحت پوشش بوده و نحوه پرداخت و میزان تعیین شده که بیمار باید بپردازد را توضیح می‌دهد.

۲-۱ تعریف مسئله

یکی از بزرگترین چالش‌های پیش روی شرکت‌های بیمه این است که فرم‌های صورتحساب نیازمند تحلیل هستند و باید در زمان محدودی تصمیم بگیرند کدام موارد باید بازپرداخت شوند. متاسفانه، تمام فرم‌های صورتحساب شامل اطلاعات صحیح نیستند، و عدم صحت فرم‌های صورتحساب هزینه بهداشت و درمان را افزایش می‌دهد. این اشتباهات میتواند خطاهای سهوی باشد، یا یک روش عمدی برای فریب دادن شرکت‌های بیمه. بنابراین، بسیاری از شرکت‌های بیمه به یک سیستم غربالگری بدون دخالت انسان برای بررسی فرم‌های صورتحساب نیاز دارند. این سیستم می‌تواند تصمیم بگیرد کدام صورتحساب‌ها باید دقیق‌تر بررسی شوند. این سیستم‌های تشخیص اولیه برای شکار ناهنجاری‌ها و بالا بردن پرچم قرمز با استفاده از روش‌های جدید مانند داده‌کاوی و روش‌های آماری معمولی طراحی شده‌اند.

سازمان‌های بیمه‌گر پس از دریافت اسناد هزینه از ارائه‌دهندگان خدمات سلامت اعم از مراکز درمانی، پزشکان، داروخانه‌ها، آزمایشگاه‌ها فرآیند بررسی هزینه‌ها و تطبیق آنها با معیارها و جداول هزینه شده توسط آنرا را که به آن رسیدگی به اسناد میگویند را آغاز مینمایند. با توجه به حجم اسناد قابل رسیدگی و کند بودن روش های مبتنی بر الگوهای ذهنی افراد خبره و همچنین کمبود منابع انسانی در صورتیکه بتوان بر اساس روش‌های مبتنی بر تحلیل داده‌ها، نسبت به کشف داده‌های تقلبی اقدام کرد، حجم بیشتری از هزینه‌های غیر قابل پرداخت را در زمان کوتاهتری از سبد هزینه سلامت حذف میگردد. همچنین با وجود حجم زیاد داده‌ها و متنوع بودن داده‌ها در حوزه سلامت، روش‌های سنتی یادگیری ماشین برای کشف تقلب در حوزه‌ی سلامت کافی نیستند. به همین دلیل استفاده از روش‌های تحلیل دادگان انبوه میتواند به فهم دقیقتر داده‌ها کمک کند. علاوه بر این، تحلیل

دادگان انبوه نه تنها قادر به پردازش حجم انبوه داده هستند بلکه به خوبی از پردازش موازی داده‌ها پشتیبانی میکند[۳].

تقلب در حوزه سلامت یک جرم بزرگ است و هزینه‌های شخصی و بودجه‌ای قابل توجهی به افراد، دولت‌ها و جامعه وارد میکند. بنابراین، کشف موثر تقلب برای کاهش هزینه‌ها و بهبود کیفیت سیستم سلامت بسیار مهم است. به منظور دستیابی به کشف موثرتر تقلب، بسیاری از پژوهشگران رویکردهای ضد تقلب پیچیده‌ای بر پایه داده‌کاوی، یادگیری ماشین و دیگر روش‌های تحلیلی توسعه دادند. این رویکردهای جدید ارائه شده دارای مزیت‌هایی مانند یادگیری خودکار الگوهای تقلب از داده‌ها، مشخص کردن احتمال تقلب برای هر مورد و شناسایی گونه‌های جدید تقلب دارند[۳].

کاهش ۱۰ ٪ هزینه‌های سلامت از طریق حذف اسناد تقلبی میتواند منجر به افزایشی به همین میزان در کیفیت و کمیت خدمات سلامت به بیمه‌شدگان باشد. ایجاد یک انباره داده حاصل از فرآیند فراخوانی، پالایش و بارگذاری داده ضمن استنادپذیر کردن داده‌های موجود در پایگاه‌داده‌ای سازمان‌های بیمه‌گر و ایجاد بستر داشبورد برای برپایی سامانه‌های هوش تجاری امکان تجزیه تحلیل و بهرمندی از روش‌های داده‌کاوی برای کشف تقلب را نیز فراهم میکند.

۳-۱ انگیزه‌های پژوهش

کلاهبرداری در حوضه بهداشت و درمان (HCF) با احتساب ۹۸ بیلون دلار از هزینه‌های سالانه که به بیمه پزشکی سالمندان (Medicare) و بیمه بهداشت مستمندان (Medicaid) در ایالات متحده مصرف می‌شود، یک تخلیه چند بیلون دلاری در هزینه های مراقبت های بهداشتی است[۵]. حجم بالای HCF به نسبت منابع موجود برای تحقیق و پیگرد این فعالیت‌های کلاهبرداری، پیگیری این موضوع را در اولویت قرار می‌دهد. یکی از منابع هزینه‌های قابل توجه سازمان‌های بهداشت، پرداخت سهم بیمه داروهای تجویز شده برای بیماران تحت پوشش است. هر ساله میلیون‌ها تقلب در نسخه تجویز شده و در نتیجه میلیاردها دلار هزینه برای این سازمان‌ها ایجاد می‌شود. در این بین افراد و نهادهای سودجو از جمله افراد تحت پوشش، پزشکان، شرکت‌های تولید دارو و داروخانه‌ها، به طرق مختلف به دنبال کسب منفعت و سودجویی برای خود هستند. در نتیجه، شناسایی و عدم پرداخت نسخ جعلی میتواند باعث کاهش هزینه قابل توجهی شود. از طرفی بررسی موردی همه‌ی نسخ دارویی توسط متخصصین بسیار پر هزینه و از نظر زمانی تقریبا ناممکن خواهد بود.

تجزیه و تحلیل گراف به دلایل مختلفی یک چارچوب امیدوارکننده برای ارزیابی خطر وقوع HCF است. اغلب چندین نهاد متخلف در وقوع HCF درگیر هستند. الگوریتم‌های نمایشی گراف با ایجاد روابط بین نهاد متخلف آشکار، تشخیص فعالیت‌های هماهنگ شده و گسترش نفوذ اجتماعی را تسهیل می‌کند. علاوه بر این، تجزیه و تحلیل گراف دارای سابقه اثبات شده در برنامه‌های اجرای قانون و تجزیه و تحلیل هوشمند اطلاعات است و با توجه به پژوهش‌های اخیر می‌توان گفت که آن‌ها می‌توانند در حوزه HCF مفید باشند.

در این پژوهش تلاش می‌شود تا با استفاده از روش‌های تحلیل گراف، پایگاه داده‌های بیمه پزشکی سالمندان آمریکا (و سایر برنامه های مراقبت های بهداشتی فدرال) که ماهانه توسط دفتر بازرس کل (OIG) منتشر می‌شود، بررسی شود و نسخه دارویی سسست‌هنجار و مشکوک به تقلب شناسایی شود و برای بررسی بیشتر در اختیار متخصصین بیمه قرار بگیرد.

۴-۱ اهداف

بررسی دقیق میزان استفاده از منابع موجود برای مراقبت درمانی و کشف الگوهای ناهنجار موجود در داده‌های درمانی با توجه به محدودیت‌های مالی در سازمان‌ها، از اهمیت بالایی برخوردار است. با توجه با اینکه امروزه داده‌های زیادی در سازمان‌های سلامت تولید می‌شود که دارای پیچیدگی‌های فراوانی هستند و با روش‌های سنتی قابل تحلیل نیستند، نیاز به الگوریتم‌های هوشمند بیش از پیش احساس می‌شود. با توجه به انسانی بود فعالیت فوق، محدودیت‌هایی نظیر خطای انسانی، کمبود نیروی انسانی خبره، محدودیت‌های زمانی فعالیت انسانی، عدم کیفیت یکسان در رسیدگی، احتمال وجود تعاملات انسانی ارزیابی شونده و سایر موارد بر رسیدگی تاثیر گذار است. حجم زیاد پرونده‌ها نیز بر مشکل افزوده و احتمال کشف موفق تقلب‌های پیچیده را کاهش می‌دهد. در نتیجه با وجود حجم، سرعت تولید و تنوع، داده‌های حوزه سلامت، این دسته از داده‌ها در گروه دادگان انبوه قرار میگیرند و برای تحلیل آن‌ها باید از روش‌های تحلیل دادگان انبوه استفاده کرد.

از ابزارهای قدرتمند برای تحلیل دادگان انبوه می توان به تکنیک‌های تحلیل گراف وروشهای مبتنی بر گراف اشاره کرد. تحلیل های بدست آمده از این روش می‌تواند، اطلاعات و الگوهای مفیدی را کشف کند. اطلاعات مفیدی که بسیاری از سازمان‌ها به راحتی قادر به کشف آنها نیستند. به طورکلی، روش‌های مبتنی بر گراف ابزار قدرتمندی برای تحلیل داده‌های با حجم و پیچیدگی زیاد است. به این ترتیب، در این پژوهش سعی بر آن است

که با شناسایی الگوهای موجود در تجویز نسخ دارویی و کشف موارد مشکوک به تقلب، همزمان با حفظ و حتی بهبود خدمات، هزینه‌های سیستم بیمه سلامت به صورت قابل توجهی کاهش یابد.

۱-۵ چالش‌های این مبحث

تقلب و سوءاستفاده، که به موضوع بزرگی در راستای توسعه سیستم‌های اطلاعاتی تبدیل شده است، در حال مختل کردن صنایع زیادی است. صنایع بهداشت و درمان و مخابرات، مانند صنعت بانکداری، از تقلب و سوءاستفاده مکرر رنج می‌برد. البته مردم زیادی تقلب را با سوءاستفاده اشتباه می‌گیرند؛ این واژه‌ها نمی‌توانند با هم ترکیب شوند. تقلب به عنوان یک فریب عمدی یا ارائه اطلاعات نادرست تعریف می‌شود که توسط شخصی که میداند این فریب یا ارائه نادرست اطلاعات ممکن است سود غیرمجازی برای او یا شخص دیگری داشته باشد انجام می‌گیرد (راهنمای تقلب بهداشت و درمان امریکا، ۱۹۹۱). به طور مختصر، تقلب گفته‌ای غلط است که عمدا برای رسیدن به چیزی غیرمنصفانه و غیرقانونی، رواج داده شده است. درحالیکه سوءاستفاده به عنوان رفتاری متناقض و نامناسب با هدفی غیر قانونی تعریف می‌شود بدون اینکه لزوما عواقب قانونی داشته باشد.

هشتاد درصد هزینه بهداشت و درمان مربوط به تصمیم پزشکان درباره خدماتی است که بیماران نیاز دارند. بنابراین، تقلب و سوء استفاده رخ داده توسط پزشکان می‌تواند خیلی قابل توجه باشد[۴]. البته دلایل و انگیزه هایی وجود دارد که چرا پزشکان، قانون مربوط به تقلب و سوءاستفاده را زیر پا می‌گذارند. دیدی که پزشکان از فعالیت خود به عنوان کسب و کار دارند، میتواند نقشی حیاتی در ارتکاب به تقلب یا سوء استفاده ایفا کند. برای مثال، هزینه صدور صورتحساب میتواند انگیزه بزرگی برای پزشکانی باشد که خودشان را به عنوان یک فروشنده می بینند. پزشکان می توانند اقدامات غیرضروری برای افزایش هزینه ها انجام دهند. اگرچه این روش ها بر سابقه پزشکی بیمار اثر میگذارد و آن را تحریف می کند و ممکن است منجر به درمان اشتباه در آینده شود. از طرف دیگر، پزشکان ممکن است در شرایط دشواری بین انتخاب تعهد حرفه‌ای در مقابل بیماران یا قوانین پوشش مندرج در قرارداد شان قرار گیرند. برای مثال، برخی پزشکان ممکن است در شرایط بیمار اغراق کنند یا درخواست آزمایشی را بکنند که نشان دهد این دارو یا درمان برای بیمار ضروری است، تا در بدست آوردن پوشش اضافه به آنها کمک کنند[۴].

اگرچه تشخیص سندسازی و تقلب در بیمه حیاتی و به شدت مورد نیاز است، چالش‌ها و محدودیت‌های زیادی هستند که این کار را سخت می‌کنند. اول، تشخیص تقلب و سوء استفاده از بیمه سلامت نیازمند کارشناسانی است که از دانش پزشکی در سطح بالایی برخوردار باشند[۶]. بیشتر شرکت‌های بیمه از روش‌هایی استفاده می‌کنند که برای تشخیص فعالیت‌های متقلبانه یا سوءاستفاده‌گرانه بالقوه، نیازمند نیروی انسانی جهت ارزیابی مدارک است. این روش‌ها که مبتنی بر دانش افراد خبره است، نیاز به کارکنان خبره‌ای دارد که به اندازه کافی در دسترس نیستند. به علاوه، تکنیک‌های تشخیص دستی تقلب، به تلاش، زمان و تخصص انسانی زیادی نیاز دارد که منجر به تاخیر در اثبات یا رد صورتحساب می‌شود. علاوه بر این، آزمایشات و تشخیص‌های دستی بسیار هزینه بر هستند. با استفاده از تکنیک‌های اتوماتیک هر مورد با قوانین ساده‌ای که برای تست استفاده می‌شوند، کشف می‌شود. اگرچه، تشخیص تقلب و سوءاستفاده مستلزم بررسی متغیرها و ابهامات زیادی است. این ابهامات به فناوری اطلاعات دقیق و جامعی نیاز دارد تا بتواند صحت صورتحساب را آزمون کند[۷]. با اینکه صورتحساب‌های پزشکی و مستنداتی که الکترونیکی ارائه شده‌اند کشف تقلب را ساده تر می‌کنند، اما چالش‌های دیگری نیز وجود دارد. برای مثال، ارائه‌دهندگان خدمات بهداشتی و درمانی و بیمارستان‌ها انتظار دارند شرکت‌های بیمه به صورتحساب‌های ارائه شده از سوی آن‌ها پاسخی سریع بدهند. حقیقت این است که سرعت عمل در پردازش صورتحساب در شرکت‌های بیمه احتمال اشتباه را بالا میبرد، و باعث می‌شود برخی صورتحساب‌های متقلبانه کشف نشوند. چالش دیگر تشخیص تقلب این است که در روش‌ها و نظارت‌های کنونی، داده‌هایی که نیازمند تحلیل هستند، می‌توانند نسبت به هر تغییری حساس باشند. این نوسانات و بی‌ثباتی در سیستم بیمه سلامت، مانع از بررسی صورتحساب‌های بیمه میشود. از این رو، کشف رفتارهای مشکوک در بهداشت و درمان نیازمند تکنیک‌های انطباقی است[۸].

از دیگر چالش‌های این تحقیق می‌توان به موارد زیر اشاره کرد:

- اکثر فرم‌ها و پرونده‌ها بصورت غیرسیستمی و دستی ذخیره شده‌اند و تعداد کمی از آن‌ها بصورت الکترونیکی ذخیره و نگه داری می‌شوند و جستجوی دستی در این اسناد بسیار زمان‌بر است.
- شرکت های بیمه اطلاعات و پرونده های مشتریان خود را به آسانی در اختیار افراد خارج از سازمان قرار نمی‌دهند و اخذ مجوز برای دسترسی به این اطلاعات فرآیندی زمانبر و مشکل است.
- دسترسی به تعداد اندکی پرونده که وقوع تقلب در آن‌ها محرز شده است بسیار مشکل است زیرا این پرونده ها اغلب محرمانه هستند.

- کسب دانش مورد نیاز از افراد خبره و کارشناسان کشف تقلبات بیمه‌ای مشکل است زیرا اکثر آن‌ها مدیران بخش بیمه هستند و به ندرت وقت آزادی برای قبول و انجام مصاحبه دارند[۸].

۱-۶ ساختار تحقیق

در فصل دوم این پایان نامه هر آنچه از مبانی نظری برای درک و فهم این پروژه لازم است، مطرح می‌شود. مفاهیم کلی در مورد سیستم های پزشکی و اهمیت اطلاعات آنها مطرح می شود.

در فصل سوم مروری به تقلب در این سیستم ها داشته ومطالب مرتبط با آن، عواقب این موضوع، راه حل های پیشین برای حل این مساله وچالش های پیش رو مطرح می گردد.

در فصل چهارم این پایان نامه به بررسی روشی پرداخته خواهد شد که به عنوان راه حلی برای پیش بینی این تقلب‌ها یا ناهماهنگی‌ها در سیستم ارایه گردیده است. این روش مبتنی بر نظریه گراف است و سعی دارد علاوه بر حل چالش‌های موجود، عملکرد روش‌های مبتنی بر گراف را در حل این موضوع هم بررسی کند. بعد از مدل سازی راه حل مطرح شده، به تشریح و توضیح آن پرداخته خواهد شد. بخش‌های مختلف مدل مطرح شده و شرح آن ها نیز بیان می شود.

در فصل پنجم پس از پیاده سازی روش مطرح شده در فصل چهارم، به بررسی نتایج این روش با سایر روشهای موجود پرداخته می شود. برای مقایسه بهتر از مجموعه داده های استانداردی استفاده شده که در این موضوع بسیار شناخته شده و لذا نتایج الگوریتم‌های مختلف روی این مجموعه داده موجود است.

در نهایت در فصل ششم، به جمع بندی نتایج و دلیل ارائه این روش و همچنین پیشنهادات آتی‌ای پرداخته خواهد شد که می‌توان در ادامه کار مورد بررسی قرار داد.

۱-۷ جمع‌بندی فصل

در این فصل پس از طرح کلی مسئله و بیان اهمیت موضوع مورد مطالعه، محدودیت‌های پژوهش تشریح شد. همچنین مقدمه‌ای به سیستم نظام سلامت و تقلب در پایگاه داده‌های پزشکی اشاره شد. در پایان نیز ساختار کلی پژوهش حاضر تشریح شد که در فصل‌های آتی در مورد هر یک به تفصیل بحث خواهد شد.

فصل ۲

۲ مفاهیم و پیشینه‌ی تحقیق

۲-۱ مقدمه

در این فصل به تفصیل مباحث و مبانی مربوط به سیستم سلامت، انواع تقلب، سست هنجاری‌های سیستم بیمه سلامت و روشهای داده کاوی که میتوانند

هزینه بهداشت و درمان با توجه به جمعیت، اقتصاد، جامعه، و تغییرات قانون به سرعت در حال افزایش است. این افزایش در هزینه‌های بهداشت و درمان بر دولت و سیستم‌های بیمه سلامت خصوصی تأثیر می‌گذارد. رفتارهای متقلبانه‌ی ارائه‌دهندگان بهداشت و درمان و بیماران با تحمیل هزینه‌های غیرضروری به مشکلی جدی

برای سیستم‌های بیمه تبدیل شده است. بنابراین، حوزه سلامت به یک منبع هزینه‌ای قابل توجه در بسیاری از کشورها تبدیل شده است. وسیع بودن حوزه سلامت و حجم زیاد مالی باعث شده تا حوزه سلامت یه یک هدف جذاب برای کلاهبرداری تبدیل شود. شرکت‌های بیمه روش‌هایی را برای تشخیص تقلب ایجاد می‌کنند که عمدتا برگرفته از تجارب خبرگان بوده و کمتر به روش‌های مبتنی بر تحلیل داده متکی است. صنعت بیمه و در راس آن بیمه سلامت با هزینه ای بالغ بر ۵۰ هزار میلیارد تومان یکی از کلیدی‌ترین هزینه‌های تحت مدیریت و نظارت دولت است که با تخمینی معادل ۳ تا ۱۰ درصد یعنی ۱۵ تا ۵۰ هزار میلیارد ریال تقلب مواجه است.

سازمان‌های بیمه‌گر پس از دریافت اسناد هزینه از ارائه‌دهندگان خدمات سلامت اعم از مراکز درمانی، پزشکان، داروخانه‌ها و آزمایشگاه‌ها، فرآیند بررسی هزینه‌ها و تطبیق آن‌ها با معیارها و جداول هزینه شده توسط آن را، که به آن رسیدگی به اسناد می‌گویند، آغاز می‌نمایند. این رسیدگی هم‌اکنون به‌صورت دستی و توسط افراد خبره صورت می‌گیرد. استفاده از روش‌های تحلیل داده‌های بزرگ نظیر داده‌کاوی به ذینفعان کمک می‌کند تا بتوانند ضمن تعمیم و بهره‌برداری از الگوهای شناخته شده جهت بکارگیری الگوریتم‌های با ناظر به کشف الگوهای ناشناخته از طریق به‌کارگیری الگوریتم‌ها و مدل‌های نظارت نشده، بپردازند.

سامانه‌های مراقبت سلامت در سراسر جهان مجموعه ای از افراد، سازمان‌ها و منابع هستند که برای رفع نیازهای درمانی جمعیت هدف تشکیل می‌شوند و درحال تغییر و توسعه هستند. اغلب آن‌ها از سه بخش به نام‌های ارائه‌دهندگان خدمات، بیمار و شرکت‌های بیمه تشکیل شده‌اند[۱].

۲-۲ تقلب

تقلب در بیمه یک مساله مهم و پرهزینه برای بیمه‌گذاران و شرکت‌های بیمه در تمام بخش‌های صنعت بیمه است. در سال‌های اخیر، تشخیص تقلب توجهات و نگرانی‌های بسیاری را به خود جلب کرده‌است. دیکشنری آکسفورد تقلب را به عنوان "ارتکاب فریب عمدی که منجر به سود شخصی یا مالی می‌شود" تعریف می‌کند. تقلب به شکل‌های بسیار گوناگونی رخ می‌دهد و همانطور که تکنولوژی‌های جدید و سیستم‌های اقتصادی و اجتماعی جدید فرصت‌های جدیدی را برای فعالیت‌های تقلب‌آمیز فراهم می‌کنند، تغییر می‌کند. تعیین خسارت کلی مشاغل به دلیل فعالیت‌های تقلب‌آمیز دشوار است. [۱۱] تقلب را این‌گونه توصیف می‌کند که منجر به سوءاستفاده از منافع یک سازمان می‌شود و نه لزوما منجر به یک پیامد قانونی مستقیم.

اگرچه که یک تعریف پذیرفته شده جهانی از تقلب مالی وجود ندارد، [۱۲] آن را بعنوان یک عمل عمدی که در تضاد با قوانین و قاعده‌ها و سیاست و با هدف کسب منافع مالی غیرمجاز است، تعریف می‌کند.

۳-۲ سندسازی، تقلب و سوءاستفاده

راه‌های بیشماری برای تقلب و سوءاستفاده وجود دارد. همچنین ارتباطی قوی بین سندسازی، تقلب و سوءاستفاده وجود دارد. برای مثال بیشتر دلایلی که یک صورتحساب در بیمه رد میشود، این است که شاخص‌های مشکوک دارد. در این شرایط، بیمه‌گر از ارائه کننده خدمات سلامت یا بیمه شده می خواهد تا اطلاعات ارائه شده را تایید کند. بنابراین، تعیین و طبقه‌بندی دقیق این پارامترها حیاتی است. انواع تقلب‌های شناخته شده در جدول (۱) است[۴].

جدول ۱ انواع تقلب در بیمه سلامت[۴]

	انواع تقلب
۱	کدگذاری اشتباه خدمات درمانی
۲	صدور مجدد صورتحساب
۳	تجزیه یک فعالیت ترکیبی با کد واحد به فعالیتهای جزئی تر
۴	صورتحساب مواردی که تحت پوشش نیستند
۵	ارایه خدمات غیر ضروری
۶	عدم تطبیق تشخیص و درمان
۷	ارایه خدمات بیش از ظرفیت
۸	ارجاع منفعت طلبانه

کدگذاری اشتباه فعالیت‌ها، می‌تواند سرنخ‌هایی از تقلب و سوءاستفاده داشته باشد. کدگذاری فعالیت‌ها زمانی رخ می‌دهد که ارائه‌کنندگان خدمات بهداشتی و درمانی از کدی استفاده می‌کنند که گران‌تر از خدمات بهداشت و درمان، تست‌ها، یا آیتم‌هایی است که واقعا برای بیمار انجام شده است. برای مثال، کد ۹۹۲۱۱ برای یک مشکل پزشکی ساده و یک ویزیت کوتاه است که ۲۰ دلار هزینه دارد، در حالیکه کد ۹۹۲۱۵ نشان دهنده یک مشکل پیچیده و ویزیتی طولانی با هزینه ۱۴۰ دلار است. در نتیجه، چک کردن خطاهای صورتحساب مربوط به کدگذاری فعالیت‌ها برای کاهش هزینه بهداشت و درمان و جلوگیری از تقلب و سوءاستفاده، حیاتی است. از طرفی دیگر، بسیاری از پزشکان معتقدند که دقت در کدگذاری درست در صورت حساب به اندازه ویزیت بیمار زمان می‌برد و آنرا بهانه‌ای برای عدم دقت و بروز اشتباه می‌دانند. در ایران از سال ۱۳۸۴ اقداماتی در خصوص یکسان سازی نرخ خدمات درمانی شکل گرفته که نتیجه آن تولد کتاب ارزش نسبی خدمات و مراقبت‌های سلامت است که بر اساس فرآیندی با همین هدف از کشور امریکا اقتباس شده است[۶].

هر چند هدف کدینگ واحد پیگیری نمی‌شود ولی از نتایج مشخص آن رویکرد یکسان سازی کدینک و کاهش این گونه از تقلب‌ها می‌باشد. صدور مجدد صورتحساب، که به صدور دوباره صورتحساب برای یک فعالیت در یک زمان با تغییراتی کوچک گفته می‌شود، مانند تاریخ، هم می‌تواند یک اشتباه ساده باشد، هم می‌تواند یک سوء استفاده باشد. در هر صورت، ارزش بررسی مجدد و حذف را دارد. تجزیه یک فعالیت ترکیبی با کد واحد به فعالیت‌های جزئی‌تر به چندین کد جزئی‌تر، روشی دیگر برای افزایش هزینه و بدست آوردن منفعت غیر مجاز است. درمان‌ها یا آزمایش‌هایی وجود دارند که شامل بیش از یک خدمت است. وقتی این خدمات با هم انجام شوند، تامین کننده خدمات بهداشتی و درمانی نیاز به استفاده از کدهای مشخصی دارد که دو خدمت یا بیشتر را گروه بندی کند. اگر تامین کننده خدمات بهداشتی و درمانی از این کدهای صورتحساب مشخص، برای تمام خدمات اختصاص یافته استفاده نکند و به صورت مجزا آنها را صورتحساب کند، ممکن است پولی بیشتر از خدماتی که واقعاً انجام داده دریافت کند. برای مثال، تست کامل خون شامل آزمایش‌های زیادی مانند اندازه‌گیری آنزیم‌ها و مواد معدنی مختلف است. زمانی که این آزمایش‌ها جداگانه صورتحساب شود، نرخ پرداخت ممکن است دو برابر شود. ارائه صورتحساب برای مواردی که تحت پوشش بیمه نیست به جای موارد تحت پوشش نیز یکی از فعالیت های سوءاستفاده‌گرانه و دلیلی برای سندسازی است که مکرر دیده می‌شود، زیرا تامین‌کنندگان خدمات بهداشتی و درمانی موظف هستند بهترین مراقبت ممکن را پیشنهاد بدهند، بعضی اوقات ممکن است به خاطر سلامت بیمارشان، مواردی که تحت پوشش نیستند را به جای موارد تحت پوشش صورتحساب کنند.

پزشکان اغلب قوانین بازپرداخت را دستکاری می‌کنند تا به بیمارانشان کمک کنند تا برای خدمات ضروری در طرح درمان، پوشش لازم را بگیرند[۵].

انجام خدماتی که برای رفاه بیمار ضروری نیست، به عنوان مواردی که از نظر پزشکی ضروری نیست در نظر گرفته می‌شود. بیمه‌گر پوشش را فقط برای تشخیص و درمان خدمات قانونی، منطقی و ضروری از نظر پزشکی، فراهم می‌کند. صورتحساب‌ها یا صورتحساب‌های بیمه که شامل خدمات غیر ضروری است ممکن است منجر به رد صورتحساب شود یا نیاز به تحقیق داشته باشد که بفهمیم آیا تقلب یا سوء استفاده است یا خیر. زمانی که یک طرح درمان که نیازمند شرایط پیش نیاز است برای بیماری به کار برده می‌شود که شرایط پیش نیاز را ندارد، یک نشانه قرمز می‌تواند رفتار متقلبانه یا سوءاستفاده‌گرانه بالقوه را نشان دهد. گذشته از شرایط پیش نیاز، یک عدم تطابق بین تشخیص و طرح درمان می‌تواند نشانه یک رفتار مشکوک باشد. برای مثال، تشخیصی که نیاز به داروی خاص برای بیمار ندارد ممکن است نشان دهنده تقلب یا سوء استفاده بالقوه باشد.

نسبت برخورد غیر معمول با بیمار، پارامتر دیگری برای تخمین ریسک تقلب و سوء استفاده است. برای مثال، اگر پزشکی هر روز تعداد زیادی از بیماران را ببیند که بیشتر از میزانی است که او می توانسته بپذیرد، اثبات کننده‌ی مراقبت ضعیف او از بیمارانش یا ارتکاب به تقلب باشد. یک طرح درمان ناکافی که به پزشکی که بیمارانی بیشتر از حد توانش را میبیند اختصاص یافته است، بینشی نسبت به رفتار پزشک می‌دهد. علاوه بر این، بیمارستان‌هایی که تعداد پزشکانی که استخدام کرده‌اند را بیشتر از تعداد واقعی گزارش می‌دهند، تقلب کرده‌اند، زیرا ارائه اطلاعات نادرست نیز تقلب است. ارجاع منفعت طلبانه، معرفی بیماران به پزشکی خاص یا ارائه‌دهنده خدمات بهداشتی و درمانی خاص است. برای مثال، اگر یک پزشک منفعتی شخصی از یک کلینیک داشته باشد، نمی‌تواند هیچ بیماری را به آن کلینیک ارجاع دهد. در بعضی از کشورها از جمله امریکا قانونی برای مقابله با این امر وجود دارد. در ایران اشتراک منافع پزشکان با داروخانه‌ها و آزمایشگاه‌ها و بیمارستان‌ها به تناسب قرارداد سازمان‌های بیمه‌گر ممکن است با جرایمی همراه باشد. به صورت خالصه، قرارداد مقابله با ارجاع منفعت طلبانه زمانی نقض می‌شود که ارائه‌دهنده خدمات بهداشتی و درمانی بیماران را به جایی که ارتباط مالی با آن دارد ارجاع دهد. این معرفی‌ها توسط قوانین یا قراردادهای ضد ارجاع منفعت طلبانه ممنوع شده‌اند و در صورت رخ دادن تقلب محسوب می‌شوند.

۲-۴ بازیگران نظام سلامت

ارتباط میان بازیگران[1] مختلف سیستم سلامت به واضح‌ترین شکل ممکن در شکل (۲) ارائه شده است که موارد سواستفاده بین دو یا تعداد بیشتری بازیگران را شامل می‌شود. برای مثال با دیدن بیمارستان و تعیین اینکه چه نوع سواستفاده‌هایی می‌تواند در ارتباط میان آن و سایر ذینفعان، بیماران و سایر بیمارستان‌ها صورت گیرد، یک تحلیل می‌تواند انجام شود[۱۷].

کلاهبرداری مراقبت‌های بهداشتی موضوعی خاص برای هر کشور است و رفتارهای کلاهبردانه، متفاوتی به همین نسبت تغییر می‌کند. با این وجود، انواع کلاهبرداری‌هایی که در زمینه مراقبت‌های بهداشتی انجام می‌شود، تقریباً برای همه کشورها رخ می‌دهد. همانطور که در شکل (۲) دیده می‌شود، چهار دسته اصلی در تقلب در مراقبت‌های بهداشتی نقش دارند. این دسته‌ها ارائه‌دهندگان خدمات هستند، که شامل پزشکان، شرکت‌های آمبولانس بیمارستان و آزمایشگاه‌ها، مشترکین بیمه (که شامل بیماران و کارفرمایان بیماران می شود)، و شرکت های بیمه‌ای که ادارات درمان و بهداشت دولتی و شرکت های بیمه خصوصی را در بر می‌گیرند. براساس اینکه کدام طرف مرتکب تقلب می شود، رفتارهای تقلب آمیز در ادامه دسته‌بندی و توضیح داده می‌شود. بر اساس اینکه چه کسی مرتکب تقلب می‌شود، رفتارهای تقلب آمیز به صورتی که در ادامه آمده‌اند دسته‌بندی می‌شوند[۳]:

تقلب ارائه‌دهنده‌ی خدمات[2]

شامل تخلفات ارائه‌دهندگان خدمات می‌شود که می‌تواند به عنوان مثال، جعل تشخیص و یا سابقه معالجه برای توجیه آزمایشات، جراحی‌ها یا سایر فرآیندهایی که از نظر پزشکی غیرضروری هستند، باشد. با توجه به انواع تقلب، اکثر مطالعات تا کنون برای کشف تقلب ارائه‌دهندگان خدمات مورد استفاده قرار گرفته است. از آنجا که کشف تقلب در ارائه‌دهندگان خدمات مسئله مهمی در جهت ارتقاء کیفیت و ایمنی سیستم مراقبت‌های بهداشتی است، بسیاری از محققان به این افراد توجه کرده‌اند.

---

[1] Actor

[2] Service Provider’s Fraud

تقلب مشترکان بیمه

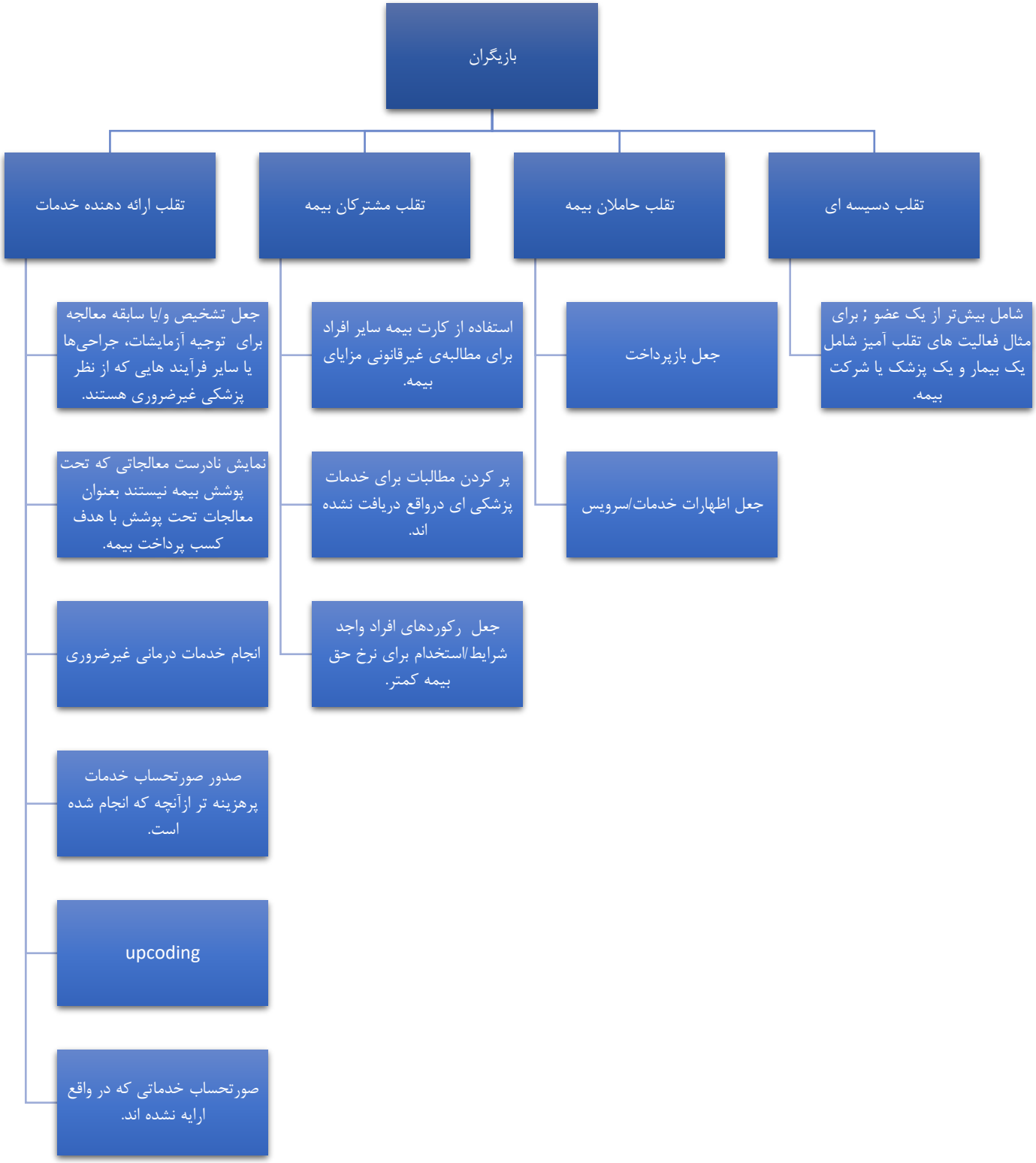
شامل افرادی می‌شود که خدمات درمانی‌ای که ادعا می‌کنند را دریافت نکرده‌اند و یا از اعتبار بیمه افراد دیگر به صورت غیر قانونی استفاده کرده‌اند و یا اسنادی را برای شرایط استخدام به منظور پرداخت حق بیمه کمتر جعل کرده‌اند.

تقلب حاملان بیمه

شامل افرادی می‌شود که جعل در بازپرداخت یا جعل در اظهاراتشان نسبت به سود یا زیان‌دهی کرده‌اند.

تقلب دسیسه‌ای

شامل تخلفاتی می‌شود که در بیشتر از یکی از سه دسته دیگر درگیر آن هستند (برای مثال: یک دکتر و یک بیمار).



شکل ۲ درخت دسته‌بندی بازیگران نظام سلامت

۲-۵ دسته‌بندی چالش‌های کشف تقلب

همانطور که اشاره شد، چالش‌ها و محدودیت‌های زیادی هستند که کار تشخیص سندسازی و تقلب در بیمه را سخت میکنند. این چالش‌ها از منظرهای سازمان بیمه، مجموعه داده و مقالات قابل بررسی هستند که در ادامه به هر کدام از آن‌ها به تفضیل پرداخته می‌شود.

۲-۵-۱ چالش‌های کشف تقلب

شرکت های بیمه می توانند با آگاهی از انواع تقلبات و فرآیندهایی که احتمال بروز تقلب در آن‌ها وجود دارد سیستم هشدار دهنده و پیشگیرانه‌ای را طراحی کنند و با آگاهی از میزان آسیب‌پذیری خود استراتژی‌های موثرتری را به‌کار گیرند، اما برای تحقق این امور شرکت‌های بیمه با محدودیت‌ها و پیچیدگی‌های زیادی مواجه‌اند[۱۸]:

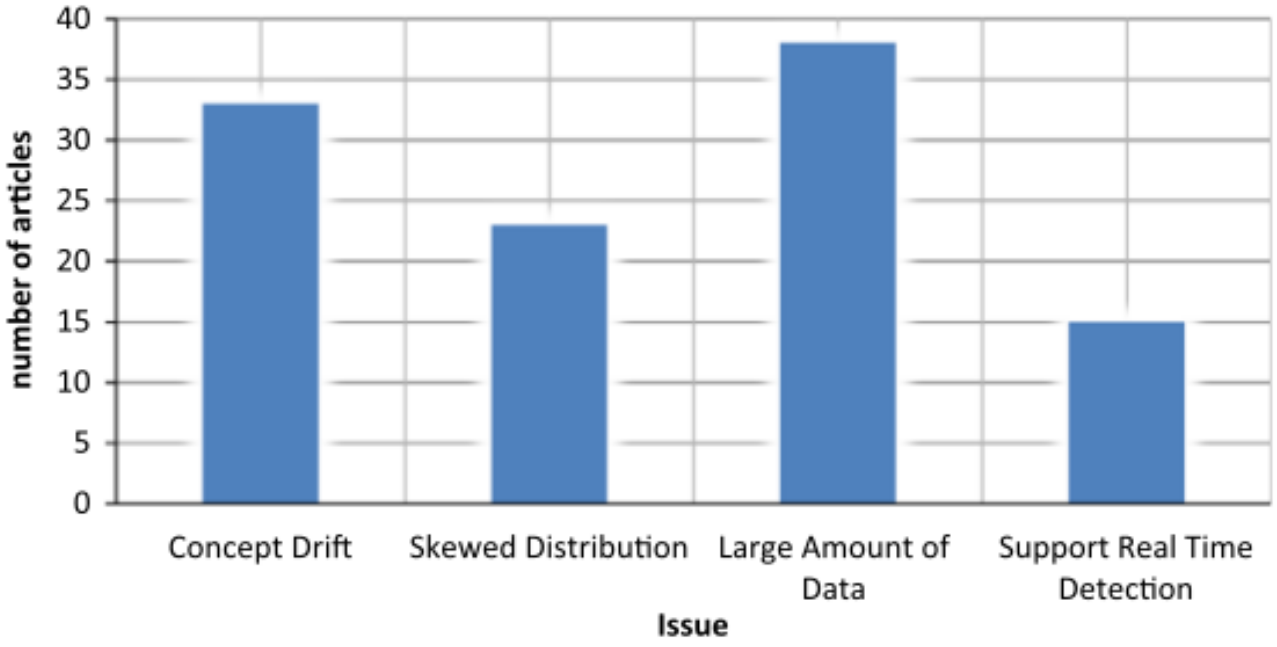
- پنهان بودن ماهیت تقلب
- پویایی و حساسیت به تغییر در تقلب (هنگام شناسایی یک سبک تقلب، کلاهبرداری با سبک دیگری در حال شکل‌گیری است)
- عدم توافق اجماع در برخی موارد بر آنچه که واقعاً به منزله‌ی تقلب در بیمه استاندارد.
- نگاه جامعه به شرکت‌های بیمه‌ای و ارائه‌دهندگان خدمات درمانی بعنوان نهادهای حمایتی
- ضعف سیستم‌های کنترلی مبتنی بر فناوری اطلاعات

۲-۵-۲ چالش‌های کشف تقلب از منظر داده

از چالش های موجود در کشف تقلب عدم وجود داده‌های با برچسب سالم و وجود تعداد بسیار کم داده‌های تقلبی جهت یادگیری است. بصورت طبیعی تعداد تقلب کم است ولی تعداد دفعاتی که ثبت شده و کدگذاری شده‌است کم‌تر است. لذا داده‌هایی آماده نشده با تعداد کمی برچسب و تعداد زیادی نمونه‌های نامشخص وجود دارد.

۳-۵-۲ چالش‌های کشف تقلب از منظر مقالات

شکل (۳) توزیع مطالعات سیستم‌های تشخیص تقلب را بر اساس چالش‌ها بر اساس تعداد مقالات منتشر شده در سال‌های ۱۹۹۴ تا ۲۰۱۴ نشان می‌دهد و به رایج‌ترین انواع تقلب‌های الکترونیکی مانند کارت‌های اعتباری، بیمه مراقبت سلامت، مخابرات، بیمه اتومبیل متمرکز است[۱۹].



شکل ۳ توزیع مقالات FDS براساس مسائل و چالش های بین سال های ۱۹۹۴ تا ۲۰۱۴

۲-۵-۴ مفاهیم پایه در پژوهش‌های کشف تقلب در سیستم سلامت

مفهوم رانش[۳]

تعاریف مختلفی برای مفهوم مسئله رانش وجود دارد. در داده‌کاوی رانش به پدیده‌ای که مدل پایه‌ی آن در طول زمان درحال تغییر است اشاره دارد. کار سیستم‌های تشخیص تقلب در محیط پویا که رفتار کاربران قانونی/غیرقانونی بطور پیوسته در حال تغییر است مفهوم پدیده رانش گفته می‌شود[۲۰].

برای مثال در حوزه کارت اعتباری رفتار صاحب کارت ممکن است به دلیل برخی عوامل خارجی تغییر کند. برای مثال مقدار تراکنش و تکرار به عادات خرج یک فرد وابستگی نزدیکی داشته باشد که درواقع تحت تأثیر شیوه زندگی، منبع درآمد فرد و ... است که در طول زمان می‌تواند تغییر کند[۲۱].

به‌علاوه مفهوم رانش سابقاً به یک سناریوی یادگیری با ناظر زمانی که رابطه بین داده ورودی و متغییر هدف در طول زمان تغییر می‌کند، ارجاع داده می‌شد. اگرچه در یادگیری با ناظر هدف پیش‌بینی یک متغییر هدف y با استفاده ازمجموعه ای از ویژگی‌های ورودی x است. در نمونه یادگیری که برای ساخت مدل استفاده می‌شود هردوی x وy در زمان پیشگویی ناشناخته‌اند و رابطه بین داده ورودی و متغییر هدف ممکن است تغییر کند[۲۲].

مفهوم رانش یک نگرانی بزرگ است، مخصوصاً در یادگیری آنلاین که مدل تشخیص فوراً به روز می‌شود، اما براساس داده‌های خروجی. بنابراین وقتی داده‌های جدید می‌رسند، مدل ممکن است گمراه شود و اخطار اشتباه دهد. توجهات در تحقیقات به مقابله با رفتار غیرایستا و بطور پویا به روز رسانی مدل تشخیص تقلب، اختصاص یافته است و در نتیجه استفاده از الگوریتم های یادگیری تطبیقی۴ برای مقابله با مفهوم رانش لازم است. الگوریتم‌های یادگیری تطبیقی می‌توانند بعنوان الگوریتم های یادگیری افزایشی توسعه یافته دیده شوند که قادر به بروزرسانی مدل تشخیص برای داده‌ی در جریان تکامل، در طول زمان می‌باشند[۲۲]،[۲۳].

[۲۴] مفهوم یادگیری افزایشی با رانش را این‌گونه بیان می‌کند رابطه (۲-۱): فرآیند یادگیری افزایشی در هر زمان t که داده‌های قبلی در دسترس هستند، یک نمونه‌ی هدف xt+۱ میرسد وظیفه‌اش این است که برچسب

---

[۳] Drift

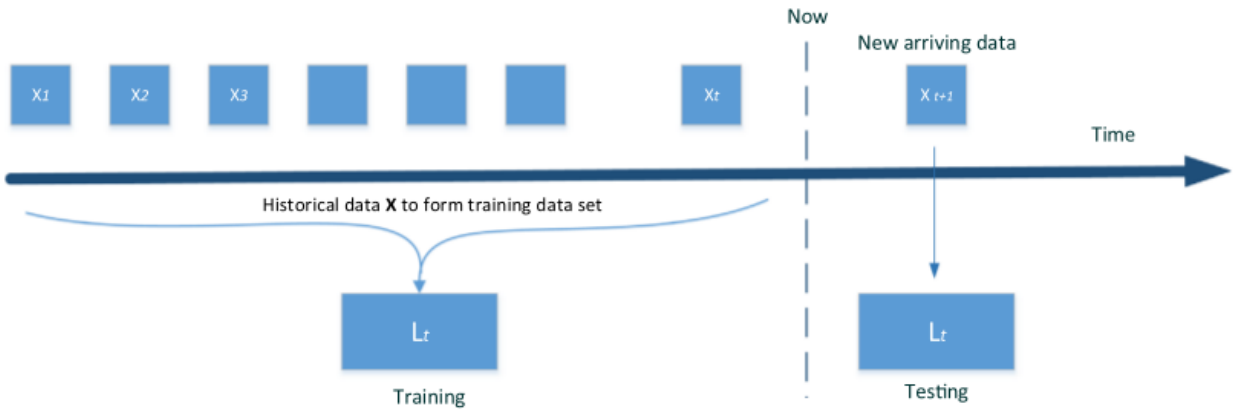
[٤] Adaptive

y_{t+1} را پیش‌بینی می‌کند. بدین منظور یادگیرنده L_t در فاز یادگیری با استفاده از همه یا انتخاب از داده‌های قبلی برچسب دار ساخته می‌شود.

رابطه (۲-۱)

$$X_{historical} = (x_1 , x_2 , \ldots, x_t)$$

که این را در شکل (۴) می‌بینیم:



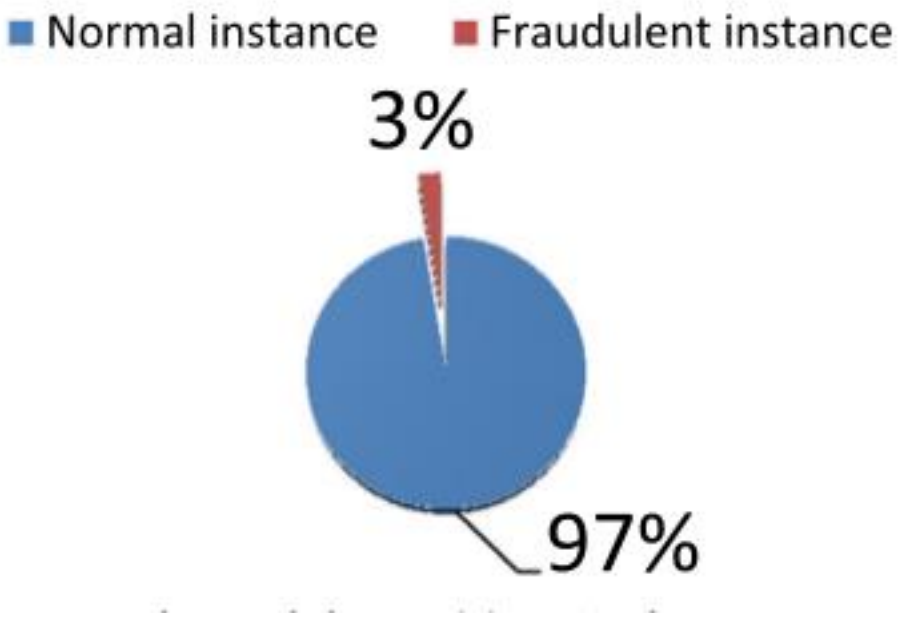
شکل ٤ فرآیند یادگیری افزایشی در هر زمان t

مفهوم توزیع اریب کلاس‌ها[۵]

مسئله‌ی داده‌های اریب یکی از مهم‌ترین مسایلی است که در سیستم‌های تشخیص تقلب با آن مواجه ایم. نامتوازن بودن داده‌ها، تأثیری جدی روی کارایی طبقه بندهایی که قرار است توسط اکثریت کلاس سراسر پوشانیده شوند و اقلیت کلاس نادیده گرفته شوند دارد[۲۵].

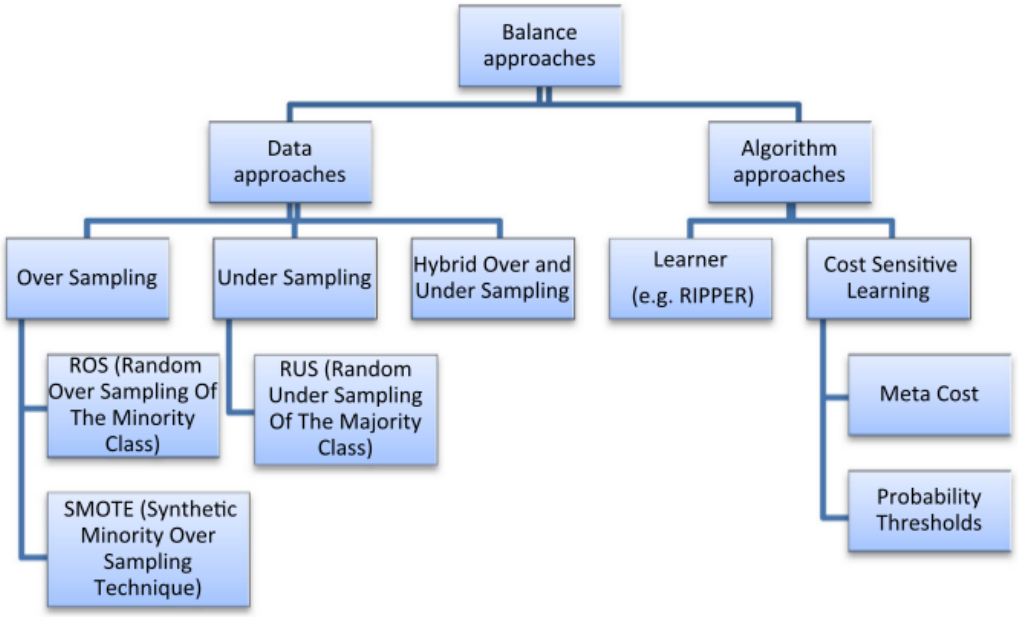
[۵] Skewed class distribution

داده‌ی رقابت‌های داده‌کاوی UCSD[۶] برای تشریح مساله نامتوازن بودن استفاده شده و یک مجموعه داده از دنیای واقعی است که برای تشخیص تراکنش های تجارت الکترونیک غیرعادی استفاده شده‌است. مجموعه داده‌ی آموزش شامل ۱۰۰۰۰۰ تراکنش از ۷۳۷۲۹ مشتری در طی ۹۸ روز است و مجموعه داده تست شامل ۵۰۰۰۰ تراکنش است. داده‌ی آموزشی به شدت نامتوازن است که شامل ۹۷۳۴۶ تراکنش طبیعی و فقط ۲۶۵۴ تراکنش تقلبی است. همانطور که در **Error! Reference source not found.** (۵) می‌بینیم درصد ت راکنش ها درحدود ۹۷ به ۳ درصد قانونی و جعلی می‌باشد. بنابراین یک مکانیزم متوازن‌سازی نیاز است تا این داده‌ها را با نرخ ۱:۱ میان طبیعی و جعلی متوازن کند. روش‌های متوازن‌سازی داده‌ها در دو سطح می‌تواند طبقه بندی شود: سطح داده و سطح الگوریتم، که تکنیک‌های آن در شکل (۶) نشان داده شده است.



شکل ۵ مجموعه داده آموزشی ناهمگون UCSD [۱۹]

University of California, San Diego6



شکل ۶ روش‌های رسیدگی به داده‌های ناهمگون[۲۶]

مقیاس بزرگ و ابعاد زیاد مجموعه داده‌ی تقلب و حضور تعداد زیاد ویژگی‌ها/ورودی ها/متغییرها فرآیند داده‌کاوی و تشخیص را بسیار دشوار و پیچیده می‌سازد[۲۷]. بعلاوه این شرایط فرآیند تشخیص را نیز کند میکند. بنابراین سیستم‌های تشخیص تقلب موجود از روش‌های کاهش داده برای کاهش حجم مجموعه داده‌ها استفاده می‌کند[۲۸]. بعلاوه داده‌ی کم‌حجم مدل و در نتیجه زمان محاسبه را کاهش می‌دهد[۲۹]. روش‌های کاهش داده شامل کاهش ابعاد(Dimension Reduction) و کاهش عددی(Numerosity Reduction) است[۳۰].

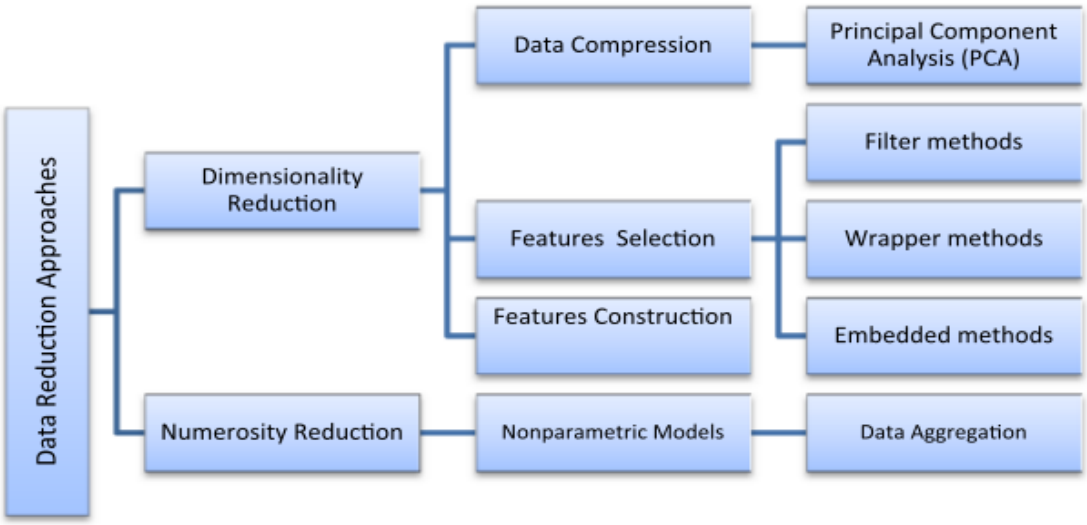
کاهش ابعاد شامل استراتژی‌های بسیاری است به نام‌های فشرده‌سازی داده(data compression)، انتخاب داده(feature selection)،ساخت ویژگی(feature construction).مرسوم‌ترین و پرتکرارترین استراتژی‌های استفاده شده در سیستم‌های تشخیص تقلب هستند. استراتژی فشرده‌سازی داده، از طریق استفاده از تکنیک‌های فشرده‌سازی مانند [۳۱] ,[۳۲] نمایش داده‌ی اصلی را فشرده می‌کند. در این میان انتخاب ویژگی یک استراتژی دیگر کاهش ابعاد است. مهمترین و مرتبطترین ویژگی‌های انتخاب می‌شوند تا در ساخت مدل استفاده شوند. انتخاب ویژگی توسط [۳۳] نام گذاری شد.

سه روش انتخاب ویژگی که در سیستم‌های تشخیص تقلب استفاده می‌شوند:

۱) روش‌های تصفیه ۲) روش‌های بسته‌بندی ۳.روش‌های جاسازی

- ✓ روش تصفیه بعنوان الگوریتم پیش‌پردازش برای رتبه بندی ویژگی‌ها عمل می‌کند که ویژگی‌های با رتبه بالا انتخاب و به یک پیش‌بینی کننده اعمال می‌شوند.
- ✓ در کاهش عددی نیز داده‌ها با نمایش کوچکتری جایگزین می‌شوند. مانند استفاده از تجمیع داده ها[۷][۲۵], [۲۸].
- ✓ روش جاسازی شامل انتخاب متغییر بعنوان بخشی از فرآیند یادگیری بدون تقسیم داده به مجموعه‌ی آموزش و تست می‌باشد[۳۴]. ساخت ویژگی در جایی است که یک مجموعه کوچک از ویژگی‌های مفیدتر از مجموعه‌ی اصلی مشتق می‌شود.

روش‌های کاهش داده شامل کاهش ابعاد و کاهش ارقام در شکل (۷) آورده شده‌اند:



شکل ۷ روش‌های کاهش داده [۳۰]

۷ data aggregation

مفهوم پشتیبانی تشخیص آنلاین

سیستم‌های تشخیص تقلب در دو حالت متفاوت کار می‌کنند که تشخیص آنلاین و آفلاین است که بر اساس انواع تقلب متفاوت است و هریک کاربرد خاص خود را دارند[۱۹].

ادغام در نظام مراقبت‌های بهداشتی

تعداد بسیار کمی از مقالات مورد بررسی تلاش برای ادغام فرآیند داده‌کاوی را در چارچوب تصمیم‌گیری واقعی انجام دادند. تأثیر کشف دانش توسط داده‌کاوی بر میزان کاری و زمان کار حرفه‌ای مراقبت‌های بهداشتی مشخص نیست. مطالعات آینده باید یکپارچگی سیستم توسعه‌یافته را در نظر بگیرند و تاثیر آن بر محیط کار را بررسی کنند[۳۵].

خطای پیشبینی و اثر "Black Swan"

در مراقبت‌های بهداشتی، پیشبینی بهتر از پیشبینی اشتباه بهتر است[۳۶]. کمی کمتر از نیمی از ادبیات که در تجزیه و تحلیل شناسایی شده است، به پیشبینی اختصاص یافته، اما هیچ یک از مقالات درباره نتیجه خطای پیشبینی بحث نشده است. دقت پیش بینی بالا برای سرطان یا هر بیماری دیگر، برنامه صحیحی را برای تصمیم گیری فراهم نمی‌کند.

علاوه بر این، مدل پیش بینی ممکن است در پیش بینی رویدادهای عادی بهتر از موارد نادر باشد. محققان باید مدل های پیشرفته ای را برای رسیدگی به غیرقابل پیش‌بینی"The Black Swan"توسعه دهند[۳۷].

یک مطالعه [۳۸] یک مسئله مشابه را در توصیه‌های مبتنی بر شواهد برای تجویز پزشکان مطرح کرد. نگرانی آنها این بود که چه مقدار شواهد باید برای تهیه یک توصیه کافی باشد.

بسیاری از مطالعات در این بررسی این مسائل برجسته را رفع نمی کند. پژوهش‌های آینده باید به چالش‌های پیاده‌سازی مدل‌های پیشبینی‌کننده بپردازد، به ویژه اینکه چگونه فرآیند تصمیم‌گیری باید در صورت اشتباهات و حوادث غیرقابل پیشبینی سازگار شود.

از دست دادن اطلاعات در پیش پردازش

پیش پردازش داده‌ها، از جمله دستکاری داده‌های از دست رفته، پرهزینه‌ترین و مهم‌ترین بخش داده کاوی است. شایع ترین روش مورد استفاده در مقالات مورد بررسی حذف یا حذف داده های از دست رفته است. در یک مطالعه، حدود ۴۶.۵٪ از داده ها و ۳۶۳ از ۴۱۰ ویژگی به دلیل مقادیر گم شده حذف شدند[۳۹]. در یکی دیگر، محققان [۴۰] تنها قادر به استفاده از ۲۰۶۴ از ۴۹۴۸ مشاهدات (۴۲٪) بودند. با حذف مقادیر از دست رفته و داده‌های پرت، ما مقدار قابل توجهی از اطلاعات را از دست می دهیم. پژوهش‌های آینده باید بر روی یافتن یک روش بهتر تخمین مقادیر از دست‌رفته نسبت به حذف تمرکز نمایند. علاوه بر این، تکنیک‌های جمع آوری داده‌ها باید به منظور جلوگیری از این موضوع توسعه یابند یا اصلاح شوند.

خودکارسازی فرآیند داده‌کاوی برای کاربران غیرمتخصص

کاربران نهایی داده‌کاوی در مراقبت های بهداشتی، پزشکان، پرستاران و متخصصین مراقبت‌های بهداشتی هستند که آموزش‌های محدودی در زمینه تحلیلی دارند. یک راه حل برای این مشکل این است که یک سیستم خودکار (یعنی بدون نظارت انسان) برای کاربران نهایی ایجاد شود [۴۱]. یک ساختار خودکار مبتنی بر ابر برای جلوگیری ازخطاهای پزشکی نیز می تواند توسعه یابد [۴۲]؛ اما این کار چالش برانگیز خواهد بود زیرا در آن زمینه‌های کاربردی مختلف وجود دارد و یک الگوریتم دقت مشابهی برای تمام برنامه‌های کاربردی ندارد [۴۱].

ماهیت بین رشته ای تحقیق و دانش متخصص حوزه

تجزیه و تحلیل بهداشت و درمان یک زمینه تحقیقاتی بین رشته‌ای است[۴۱]. به عنوان یک روش تجزیه و تحلیل، داده‌کاوی باید از ترکیبی از نظر کارشناس از حوزه های خاص مراقبت های بهداشتی و مشکل مشخص(به عنوان مثال، انکولوژی برای تحقیقات سرطان و متخصص قلب برای CVD) استفاده کند[۴۳]. تقریبا ۳۲٪ از مقالات در تجزیه و تحلیل از نظر متخصص به هیچ شکلی شکل استفاده نمی شود. پژوهش های آینده باید شامل اعضای از رشته های مختلف از جمله مراقبت های بهداشتی باشد[۳۵].

۵-۵-۲ انواع ناهنجاری در تشخیص تقلب

ناهنجاری‌ها نمونه داده‌هایی هستند که به میزان قابل توجهی با سایر نمونه داده‌ها متفاوت و ناسازگار هستند[۴۴]. ناهنجاری‌ها همچنین پرت‌ها، اختلالات، مشاهدات غیرواقعی و استثنائات نیز نامیده می‌شوند[۴۵]. در تعریف دیگری، ناهنجاری را بعنوان مشاهده یا زیرمجموعه‌ای از مشاهدات میداند که تا حدی زیادی از دیگر مشاهدات متفاوت است[۴۶]. منشأ ناهنجاری‌ها میتواند رفتار کلاهبردارانه، خطای انسانی یا شکست سامانه‌ها باشد[۴۷].

ناهنجاری‌ها از چند نظر قابل دسته‌بندی هستند. از نظر ماهیت ناهنجاری‌ها به ۴ دسته نقطه‌ای، جمعی، زمینه‌ای و افقی تقسیم میشوند[۴۸]. زمانی که یک نمونه داده خاص الگوی معمول مجموعه داده را نقض کند، ناهنجاری نقطه‌ای به وجود می‌آید. ناهنجاری جمعی رفتار نامتعارف و غیر عادی جمعی از داده‌های مشابه نسبت به سایر نمونه‌های مجموعه داده است و رفتار غیرعادی یک نمونه داده در یک زمینه خاص با سایر نمونه‌های مجموعه داده یک ناهنجاری زمینه‌ای است. تشخیص این نوع از ناهنجاری نیاز به شناخت زمینه مورد نظر دارد و به همین دلیل ناهنجاری شرطی نیز نامیده می‌شود. از نظر نوع شبکه، ناهنجاری‌ها به یکی از دو دسته ایستا، پویا تقسیم می‌شوند[۴۹].

تشخیص ناهنجاری گراف ایستا

ناهنجاری‌ها به یکی از دسته‌های ایستا ساده، ایستای با ویژگی، پویای ساده و پویای با ویژگی تقسیم می‌شوند[۴۹]. در ناهنجاریهای ایستای بدون ویژگی، هر اطلاعاتی راجع به نوع تعامل، مدت زمان آن، سن افراد درگیر و غیره نادیده گرفته میشود و تنها تعامل اتفاق افتاده بین افراد قابل توجه است. در ناهنجاری‌های ایستای با ویژگی، علاوه بر ساختار شبکه، مشخصات مرتبط با افراد و تعامل بین آنها نیز در تشخیص ناهنجاری‌ها در نظر گرفته می‌شود.

تشخیص ناهنجاری مبتنی بر ساختار

دو نوع روش تشخیص ناهنجاری مبتنی بر ساختار به نام‌های ناهنجاری در گراف‌های ساده‌ی ایستا[۸] و ناهنجاری در گراف ویژگی ایستا[۹] وجود دارد که بصورت زیر تشریح می‌شوند:

در طرح تشخیص ناهنجاری در گراف ساده‌ی ایستا، ویژگی‌های مرکزی گراف‌های مختلف مانند درجه گره، مرکزیت egonet و ... استخراج می‌شوند و یک فضای ویژگی با بقیه ویژگی‌هایی که از منابع اطلاعاتی اضافی برای تشخیص تقلب استخراج شده‌اند ساخته می‌شود. در [۵۰] یک روش تشخیص ناهنجاری ارائه کرده‌اند که از شاخص‌های گراف برای شناسایی کاربران با روابط غیرعادی نسبت به سایر کاربران در شبکه اجتماعی آنلاین استفاده می‌کنند. آن‌ها از ویژگی‌های مختلف نظریه گراف مانند تعداد گره‌های همسایه و یال‌ها، betweenness centrality وcommunity cohesiveness برای تمایز رفتارهای آنلاین افراد توسط الگوهای مصرف آن‌ها استفاده نمودند. به‌علاوه دنبال کردن ارتباطات کاربران می‌تواند الگوهای معناداری را آشکار سازد. زیرا کاربران می‌توانند هویت خود را با اطلاعات اشتباه پنهان سازند اما ارتباطات میان یکدیگر را نمی‌توانند پنهان کنند. آن‌ها از شاخص‌های محلی مانند single node (ego) و(one-level neighborhood (an egonet و two-level neighborhood (a super egonet) وbetweenness centrality و average betweenness user's egonet برای شناسایی کاربران با ساختارهای ارتباطی ناهنجار استفاده کردند.

در طرح تشخیص ناهنجاری در گراف ویژگی ایستا در [۵۱] یک روش تشخیص مبتنی بر گراف به نام GBAD ارائه دادند که اساساً مبتنی بر این نظریه است که یک فرد سعی در ارتکاب یک عمل غیرقانونی یا غیرعادی را دارد، بنابراین از رفتارهای شناخته شده‌ای پیروی و قصد واقعی خود را پنهان می‌کند. این روش شامل سه الگوریتم مختلف GBAD-MDL و GBAD-MPL و GBAD-P است. الگوریتم GBAD-MDL زیرساخت هنجاری را با استفاده از اصل بیشینه طول توصیف (MDL[۱۰]) پیدا می‌کند و زیرساختارهای مشابه را با سطح قابل پذیرشی از تغییر از زیرساختار طبیعی جستجو می‌کند. الگوریتم GBAD-MPL نیز بهترین زیرساختار را با جستجو در یال‌ها و راس‌هایی که گم شده‌اند تعیین می‌کند. الگوریتم GBAD-P از روش ارزیابی MDL برای

---

۸ static plain graphs

۹ static attributed graph

10 Maximum Description Length

کشف بهترین زیرساختار در گراف استفاده می‌کند اما به‌جای امتحان کردن همه نمونه‌ها برای مشابهت، این روش همه‌ی بسط‌ها برای زیرساختارهای طبیعی را در جستجوی بسط با کمترین احتمال، امتحان می‌کند. نویسنده از این روش برای کشف کارمندان مشکوک و اعمال آن‌ها به‌عنوان یک ابزار برای پشتیبانی تحقیقات جرم استفاده نموده است.

تشخیص ناهنجاری مبتنی بر اجتماع

در [۵۲] یک روش تشخیص ناهنجاری مبتنی بر اجتماع ارائه دادند، با شناسایی اجتماعاتی که برای مرزهای اجتماعی اهمیتی قایل نیستند. این کار بر اساس یک نظریه تعلق گره‌های دارای سورفتار متمایل به چندین اجتماع است. نویسنده جداسازی اجتماعات را بهبود داده است که هر گره فقط به یک اجتماع تنها تعلق داشته باشد. در [۵۳] یک روش خوشه بندی متمرکز و تشخیص ناهنجاری در گراف‌ها به نام FocusCo ارائه نمودند. الگوریتم شامل سه گام است ۱. استنتاج وزن‌های ویژگی‌ها ۲. استخراج خوشه‌های گراف‌های ویژگی متمرکز ایستا ۳. تشخیص ناهنجاری.

به‌طور مختصر هدف این است که خروجی مجموعه‌ای از گره‌های ارائه شده توسط کاربر که مربوط به ویژگی‌های متمرکز هستند توافق کنند. در این روش یک خوشه از گره‌های متصل به هم به نام خوشه‌های متمرکز، با توجه به ویژگی‌های متمرکز یافت می‌شود و بر اساس خوشه‌های متمرکز یک ناهنجاری به‌عنوان گره‌ای که از نظر ساختاری متعلق به خوشه هست اما انحراف زیادی در ویژگی‌های متمرکز دارد. آن‌ها همچنین نشان دادند که این روش برای گراف‌های ساختگی و واقعی بسیار مؤثر و مقیاس پذیر است.

تشخیص ناهنجاری گراف پویا

گراف‌های دنیای واقعی به‌طور مداوم در حال تغییرند. تشخیص ناهنجاری در این نوع از گراف‌های پویا کاری بسیار چالش برانگیز است. ناهنجاری‌ها در گراف پویا به یکی از دسته‌های پویای مبتنی بر فاصله، پویای مبتنی بر فشرده‌سازی ، پویای مبتنی بر تجزیه، پویای مبتنی بر خوشه یا اجتماع، پویای مبتنی بر مدل‌های احتمالاتی و پویای مبتنی بر پنجره تقسیم می‌شوند.

تشخیص ناهنجاری مبتنی بر فاصله

معیار مبتنی بر فاصله می‌تواند برای اندازه‌گیری تغییر بین دو شی به‌کار رود. دو شی که در معیار اندازه اختلاف کمی دارند، یکسان نامیده می‌شوند. معیارهای مختلفی برای تشخیص ناهنجاری وجود دارد. فاصله‌ی خطای اصلاح تطابق گراف[۱۱]، بیشینه زیرگراف مشترک[۱۲]، فاصله ماتریس همسایگی[۱۳]، فاصله ویرایش گراف[۱۴]، فاصله همینگ برای ماتریس‌های همسایگی گراف‌ها و ... [۵۴].

تشخیص ناهنجاری مبتنی بر فشرده‌سازی

در این فرآیند یک نمایش گراف فشرده با استفاده ازحداقل طول توصیفی و روش فشرده‌سازی با بهره‌گیری از الگوها و تنظیمات داده‌ها با کمترین هزینه‌ی رمزگذاری به‌دست می‌آید. سپس ناهنجاری‌ها به‌عنوان گراف‌هایی که مانع فشرده‌سازی هستند تعریف می‌شوند[۵۴].

تشخیص ناهنجاری مبتنی بر تجزیه

این روش ناهنجاری‌های موقتی را با نمایش مجموعه‌ای از گراف‌های تکامل زمانی[۱۵] بعنوان یک تنسور یا آرایه‌ی چند بعدی تشخیص می‌دهد و factorization یا کاهش بعد انجام می‌دهد. یک روش جدید تجزیه ماتریس فشرده[۱۶] برای محاسبه تقریب‌های کم مرتبه خلوت در [۵۵] ارائه شده است. خطای بازسازی هرگراف خلوت در طول زمان پیگیری شده و در جایی که تغییر زیاد باشد، گراف منطبق ناهنجاری خواهد بود.

---

۱۱ Error correcting graph matching distance

۱۲ Maximum Common Sub graph (MCS)

۱۳ distance of adjacency matrices

۱۴ Graph Edit Distance (GED)

۱۵ time evolving

۱۶ Compact Matrix Decomposition (CMD)

تشخیص ناهنجاری مبتنی بر خوشه یا اجتماع

در مورد روش مبتنی بر اجتماع یا خوشه، به جای زیر نظر گرفتن تغییرات در کل شبکه، یک اجتماع در هر زمان، برای هر حادثه‌ی غیرعادی ای زیر نظر گرفته می‌شود. در [۵۶] یک برنامه‌ی تشخیص داده‌ی پرت ساختار یافته در جریان‌های شبکه‌ای وسیع ارائه دادند که با تقسیم پویای شبکه برای ساخت مدل‌های آماری مقاوم در برابر رفتار ارتباطی است.

تشخیص ناهنجاری مبتنی بر مدل‌های احتمالاتی

نظریه احتمال مدلی می‌سازد که می‌تواند طبیعی تلقی شود و انحراف از این مدل به‌عنوان ناهنجاری علامت گذاری شود. یک روش دو مرحله‌ای در [۵۷] ارائه شد که گام اول شامل مدل‌های بیزین مزدوج برای فرآیندهای شمارش زمان گسسته برای ردیابی جفت پیوندهای میان همه‌ی گره‌ها در گراف برای ارزیابی طبیعی بودن رفتار است. در گام دوم استنتاج شبکه استاندارد روی کاهش زیرمجموعه‌ی گره‌های بالقوه غیرعادی اعمال می‌شود.

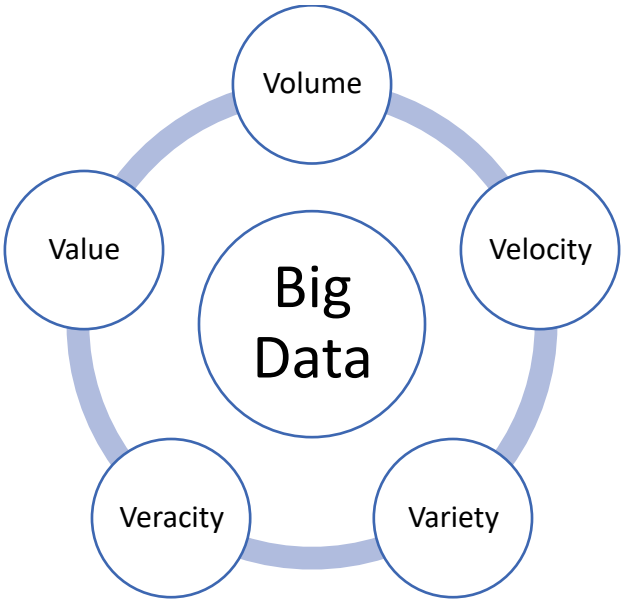
تشخیص ناهنجاری مبتنی برپنجره

الگوریتم‌های تشخیص ناهنجاری برخی روش‌ها را ارائه می‌کنند که محدود به یک چارچوب زمانی است. در [۵۸] روشی ارائه دادند به نام یادگیری الگو و تشخیص ناهنجاری روی جریان‌ها۱۷که یک روش تقسیم‌بندی و پنجره‌گذاری است که گراف را همانطور که در طول زمان در جریان است تقسیم می‌کند و ناهنجاری‌ها و الگوهای غیرقانونی که متعلق به مجموعه‌ای از الگوهای قانونی یافت شده در پنجره زمانی جاری است را حفظ می‌کند.

۲-۶ تعریف Big Data و کاربرد آن در کشف تقلب

۱۷ Pattern Learning and Anomaly Detection on Streams (PLADS)

تعریف Big Data در تحقیقات مورد توافقی جهانی نیست. بنابراین، از تعریف فراگیر دیمچنکو و همکاران استفاده می‌کنیم. همانطور که در شکل (۸) دیده میشود، داده‌های بزرگ را با پنج V تعریف می‌کنند: حجم (Volume)و سرعت (Velocity) و تنوع(Variety)، صحت(Veracity) و ارزش(Value).



شکل ۸ پنج V داده‌های بزرگ[۵۹]

حجم وابسته است به مقدار زیاد داده‌ها، سرعت مربوط است به سرعت بالایی که در آن داده‌های جدید تولید می‌شود، تنوع مربوط به سطح پیچیدگی داده‌ها (به عنوان مثال ترکیب داده‌ها از منابع مختلف)، صحت بیانگر اصالت داده‌ها، و مقدار نشان دهنده‌ی اینکه چقدر کیفیت داده‌ها با توجه به نتایج مورد نظر، خوب است.

مجموعه داده‌های منتشر شده توسط CMS بسیاری از این خصوصیات Big Data را به معرض نمایش می‌گذارد. این مجموعه داده‌ها واجد شرایط Big Volume هستند زیرا در برگیرنده سوابق سالانه مطالبات پزشکان ارائه‌دهنده خدمات پزشکی در کل ایالت متحده می‌باشند.

هر ساله CMS داده‌های سال گذشته را منتشر می‌کند که حجم وسیعی از داده‌های موجود را افزایش می‌دهد. مجموعه داده‌ها حاوی حدود ۳۰ ویژگی است. هر کدام از ۳۰ ویژگی، از مشخصات جمعیت ارائه‌دهنده و انواع پروسه‌های درمانی، تا مبالغ پرداخت و تعداد خدمات انجام شده می‌باشد؛ بنابراین واجد شرایط Big Variety می‌باشند. علاوه بر این، مجموعه داده ترکیبی مورد استفاده در این تحقیق، ذاتاً شرط داده‌های Big Variety را

فراهم می‌کند، زیرا سه منبع اصلی (اما متفاوت) Medicare را با هم ترکیب می‌کند. از آنجایی که CMS یک برنامه دولتی با کنترل کیفیت شفاف و مستندات دقیق است، برای هر مجموعه داده، این مجموعه داده‌ها قابل اعتماد، معتبر و نمایانگر کلیه مطالبات ارائه‌دهندگان شناخته شده‌ی Medicare هستند که نشان از صحت Big Veracity داده‌ها دارد. مجموعه داده LEIE می‌تواند به عنوان Big Value در نظر گرفته شود زیرا شامل بزرگترین مخزن کلاهبرداران ارائه‌دهنده خدمات پزشکی شناخته شده دنیای واقعی در ایالت متحده می‌باشد.

۱-۶-۲ برخی معیارهای تحلیل شبکه‌ی پزشکان

معیارهای مرکزی، برای تجزیه و تحلیل اهمیت نسبی پزشکان در شبکه هستند. برای بررسی اینکه همکاری میان آنها در درمان بیماران وجود دارد یا خیر و عبارتند از:

۱. درجه: اهمیت پزشک متناسب با تعداد بیماران مشترک با پزشکان دیگر است.

۲. مقدارخاص[۱۸]: هرچه تعداد بیماران مشترک با سایر پزشکان مهم بیش‌تر باشد، پزشک مهم‌تر درنظر گرفته می‌شود. اگر پزشک تعداد زیادی از بیماران را به اشتراک بگذارد، اما با پزشکانی که در شبکه مهم نیستند، پزشک مهم در نظر گرفته نمی‌شود[۶۰].

۳. بینابینی[۱۹]: با توجه به اینکه پزشکان می‌توانند برای دیگر پزشکانی که نزدیکترند تاثیرگذار باشند، مثلا زمانی که پزشک یک پزشک دیگر را به بیمار خود نشان می‌دهد، تعریف معیاری که این مجاورت را نشان دهد امکان‌پذیر است[۶۱].

۴. نزدیکی[۲۰]: پزشکان می‌توانند توسط مقدار سایر پزشکان که در میان یکدیگر شناخته شده‌اند، غیرمستقیم و بدون اطلاع آن‌ها، مرتبط باشند. بنابراین، با توجه به تعداد پزشکان مورد نیاز برای اتصال دو پزشک، می‌توان میزان جدایی پزشکان را محاسبه کرد[۶۲].

---

۱۸ eigenvalue

۱۹ betweenness

۲۰ closeness

۲-۷ کلان داده‌ها و تشخیص تقلب

از نظر اقتصادی تقلب در بیمه به یک مساله جدی در حال افزایش تبدیل شده‌است. براساس گزارش خبر BBC در سال ۲۰۰۷ مطالبه‌های بیمه تقلبی ۱.۶ بیلیون پوند در سال برای بیمه‌گذاران انگلستان هزینه دارد. خسارات کلی ناشی از تقلب توسط بیمه غیرقابل محاسبه است. تشخیص تقلب بیمه برای جلوگیری از نتایج مخرب تقلب بیمه مهم است. تشخیص تقلب بیمه شامل بررسی مطالبه‌های جعلی از مطالبه‌های اصلی است. به این ترتیب، افشای رفتار یا فعالیت جعلی، تصمیم‌گیرندگان را برای توسعه استراتژی‌های مناسب برای کاهش اثر تقلب قادر می‌سازد. داده‌کاوی یک نقش مهم در تشخیص تقلب بیمه دارد، همانطور که اغلب برای استخراج و پرده‌برداری از حقایق پنهان، مقدار زیادی داده به‌کار می‌رود. داده‌کاوی درباره‌ی پیداکردن روش‌هایی است که قابل اعتمادند، قبلا ناشناخته بودند و از داده‌ها قابل اجرا هستند. این داده باید دردسترس و مرتبط و کافی و تمیز باشد. هم‌چنین مسئله داده‌کاوی باید به خوبی قابل تعریف باشد و با ابزارهای پرس‌وجو قابل حل نباشد و با یک مدل پردازش داده‌کاوی تعلیم داده شود[۱۳]. داده‌کاوی را به این صورت تعریف می‌کنند: پروسه شناسایی الگوهای مورد علاقه در پایگاه داده بطوریکه بعدا بتوانند در تصمیم‌گیری استفاده شوند[۱۴].

[۱۴] داده‌کاوی را پروسه‌ای تعریف می‌کند که از ریاضیات آماری و هوش‌مصنوعی و تکنیک‌های یادگیری ماشین برای استخراج و شناسایی اطلاعات مفید استفاده می‌کند و بطور پیوسته از یک پایگاه‌داده بزرگ تجربه کسب می‌کند.[۱۵] بیان می‌کند که هدف داده‌کاوی به دست آوردن اطلاعات مفید و غیرآشکار از داده‌های ذخیره شده در مخازن بزرگ است. [۱۶]مشخص می‌کند که یکی از مزایای مهم داده‌کاوی این است که می‌تواند برای توسعه یک کلاس جدید از مدل‌ها برای شناسایی جملات، قبل از اینکه توسط متخصصان شان تشخیص داده شوند، استفاده شود.[۱۱] اشاره دارد که تشخیص تقلب یکی از بهترین کاربردهای داده‌کاوی در صنعت و دولت است. تکنیک‌های داده‌کاوی گوناگونی در تشخیص تقلب بیمه به‌کار گرفته شده‌اند مانند شبکه‌های عصبی، مدل‌های رگرسیون منطقی، روش‌های نایوبیز و درخت تصمیم.

در داده‌های کلان، فقط موضوع حجم [۲۱] مطرح نیست و باید سایر موارد از قبیل تنوع[۲۲] داده‌ها و سرعت [۲۳], برقرار باشد تا بتوان به دنیای یادگیری عمیق وارد شد. در موضوع کشف تقلب حوزه درمان نه‌تنها با حجم انبوهی از داده‌های متنوع مواجه‌ایم، بلکه این داده‌ها و الگوهای مربوط روزبه روز درحال تغییرند.

در سال‌های اخیر علاقه رو به افزایش در کاوش داده‌های مراقبت سلامت برای تشخیص تقلب شکل گرفته است. سیستم‌ها برای پردازش مطالبات الکترونیک پیاده‌سازی شده‌اند تا بصورت اتوماتیک بازرسی و مرور از داده‌های مطالبات را انجام دهند. این سیستم‌ها برای تشخیص اعمال تقلبی، صورتحساب اشتباه، مطالبات تکراری و سرویس‌هایی که تحت پوشش درمانی نیستند، طراحی شده‌اند. قابلیت‌های تشخیص تقلب این سیستم‌ها معمولاً محدود است زیرا تشخیص بطور عمده مبتنی بر قوانین ساده از پیش تعریف‌شده توسط متخصصان امر است. برای رسیدن به تشخیص موثرتر، بسیاری از محققین روش‌های پیچیده‌تر مقابله با تقلب را توسعه داده‌اند که بر اساس داده کاوی، یادگیری ماشین و سایر روش‌های تحلیلی است. روش‌های جدید ارایه شده برخی مزیت‌های اصلی مانند یادگیری خودکار الگوهای تقلب از داده‌ها و تعیین احتمال تقلب برای هر مورد و شناسایی انواع جدید تقلب که قبلاً ثبت نشده‌اند را دارا هستند[۱۰],[۲۵].

۲-۸ رویکردهای کلی کشف تقلب

رویکرد های کلی برای حل مساله پیدا کردن تقلب به صورت زیر قابل دسته بندی است.

۲-۸-۱ الگوریتم‌های خوشه‌بندی

خوشه‌بندی اولین بار روی داده‌های پزشکی برای بخش‌بندی درمان پزشکان عمومی توسط [۷۰] اعمال شد. در [۷۱] و [۲۵] از داده‌های جغرافیایی در یک رویکرد مبتنی بر خوشه‌بندی استفاده نمودند. الگوریتم گروه‌خوشه‌ای بیزین برنولی[۲۴][۷۲] با تمرکز بر وقوع ویزیت میان ارائه‌دهندگان و ذینفعان، داده‌های دوتایی را مدل

---

۲۱ volume

۲۲ variety

۲۳ velocity

۲۴ Bayesian Bernoulli co-clustering

می‌کند. این به طور بالقوه می‌تواند یک نوع تقلب در حال ظهور به نام "تقلب توطئه" را نشان دهد که شامل ویژگی‌های بیش از یک عضو از سیستم پزشکی است. این الگوریتم های خوشه بندی به بازرسان کمک می‌کند که صورتحساب و متغیر مطلوبشان را گروه‌بندی کنند. [۷۳]

یک مرور کلی از روش های تشخیص داده پرتدر برخی از آزمایش‌ها برای ارزیابی اثربخشی آن ارائه می‌کند. این روش‌های تجزیه تحلیل شامل مدل‌های خطی، طرح جعبه‌ای[۲۵]، تحلیل قله[۲۶]، خوشه‌بندی چند متغیره و ارزیابی متخصص می‌باشد. [۷۴] یک روش تشخیص داده پرت مبتنی بر چگالی محلی برای شناسایی الگوهای پرداخت نامناسب در سیستم پزشکی استرالیا ارائه می‌دهد. [۷۵] یک رویکرد یکپارچه که ترکیبی از انتخاب ویژگی، خوشه‌بندی، تشخیص الگو و تشخیص بیرونی است برای شناسایی تقلب در سیستم پزشکی استرالیا ارائه نمودند. [۷۶] یک روش تشخیص ناهنجاری دو مرحله‌ای برای شناسایی بیمارستان‌های جعلی در سیستم مراقبت بهداشت عمومی برزیل ارائه می‌کند. همچنین شامل مطالعات تشخیص داده پرت با داده های تجویزی است. [۷۷] یک مدل رفتاری پایه نرمال را برای شناسایی ناهنجاری ها برای شناسایی ناهنجاری‌های مربوط به هر نسخه ایجاد می‌کند. [۷۸] یک مدل تشخیص داده پرت مبتنی بر استنتاج بیزی است که با استفاده از توزیع احتمالات و فواصل قابل قبول برای ارزیابی ارجاعات ارائه می‌دهد. [۷۹] استفاده از یک تابع غلظت[۲۷] را به عنوان یک ابزار تشخیص پیش نمایش داده پرت برای کمک به ارزیابی تقلب پزشکی ارائه می‌دهد. علاوه بر این، ابزارهای صنعتی مبتنی بر تجزیه و تحلیل گراف، تجزیه و تحلیل ارتباطات و انجمن‌ها، ممکن است به بازرسان کمک کند تا روابط، پیوندها و الگوهای پنهان به اشتراک گذاری اطلاعات و تعاملات در گروه های بالقوه جعلی ارائه دهندگان و بیماران را آشکار سازند. تعداد و کیفیت ارتباط بین مشاغل را می توان با استفاده از شباهت در اطلاعات ارتباطی آن‌ها، مکان، ارائه دهندگان خدمات، دارایی‌ها و وابستگی‌ها تجزیه و تحلیل نمود. ارتباطات بالقوه با بازیکنان درگیر در تقلب ممکن است پرچم های قرمز را به ارمغان بیاورند و منجر به تحقیقات آتی گردند. این به طور خاص می‌تواند برای آشکارسازی شبکه‌های سازمان یافته، پیچیده و هماهنگ ارائه دهندگان و بیماران مفید باشد. رویکردهای بدون ناظر به طور کلی مورد استفاده قرار می‌گیرد تا قبل از اینکه متخصصان حوزه را به تحقیق بفرستند فعالیت‌های جعلی را به طور بالقوه برچسب بزنند. بنابراین، یک همکاری نزدیک بین پزشکان، آمارگیران و افرادی که در

---

۲۵ boxplots

۲۶ peak analysis

۲۷ concentration function

تصمیم‌گیری شرکت دارند، در مراحل تعیین و تنظیم مدل و تجزیه و تحلیل و تفسیر نتایج سودمند خواهد بود[۶۹].

۲-۸-۲ الگوریتم Apriori

الگوریتم Apriori یکی دیگر از تکنیک‌هایی است که در تشخیص تقلب استفاده می‌گردد. این الگوریتم (Agrawal و همکاران ۱۹۹۳)، مهم‌ترین الگوریتم کلاسیک برای کاوشاقلاممکرر است. Apriori برای یافتن همه اقلام مکرر در پایگاه داده داده شده DB استفاده می شود. بر اساس اصل Apriori هر زیر مجموعه‌ای از اقلام مکرر نیز باید مکرر باشد. به عنوان مثال: اگر XY مجموعه اقلام مکرر است، هر دو A و B باید مجموعه های مکرر باشند. ایده کلیدی الگوریتم Apriori این است که چند گذر از پایگاه داده را انجام دهیم. که یک رویکرد تکراری که به نام جستجوی اول-پهنا[۲۸](جستجوی سطح هوشمندانه) شناخته می‌شود که در آن k-آیتمبرای کشف (k + ۱) آیتم بکار می روند. در ابتدا، مجموعه اقلام ۱-تکراری یافت می‌شود که آستانه پشتیبانی را برآورده می‌کند، توسط L۱ نشان داده می‌شود. در هر گذر بعدی، ما با یک مجموعه بذر از اقلام موجود در گذر قبلی که بزرگ بوده است، شروع می‌کنیم. این مجموعه بذر برای تولید مجموعه های جدید بالقوه بزرگ استفاده می‌شود که به نام‌های مجموعه اقلام کاندید شناخته می‌شود. در پایان گذر، تعیین می‌شود که کدامیک از اقلام نامزدها واقعا بزرگ (مکرر) هستند، و آنها تبدیل به دانه برای گذر بعدی می‌شوند. بنابراین، L۱ برای پیدا کردن L۲ استفاده می‌شود، مجموعه ای از مجموعه‌های مکرر ۲-آیتم که برای پیدا کردن L۳ و غیره استفاده می‌شود، تا زمانی که هیچ مجموعه مکرر k-آیتمی موجود نباشد[۸۰].

روش های مختلفی برای بهبود کارایی الگوریتم Apriori مانند جدول کاهش تراکنش، تقسیم بندی و ... استفاده می‌شود [۸۱],[۸۲]. در[۸۳]، نویسندگان روشی را برای تفسیر ویژگی‌هایی که مقادیر پیوسته دارند با استفاده از فاصله مساوی عرض باند داخلی[۲۹] ارائه نمودند که براساس نظر متخصصین پزشکی انتخاب شده است. یک تحقیق دیگر[۸۴]، صورتحساب پزشکی را با استفاده از الگوریتم Apriori تحلیل می‌کند.، برخی از اصلاحات را در الگوریتم Apriori موجود پیشنهاد دادند و سپس از اثربخشی آن در اطلاعات مفید ساخته شده در صورتحساب

---

۲۸ breadth-first search

۲۹ width binning interval

پزشکی استفاده کردند. همچنین از الگوریتم Apriori برای کشف بیماری های مکرر در اطلاعات پزشکی استفاده می‌کند. در [۸۵] روشی برای تشخیص وقوع بیماری با استفاده از الگوریتم Apriori در نقاط خاص جغرافیایی در دوره زمانی خاص ارائه شده است.

بولتن و هاند [۸۶] در سال ۲۰۰۱، PGA[۳۰]را به عنوان یک روش نامزد برای یک تکنیک تشخیص تقلب بدون ناظر ارائه نمودند. که ترکیبی است از تحلیل خوشه‌بندی و نمایه‌سازی. تحلیل خوشه یک کار توصیفی مرسوم برای شناسایی یک مجموعه محدود از دسته‌ها یا خوشه‌ها برای توصیف مجموعه داده است[۸۷]. برای کشف اینکه آیا یک ارائه‌دهنده یک رفتار مطالبه‌ای مشکوک دارد یا نه باید با سایر متخصصان زمینه مشابه مقایسه شود. PGA ابزاری محبوب است برای فهم اینکه چگونه رفتار یک پزشک خاص با رفتار سایر پزشکان در یک گروه خاص مرتبط می‌شود. یک جنبه مهم از این تحلیل این است که چگونه ارائه‌دهندگان با یکدیگر گروه می‌شوند و چرا با هم گروه می‌شوند. به‌طور خاص در صنعت پزشکی متخصصان فوق تخصص بسیاری وجود دارد و گروه‌بندی تقریبی برخی از آن‌ها باهم کار ساده‌ای نیست. بنابراین استفاده از متخصصان برای تشکیل گروه برای اعمال تحلیل PGA نیاز است. فرض بر این است که گروه‌بندی پزشکان ممکن است و سپس مقایسه رفتار می‌تواند آغاز گردد. به‌عنوان مثال ۲۰ دندانپزشک در کدپستی خاصی باهم گروه بندی شده‌اند و توزیع معالجات بررسی شده است، نتیجه میانگین گروهی تعداد دفعات معالجات ارائه شده در هر گروه از بیماران است. خلاصه‌سازی‌های میانگین‌های گروهی در جدول (۲) نشان داده شده‌اند و این مثالی است از اینکه چگونه PGA می‌تواند بکار رود. حال نمایه‌های یک دندانپزشک خاص می‌تواند با میانگین گروه مقایسه شود. اگر دندانپزشکی در طول یک دوره زمانی خاص به‌طور غیرعادی تعداد بالایی کانال ریشه انجام داده باشد، به این معنی نیست که مرتکب تقلب شده است و می‌تواند بر تحقیقات بیشتر روی آن دندانپزشک دلالت داشته باشد. از آنجایی که کانال ریشه یک جراحی دهانی در نظر گرفته می‌شود نرخ بازپرداخت آن به‌طور قابل ملاحظه ای بالاتر از سایر معالجات است و بنابراین محتمل است که دندانپزشک با قصد منحرف از این معالجه برای صورتحساب بیشتر استفاده کرده باشد.

۳۰ Peer Group Analysis

جدول ۲ مثالی از میانگین گروهی برای PGA [۸۸]

Treatment	Average per month	Reimbursement Rate
Cavity treatment	150	**$ 50**
Pulling teeth	15	**$ 300**
Root canal	3	**$ 1200**

۳-۸-۲ روش‌های کلاس بندی

روش‌های کلاس بندی که برای تشخیص اختلاف بین مطالبات جعلی و قانونی آموزش داده شده‌اند فرصتی را برای استفاده از تشخیص تقلب در حوزه پزشکی فراهم می‌کنند. روش‌های با ناظر در تشخیص تقلب در کارت‌های اعتباری و در حوزه مخابرات نسبت به این روش‌ها در بخش‌های پزشکی خاص مانند مراقبت‌های بیمارستانی که تشخیص اینکه آیا فرآیند ارائه شده واقعاً رخ داده یا ضروری بوده یا خیر ، ساده‌تر هستند. زمانی‌که تقلب کننده، مطالبات جعلی که مشابه قانونی هستند را ارائه می‌کند یک روش شناسایی که تشخیص دهد آیا درمانی صورت گرفته یا نه، نیاز است. در این‌گونه موارد زمانی‌که هیچ رفتار صورتحسابی افراطی‌ای وجود ندارد، تشخیص روش‌هایی مانند Profiling و تشخیص outlier با شکست مواجه خواهند شد. روش‌های بدون ناظر مانند تشخیص داده پرت بر غیرعادی بودن و پرت بودن متمرکز هستند که اگر صورتحساب طبیعی باشد رخ نمی‌دهند. Profilingو monitoring متمرکز بر تشخیص تغییرات در رفتار هستند و ابزار موثری برای مقابله با این نوع تقلب به نظر نمی‌رسند [۸۹].

در سال ۲۰۱۶، در مقاله [۹۰] چگونگی بکارگیری تکنیک‌های بدون ناظر در مرحله پس از پرداخت برای شناسایی الگوهای تقلب در بیمه ارائه شده است. در این مقاله تأکید ویژه‌ای بر معماری سیستم، معیارهای طراحی

شده برای تشخیص داده‌های پرت و علامت‌گذاری ارائه‌دهندگان مشکوک به تقلب را نشان می‌دهد. این الگوریتم‌ها بر روی داده‌های Medicaid شامل ۶۵۰،۰۰۰ ادعای مراقبت‌های بهداشتی و ۳۶۹ دندانپزشکان یک ایالت مورد آزمایش قرار گرفتند. دو کارشناس کلاهبرداری در امور بهداشتی، پرونده‌های علامتدار را ارزیابی کردند و نتیجه گرفتند که ۱۲ از ۱۷ ارائه‌دهنده که در صدر لیست قرار دارند (۷۱٪)، الگوهای ادعای مشکوک را ارائه کرده‌اند و برای تحقیقات بیشتر باید به مقامات ارجاع شوند. ۵ ارائه‌دهنده باقیمانده (۲۹٪) را می توان طبقه‌بندی نادرست دانست زیرا الگوهای آنها با ویژگی‌های بخصوص ارائه‌دهنده قابل توضیح است. انتخاب ارائه‌دهندگان علامتدار در صدر جدول، به عنوان یک روش هدفمند، ارزشمند است و تجزیه و تحلیل فردی ارائه‌دهنده، مواردی از کلاهبرداری بالقوه را آشکار می‌کند. این مطالعه نتیجه‌گیری می‌کند که، از طریق تشخیص داده‌های پرت، می‌توان الگوهای جدید کلاهبرداری بالقوه را با مکانیسم‌های شناسایی خودکار آینده، شناسایی کرد. اگرچه تکنیک تشخیص داده‌های پرت، نیاز به همکاری متخصصان امر برای طراحی معیارها و بخصوص تفسیر نتایج دارد. در همین سال، در مقاله [۹۱] یک الگوریتم مبتنی بر PageRank برای تشخیص کلاهبرداری و ناهنجاری‌های مراقبت‌های بهداشتی ارائه شده است. این الگوریتم در مجموعه داده‌های Medicare-B، داده واقعی با ۱۰ میلیون ادعای بیمه خدمات درمانی، اعمال شده است. این الگوریتم با موفقیت، ده‌ها ناهنجاری قبلاً گزارش نشده را شناسایی می‌کند.

یک سال بعد، در مقاله [۹۲] در سال ۲۰۱۷ یک الگوریتم بهبود یافته برای تشخیص داده‌های پرت مبتنی بر خوشه بندی K-means به منظور شناسایی تقلب پزشکی مشکوک در گزارش‌های بیمه سلامت ارائه شده است. در این مقاله به چگونگی پیش پردازش داده‌ها برای کلاهبرداری در بیمه سلامت پرداخته شده است. از مزایای این روش میتوان به کاهش زمان اجرا و استفاده کردن از داده‌های واقعی اشاره کرد. از طرفی دیگر، از معایب این روش میتوان گفت ویژگی‌های استفاده شده فقط مربوط به بازپرداخت بیمه سلامت هستند و شامل اطلاعات بیمار و نسخه نمی‌باشند. در همین سال، در مقاله [۹۳]، انواع مختلفی از روابط را مورد مطالعه و مورد بحث قرار داده شده و بر روی روابط کوچک اما انحصاری که مشکوک هستند و ممکن است نشانگر تقلب‌های بالقوه مراقبت‌های سلامت باشد، تمرکز شده است. دو الگوریتم برای شناسایی این جوامع کوچک و اختصاصی در این مقاله استفاده شده‌اند. این الگوریتم‌ها می‌توانند در مجموعه داده‌های بزرگتر اعمال شوند و بسیار مقیاس پذیر هستند. از نقاط ضعف این کار می‌توان به آزمایش الگوریتم‌ها با مجموعه داده‌های سنتز شده آزمایشگاهی اشاره کرد. [۹۳] یک الگوریتم بدون ناظر مبتنی بر فاصله برای ارزیابی خطر تقلب نسخه‌ها ارائه نموده است. در مقاله [۹۳] ماتریس ارتباط میان هر دو پزشک محاسبه می‌شود و پزشکانی که در یک شبکه به‌هم متصلند و با سایر

پزشکان ارتباطی ندارند، بعنوان شبکه تقلب‌آمیز شناسایی می‌شوند، و همچنین از اپراتورهای DB به جای loop برای بیش از دو پزشک استفاده شده است. آزمایشات بر روی بانک اطلاعاتی جراحی قلب بزرگسالان انجام شده است. نتایج به دست آمده از آزمایشات نشان می‌دهد که مدل پیشنهادی با نرخ مثبت واقعی ۷۷.۴٪ و نرخ مثبت کاذب ۶٪ برای نسخه‌های پزشکی متقلب عملکرد خوبی دارد. مدل ارائه شده دارای مزایای بالقوه از جمله دقت بالای پیشبینی خطرات در تقلب در نسخه پزشکی، تجزیه و تحلیل غیر خطی از نسخه‌های پرخطر توسط متخصصان انسانی و توانایی یادگیری با بروزرسانی‌های منظم از مجموعه‌داده‌های یکپارچه است. همچنین ترکیب چنین سیستمی در مراجع بهداشتی، سازمان‌های تأمین اجتماعی و شرکت‌های بیمه می‌تواند کارآیی را برای اطمینان از رعایت قانون بهبود بخشد، و هزینه‌های حسابرسی متخصص انسانی را بطور چشمگیری کاهش دهد. البته در این روش پزشکان بر اساس تخصصشان تقسیم‌بندی نشده‌اند و از مجموعه داده‌ی واقعی در این تحقیق استفاده نشده است. در مطالعه‌ی [۹۴]، از رویکرد داده‌کاوی در یک مجموعه‌داده وسیع سازمان بیمه درمانی از ادعاهای تجویز پزشکان عمومی بخش خصوصی استفاده است. این روش شامل ۵ مرحله است: شفاف سازی ماهیت مسئله و اهداف، تهیه داده‌ها، شناسایی و انتخاب شاخص، تجزیه و تحلیل خوشه‌ای برای شناسایی پزشکان مشکوک و تجزیه و تحلیل تمایزکننده برای ارزیابی اعتبار رویکرد خوشه‌بندی. در مقاله [۹۵]، مشکل شناسایی تقلب در سیستم‌های ارتباطی، به ویژه موارد کلاهبرداری، با ارائه یک روش تشخیص ناهنجاری که از نگاشت پزشکان با استفاده از بیماران مشترک بعنوان یک پروکسی برای ترسیم ارتباط میان آن‌ها استفاده می‌کند، مورد بررسی قرار گرفته است. هدف اصلی این است که رفتارهای انحرافی را در زمان مفید تشخیص دهد و اساس بهتری را برای تحلیلگران کلاهبرداری فراهم کند تا در تصمیم‌گیری‌ها در زمینه ایجاد موقعیت‌های احتمالی کلاهبرداری دقیق‌تر باشد.

۴-۸-۲ روش‌های یادگیری ماشین ترکیبی

در [۹۶] یک روش تشخیص تقلب موثر هیبریدی SSIsomap و SimLOF پیشنهاد شده است. SSIsomap، درواقع isomap را برای رفتار خوشه‌ها در رفتار کلاس‌ها بهبود می بخشد و SimLOF که LOF را بهبود می بخشد تا تشخیص داده پرت را بهبود بخشد، سپس از شواهد تئوری DempsterShafer برای ترکیب شواهد الگوی رفتاری و شواهد بیرونی استفاده می‌شود، که درجه اعتقاد به تقلب برای مطالبات جدیدی که از راه می‌رسند فراهم می‌کند. نتیجه آزمایش نشان می دهد که روش آن‌ها دقت بیشتری نسبت به روش های موجود در تشخیص تقلب بیمه های پزشکی دارد[۹۶]. در مطالعه‌ی [۹۷]، از چندین روش شناخته شده داده‌کاوی استفاده

شده است، مانند روش مقایسه پردازش تحلیلی سلسله مراتبی (AHP) برای وزن‌دهی بازیگران و ویژگی‌ها، حداکثر انتظار (EM) برای خوشه‌بندی بازیگران مشابه، ذخیره دو مرحله‌ای داده‌ها برای کنترل ریسک محاسبات، ابزارهای تصویرساز برای تجزیه و تحلیل مؤثر و امتیاز Z برای استانداردسازی. در این مقاله، متخصصان در تمام مراحل مطالعه شرکت می‌کنند و شش نوع رفتار غیرطبیعی و متفاوت را با استفاده از صفحه داستانی(storyboards) تولید می‌کنند. چارچوب ارائه شده با داده‌های واقعی برای شش نوع رفتار غیرطبیعی متفاوت برای نسخه‌ها با پوشاندن کلیه بازیگران و کالاهای مربوطه ارزیابی می‌شود. علاوه بر این، یک مدل صرفه جویی در هزینه نیز ارائه شده است. چارچوب توسعه یافته، یعنی مجموعه eFAD ، مستقل از بازیگر و کالاها و قابل تنظیم(یعنی به راحتی در محیط پویا کلاهبرداری و رفتارهای ناهنجار سازگار است) است. در این روش زمان اجرا، به طور قابل توجهی کاهش یافته است.

در جدول (۳) هریک از رویکردها و روش‌هایی که پیش‌تر بحث شد به تفکیک مزایا و معایب بیان شده است:

جدول ۳ انواع رویکردها و روش‌های موجود در کشف تقلب سیتم سلامت

ردیف	نام روش	ایده	مزایا	معایب	پارامترها
۱	Outlier Detection in HealthCare Fraud-A Case Study in The Medicaid Dental Domain [۹۰]	اعمال فیلترینگ ویژگی‌ها برای جداسازی بازپرداخت‌های کم، تعداد بیماران کم، تعداد مطالبات کم و استفاده از تکنیک‌های تحلیل و آنالیز و استفاده از تکنیک های تشخیص Outlier شامل انحراف از مدل خطی، انحراف خوشه، انحراف از خوشه تکی، انحراف گرایشی، حداکثر انحراف	تست روی مجموعه داده واقعی استخراج ویژگی‌ها و ارزیابی با استفاده از انجام مصاحبه با متخصصان و پزشکان تفسیر نتایج توسط تحلیل گران	اعتبارسنجی دشوار اثربخشی تکنیک تشخیص outlier نیاز به همکاری متخصصان امر برای طراحی معیارها و به‌خصوص تفسیر نتایج دارد تکنولوژی outlier هنوز در مرحله آزمایشگاهی است و خود را در عمل و در اجرای طولانی ثابت نکرده	تعداد مطالبات بازپرداخت هر ذینفع مقدار مطالبات بازپرداخت هر ذینفع مقدار مطالبات در ایام تعطیل میانگین تعداد مطالبات بازپرداختی هر ذینفع میانگین مقدار مطالبات بازپرداختی هر ذینفع کد نسخه

					کد دندان هزینه نسخه میانگین تعداد نسخه‌های بازپرداختی هر مطالبه
۲	A Novel Page Rank-Based Algorithm to Identify Anomalies [۹۱]	استفاده از یک الگوریتم personalized page rank و محاسبه یک specialty centric personalized page rank برای هر نود و سپس اتصال نودها براساس آن برای بدست آوردن آنومالی	Page rank که سابقاً در حوزه تحلیل شبکه به کار رفته بود در حوزه تقلب در مطالبات بیمه نیز مؤثر عمل می‌کند	وقوع برخی FPها به دلیل اشتراک مشخصه‌های کلی در CPT (نسخه‌ی) پزشکان با تخصص‌های مختلف	Medical procedure code national provider identifier specialty تعداد procedureها در هر سال
۳	Medical Insurance Fraud Recognition Based on Improved Outlier Detection Algorithm [۹۲]	استفاده از یک الگوریتم تشخیص Outlier بهبود یافته بر اساس خوشه‌بندی k-means	استفاده از مجموعه داده واقعی کاهش زمان اجرای الگوریتم با یافتن مقدار بهینه k با پیچیدگی زمانی از مرتبه o(I*k*m*(no)) و پیچیدگی مکانی از مرتبه o((n-o)+k)*m که نسبت به الگوریتم CBLOF بهبود داشت.	ویژگی‌های استفاده شده فقط مربوط به بازپرداخت بیمه سلامت هستند و شامل اطلاعات بیمار و نسخه نمی‌باشند	بازپرداخت مربوط به برونشیت مزمن بازپرداخت مربوط به بیماری‌های قلبی-ریوی بازپرداخت مربوط به ذات الریه
۴	Community Detection Algorithm to Find Suspicious Group of Provider Community [۹۳]	ساخت ماتریس ارتباط میان دو پزشک و شناسایی پزشکانی که در یک شبکه بهم متصلند و با سایر پزشکان ارتباطی ندارند، بعنوان شبکه تقلب	تخصیص احتمال (Likelihood/ احتمال تشکلیل شبکه انحصاری) به هر پزشک سرعت تشخیص بالا	استفاده از داده‌های ساختگی استفاده از برخی پکیج‌های نرم افزاری برای محاسبه ماتریس روابط، که برای مجموعه	procedure code صورتحساب هر procedure code

		آمیز، و همچنین استفاده از اپراتورهای DB به جای برای loopبیش از دو پزشک		داده‌های بزرگ بهینه نیست عدم تفکیک پزشکان طبق تخصص آن‌ها post payment	health claim data تاریخ ارایه خدمات کد پزشک تعداد ویزیت‌های هر بیمار به ازای هر پزشک
۵	Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study [۹۴]	انجام عمل خوشه بندی بر اساس hierachical clustering method و محاسبه تعداد بهینه خوشه ها بر اساس معیار فاصله Euclidian distance با measures استفاده از شاخص اعتباری بیشینه مقدار ضریب همبستگی سیلوعت	انجام تحقیق روی پزشکان هردو بخش عمومی و خصوصی استفاده از مجموعه داده واقعی	حذف داده‌های ناشناس عدم استفاده از روش‌های آماری برای پرکردن داده‌های ازدست رفته	درصد بیمارانی که بیش از یکبار درماه ویزیت شده‌اند میانگین اقلام دارو در یک نسخه میانگین هزینه نسخه دارویی پزشک تعداد نسخ تزریقی/حاوی آنتی بیوتیک تعداد نسخ تزریقی/حاوی آنتی بیوتیکس
۶	A Social Network Analysis Framework for Modeling Health Insurance Claims Data [۹۵]	به‌کارگیری تکنیک‌های برای تحلیل مطالبات بیمه سلامت از طریق نگاشت پزشکان با استفاده از بیماران مشترک بعنوان یک پروکسی برای ارتباط میان آن‌ها	در نظر گرفتن رابطه میان پزشکان، پزشک و بیمار، پزشک و ارایه دهندگان خدمات استفاده از مجموعه داده واقعی بهبود درک اهمیت ویژگی‌های میان افراد ارزیابی مدل توسط تحلیلگران فرآیند و پزشکان و متخصصان	داده‌ی استفاده شده فقط مرتبط به خدماتی است که توسط پزشک ارائه شده و شامل مطالبات مرتبط با تحلیل‌های کلینیکی، آزمایش‌ها و عکس برداری یا بستری در بیمارستان نیست	ID بیمار نوع نسخه اطلاعات زمانی مرتبط به حادثه ID مشاغل درگیر در فرآیند هزینه نسخه تخصص پزشک

۷	An Interactive Machine Learning Based Electronic Fraud and Abuse Detection System in HealthCare Insurance [۹۷]	تولید سناریو توسط متخصصان و پزشکان برای رفتارهای غیرطبیعی و سپس وزن‌دهی actorها با روش وزن‌دهی binary pairwise comparison و محاسبه امتیاز خطای actorها و مطالبات، استفاده از ابزار visualization توسعه یافته تحت QlikView که برای تحلیل بکار می‌رود	سازگار در یک محیط پویا قابل استفاده برای تحلیل proactive , reactive کاهش زمان تحلیل نتایج خروجی توسط کاربران به دلیل استفاده از ابزار visualization استفاده از بیش از یک actor	تغییر مداوم وزن‌ها توسط متخصصان برای شناسایی انواع جدید تقلب تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب	نسبت تعداد نسخه‌ها به تعداد مشخصی از بیمه شدگان نسبت تعداد نسخه‌ها به تعداد مشخصی از پزشکان تعداد کل نسخه‌ها
۸	Multi Stage Method to Detect Provider and Patient Fraud [۹۸]	محاسبه یک معیار ریسک بر اساس فاصله مهالنوبیس و چگالی‌ها و محاسبه ریسک و ساخت درخت تصمیم آن	دقت بالا در مقایسه با روش دقت بالا روی همه ۴ تخصص	عدم آزمایش روش پیشنهادی با تعداد متخصصان بیشتر	تخصص پزشکان(چشم، اعصاب، حلق، عمومی) نرخ شکایت از پزشکان مدت زمان هر ویزیت تعداد ویزیت‌ها تعداد تشخیص(نسخه) تعداد سرویس‌ها و خدمات تعداد دارو

۹-۲ جمع بندی فصل

یکی از مهمترین مشکلات حوزه سلامت، کلاهبرداری است که خسارات قابل توجهی به بار می‌آورد. با توجه به حجم اسناد و ارائه‌دهندگان خدمت، کشف تقلب به صورت سنتی غیرممکن است .در نتیجه، روش‌های داده‌کاوی و تحلیل شبکه با شناسایی الگوهای موجود در داده های کنونی و کشف موارد مشکوک به تقلب، همزمان با حفظ و حتی بهبود خدمات، هزینه‌های این کار را به صورت قابل توجهی کاهش می‌دهد. در این فصل پس از بررسی مفاهیم مربوط به سلامت و تقلب و ... در این حوزه، روشهایی که برای داده‌کاوی و پیداکردن الگوهای تکراری از میان این داده‌ها تا کنون انجام شده، بررسی و چالش‌های آنها مطرح شده است. موارد زیر نکاتی هستند که در پژوهش‌های انجام شده دیده می‌شوند [۴۳] :

- داده یک مساله مهم در زمینه مراقبت سلامت است. عمده داده شامل داده‌های مطالبات از منابع دولتی و شرکت‌های بیمه خصوصی هستند.
- سیستم‌های مراقبت سلامت هر کشور متفاوتند و بطور مداوم در حال تغییر و توسعه‌اند.
- چندین تحقیق روی برخی کشورهای توسعه یافته مانند آمریکا و استرالیا انجام شده است. به عبارت دیگر کشورهای مختلف باید بعنوان منابع داده‌ی جدید در نظر گرفته شوند.
- تشخیص تقلب مراقبت سلامت عمدتا با استفاده از یادگیری ماشین و داده‌کاوی انجام شده است. روش‌های یادگیری ماشین به سه دسته تقسیم می‌شوند با ناظر، بدون ناظر و نیمه نظارتی.
- بیشتر مطالعات از روش‌های یادگیری بدون ناظر استفاده کردند. در برخی موارد روش‌های یادگیری نیمه نظارتی ارائه شده نیز می‌توانند در تشخیص تقلب مراقبت سلامت مفید باشند.
- تحقیقات بررسی شده نشان می‌دهند الگوریتم‌های شناخته شده مانند SVM و KNN و بیزین برای کلاس‌بندی، خوشه‌بندی و تشخیص موارد غیرعادی (abnormal) در تشخیص تقلب مراقبت سلامت استفاده شده‌اند.
- اگرچه الگوریتم‌های متفاوتی برای تشخیص تقلب مراقبت سلامت به کار گرفته می‌شوند، اما الگو یا روش استانداردی که همه موارد را پوشش دهد وجود ندارد.
- با توجه به انواع تقلب، اکثریت تحقیقات روی تشخیص تقلب ارائه‌دهندگان خدمات انجام شده زیرا تقلب ارائه‌دهندگان خدمات یک مساله مهم برای بهبود کیفیت و امنیت یک سیستم مراقبت سلامت است، محققان زیادی به آن‌ها توجه نشان می‌دهند.

- بطور ویژه تحقیق زیادی روی تشخیص تقلب دسیسه‌ای[۳۱] صورت نگرفته، اگرچه که چنین تحقیقاتی می‌توانند برای عواقب سخت تقلب دسیسه گران و کاهش هزینه‌های مراقبت سلامت بسیار مفید واقع شوند.
- مرسوم ترین منبع داده‌ی استفاده شده در آمریکا HFCA، .در استرالیا HCI و در تایوان NHI است.

در [۴۳] برای تشخیص تقلب مراقبت سلامت روش یادگیری بدون ناظر پرتکرارترین روش استفاده شده است زیرا به‌دست آوردن داده‌ی برچسب‌دار در تشخیص تقلب حوزه سلامت بسیار دشوار و پرهزینه است. بطور کلی میتوان نتیجه‌گیری‌های زیر را از بررسی پژوهش‌هایی که به آن‌ها اشاره شد، کرد:

- داده‌ی مراقبت سلامت در حال حاضر بعنوان مجموعه‌ای از داده‌های بزرگ از انواع داده‌ها درنظر گرفته می‌شود. این شرایط مفهوم کلان داده را درپی دارد.

- کلان داده در تحلیل‌های مراقبت سلامت یک زمینه تحقیقاتی جدید است و مطالعات کمی در این زمینه گزارش شده‌اند.

۳۱ conspiratorial

فصل ۳

۳ روش پیشنهادی

۳-۱ مقدمه

در فصل‌های گذشته با اهمیت موضوع تقلب در داده های سلامت آشنا شده و انواع مختلف تقلب در این داده ها بررسی گردید. پژوهش های پیشین صورت گرفته در این موضوع همگی دسته بندی و مورد بحث قرار گرفتند. در این فصل با ارایه روشی مبتنی بر گراف سعی در کلاس بندی داده هایی داریم که برچسب متقلب بودن یا نبودن مربوط به آنها وجود ندارد. به عبارت دیگر ابتدا مدل مجموعه داده ای شامل دوکلاس متقلب و عادی را دریافت می کند. سپس سعی میکند شباهت ها و الگوهایی را پیدا کند که در کلاس متقلب و غیر متقلب تکرار می شود. در نهایت برای یک ورودی جدید که برچس یا کلاس آن وجود ندارد، بعد از محاسبه میزان شباهت این ورودی جدید به هرکدام از کلاس های متقلب یا غیر متقلب، الگوریتم عملیات دسته بندی را انجام می دهد.

۳-۲ مجموعه داده‌ها

در این تحقیق از دو مجموعه داده استفاده شده‌است که به شرح زیر می‌باشد.

۳-۲-۱ مجموعه داده LEIE

برای دستیابی صحیح به کارایی تشخیص تقلب، همانگونه که در دنیای واقعی عمل میکند، ما به یک منبع داده نیاز داریم که شامل پزشکانی باشد که مرتکب تقلب در دنیای واقعی شده باشند.بنابراین لیستی از اشخاص و موجودیت های اخراج شده را به کار میگیریم که شامل اطلاعات زیر است:

دلیل اخراج، تاریخ اخراج، تاریخ بازگردانی/ ابطال برای همه ی پزشکان نامناسب شناخته شده برای عمل پزشکی و بنابراین اخراج از عمل در امریکا برای یک بازه زمانی داده شده.

این مجموعه داده ایجاد شده و ماهیانه توسط اداره بازرسی عمومی (OIG[32]) مطابق با بخش ۱۱۲۸و ۱۱۵۶ قانون امنیت اجتماعی نگه داری می‌شود. در این پژوهش از آخرین نسخه‌ی سپتامبر ۲۰۲۰ استفاده شده‌است. OIG اختیار محرومیت افراد از برنامه های مراقبت سلامت فدرالی مانند بیمه پزشکی را داراست. متاسفانه LEIE فراگیر نیست و ۳۸٪ ارایه دهندگان با محکومیت تقلب به جراحی ادامه میدهند و ۲۱٪ با وجود محکومیت از عمل جراحی تعلیق نشده اند. بعلاوه مجموعه داده LEIE فقط شامل مقادیر NPI برای درصد کمی از پزشکان و موجودیت هاست. مثالی از چهار پزشک مختلف و اینکه چگونه در LEIE به تصویر کشیده شده اند در جدول زیر نشان داده شده است، که هر پزشک بدون NPI مقدار ۰ را در داده‌ی LEIE دارد.

در سطح ارایه دهنده تجمیع شده و اطلاعات خاصی با توجه به روندها، داروها یا تجهیزات مرتبط با فعالیت های کلاهبردارانه ندارد. دسته های مختلفی از محرومیت/ اخراج بر اساس شدت گناه وجود دارد که توسط شماره قوانین توصیف شده اند. ما از همه ی محرومیت ها استفاده نمی‌کنیم، بلکه ارایه دهندگان محروم شده را با قوانین انتخاب شده‌ی نشان دهنده‌ی ارتکاب تقلب فیلتر می‌کنیم. جدول ۴ این قوانین را که منطبق بر محرومیت ارایه دهندگان کلاهبردار است و طول محرومیت اجباری را می‌دهد. ما تعیین کرده‌ایم که هر رفتاری که قبل از پایان تاریخ محرومیت اجباری یا در طول آن است، تشکیل دهنده‌ی تقلب است.

[32] Office of inspector general

جدول ۴ قانون های مربوط به مجموعه داده LEIE

RULE NUMBER	DESCRIPTION
1128(A)(1)	محکومیت جرایم مربوط به برنامه
1128(A)(2)	محکومیت مربوط به سو استفاده یا بی توجهی به بیمار
1128(A)(3)	محکومیت جنایی در مورد کلاهبرداری در مراقبت های بهداشتی
1128(B)(4)	لغو یا تعلیق مجوز
1128(B)(7)	کلاهبرداری ، بیرون کردن و سایر فعالیت های ممنوع
1128(C)(3)(G)(I)	محکومیت دو جرم استثنا اجباری ۱۰ سال
1128(C)(3)(G)(II)	محکومیت سه جرم استثنا تخلفات نامعین

۲-۲-۳ مجموعه داده‌ی Medicare Provider Utilization and Payment

مجموعه داده ی Medicare Provider Utilization and Payment اطلاعات مطالبات برای هر روند(نسخه) که یک پزشک در یک سال انجام داده را فراهم میکند. اخیرا این مجموعه داده در وبسایت CMS از سال ۲۰۱۲ تا ۲۰۱۵ در دسترس است. پزشکان با استفاده از NPI[33]یکتا شناسایی شده اند، در حالی‌که روندها

[33] National Provider Identifier

بر اساس کد HCPCS[۳۴] برچسب زده می شوند.سایر اطلاعات مطالبات شامل میانگین هزینه و پرداخت ها، تعداد روندهای انجام شده و تخصص پزشکی(که بعنوان نوع ارایه دهنده نیز شناخته میشود) میباشد.CMS تصمیم گرفت که داده ی Medicare Provider Utilization and Payment را تجمیع کند با:

۱)NPI مربوط به ارایه دهنده ی خدمات

۲)کد HCPCS برای روند یا سرویس انجام شده

۳) محل خدماتی که یک تسهیلات است(F[۳۵]) یا غیرتسهیلات(O[36])، به ترتیب مانند یک بیمارستان یا یک مطب.

هر سطر در مجموعه داده شامل NPI یک پزشک، نوع ارایه دهنده،کد HCPCS تقسیم شده بر اساس محل خدمات همراه با اطلاعات خاص منطبق بر این تقسیم بندی(مانند تعداد مطالبات)، و سایر ویژگی های تغییر نیافتنی(مانند جنسیت). در عمل جراحی، پزشکانی هستند که چه در بیمارستان یا در مطب شان روندهای مشابهی انجام میدهند، همچنین تعداد کمی از پزشکان که تحت چند نوع ارایه دهنده(تخصص) مانند متخصص داخلی و متخصص قلب عمل انجام میدهند. بنابراین برای هر پزشک تعداد زیادی سطر مانند ترکیب های یکتایی از NPI، نوع ارایه دهنده، کد HCPCS، و محل خدمات وجود دارد و بنابراین داده ی Medicare Provider Utilization and Payment میتواند برای فراهم نمودن اطلاعات سطح روند در نظر گرفته شود.

این مجموعه داده لیستی از پزشکان و سایر نهادهای مراقبت های بهداشتی است که برای مدت زمان مشخصی از درگیری در Medicare منع شده اند. مجموعه داده LEIE NPI هر ارائه دهنده را نشان می دهد ، که برای برچسب زدن ادعاهای جعلی استفاده می شود. مجموعه داده های موجود از CMS ، ۲۰۱۲ تا ۲۰۱۵ ، با از پایگاه داده LEIE با در نظر گرفتن دوره های شروع و پایان موارد استثنا ، برای جلوگیری از همپوشانی و احتمال شمارش مضاعف ادعاهای جعلی ، با هم ترکیب شدند. ارائه دهندگان استثنا شده از پایگاه داده LEIE [۱۳] برای

---

[۳۴] Health Common Procedure Coding System

[۳۵] Facility

[36] Non-facility

بدست آوردن برچسب های تقلبها به مجموعه داده اضافه شدند. بانک اطلاعاتی LEIE فقط شامل موارد استثنا در سطح NPI یا ارائه دهنده است ، نه تقلب در ارتباط با اقدامات پزشکی خاص انجام شده. موارد استثنا توسط اعداد مختلف قانون طبقه بندی می شوند ، که نشانگر شدت و همچنین مدت زمان هر حذف است. ارائه دهندگان انتخاب شده ، ارائه دهندگانی بودند که به دلایل شدیدتری مستثنی شدند و موارد استثنا اجباری را توسط OIG [۷]در نظر گرفتند ، همانطور که در جدول ۴ ذکر شده است برای ساخت و آزمایش مدلهای ما ، تصور می کنیم تعدادی از پزشکان LEIEکلاهبردار تلقی می شوند و کسانی که شامل آنها نمی‌شوند تقلبی نیستند. داده‌های Medicare Provider Utilization and Payment حاوی اطلاعات در مورد هر پزشک و روش انجام شده و همچنین سایر ویژگی ها مانند محل خدمات ، مبالغ ارسالی و مبالغ پرداختی است.

همانطور که گفته شد ، داده های LEIE اطلاعات استثنا را برای یک ارائه دهنده فراهم می کند اما نه برای هر روش خاصی که توسط آن ارائه دهنده انجام شده است.

در زمان انتشار این مقاله ، هیچ مجموعه داده شناخته شده ای در دسترس عموم با برچسب های کلاهبرداری توسط ارائه دهنده و با توجه به هر روش انجام شده وجود ندارد.

به همین دلیل ، داده های Medicare Provider Utilization and Payment گروه بندی و در سطح NPI جمع شدند و از مجموع این دو مجموعه داده ما توانسیم اطلاعات و برچسب را باهم داشته باشیم.

از آنجایی که ویژگی‌های عددی در این الگوریتم برای محاسبه درصد شباهت مورد استفاده قرار می گیرد بذا جدولی از مجموعه داده‌ی نهایی که در این پایان نامه استفاده شد در جدول (۵) ارایه می شود.

جدول ۵ جدول مربوط به ستونهای مجموعه داده

شماره	اسم ستون
۱	NPI
۲	FRAUD_LABEL
۳	NPPES_PROVIDER_GENDER
۴	BENE_UNIQUE_CNT
۵	AVERAGE_MEDICARE_STANDARD_AMT

۶	AVERAGE_MEDICARE_PAYMENT_AMT
۷	AVERAGE_SUBMITTED_CHRG_AMT
۸	AVERAGE_MEDICARE_ALLOWED_AMT
۹	BENE_DAY_SRVC_CNT
۱۰	LINE_SRVC_CNT
۱۱	PROVIDER_TYPE

همچنین در این مجموعه داده ویژگی های nppes_provider_genderو provider_type به دلیل اسمی بودن قابل استفاده در مدل به صورت مستقل نیستند لذا آن‌ها را به متغییر دسته ای [۳۷] تبدیل و سپس در مدل لحاظ شدند.

۳-۳ آماده‌سازی داده

همان‌طور که توضیح داده‌شد داده‌ی LEIE برچسب‌ها را برای هر شماره نظام پزشکی ملی نگهداری می‌کند و داده‌ی Medicare Provider Utilization and Payment تجمیعی از اطلاعات ثبت شده برای شماره نظام پزشکی‌ها و افراد است. اما تعداد زیادی از افراد ممکن است نام و مشخصات یکسانی داشته باشند و

[۳۷] Categorical variable

همچنین یک فرد می‌تواند با اخذ تخصص‌های بیش‌تر شماره‌های نظام پزشکی بیشتری دریافت کند. لذا مشخصاتی همچون اسم نمی‌تواند کلید اولیه و منحصربفرد مناسبی برای این داده باشد. لذا همچون پژوهش‌های پیشین تنها نمونه داده‌هایی به کار رفته‌است که دارای شماره‌ی نظام پزشکی (NPI) باشند و مابقی کنار گذاشته شده‌است. از داده‌های باقی مانده تنها NPI نگهداری شد که برای آن‌ها برچسب تقلب یا عدم تقلب وجود داشت. سپس داده‌های این دو مجموعه براساس NPI باهم ترکیب شد. جنسیت افراد (nppes_provider_gender) به متغیر دسته‌ای ۰ و ۱ برای خانم‌ها و آقایان تبدیل شد. تخصص پزشکان (provider_type) نیز به متغییر دسته‌ای تبدیل شد. سپس نمونه‌داده‌هایی که مقادیر تهی داشته و داده‌ی خود را از دست داده‌اند یا ثبت نشده‌است حذف شد. داده‌ی آماده شده شامل حدود ۲۹۰۰ نمونه است که در مقایسه با حجم اولیه‌ی داده (درحدود چند گیابایت) کوچک می‌باشد. متاسفانه داده‌های اولیه‌ی قابل استفاده حجم کمی داشته و نمی‌توان از روش‌های یادگیری ماشین پیچیده برای آموزش و ایجاد مدل استفاده کرد.

۳-۳ نیازمندی‌های روش پیشنهادی

برای پیش بینی اینکه آیا یک ورودی جدید در دسته متقلب ها قرار می گیرد یا خیر، الگوریتم زیر طراحی گردیده است. این الگوریتم در ابتدا مجموعه داده ای دریافت کرده و آن را بر اساس برچسب هایش به دو دسته تقسیم می کند. در داده یک گره در گراف محسوب می شود. و داده هایی که در یک دسته قرار دارند توسط یال به هم وصل می شوند. پس وجود یال بین دو داده نشان از این دارد که این دو داده برچسب یکسان دارند. در این پایان نامه فرض بر وجود یا عدم وجود یال است .

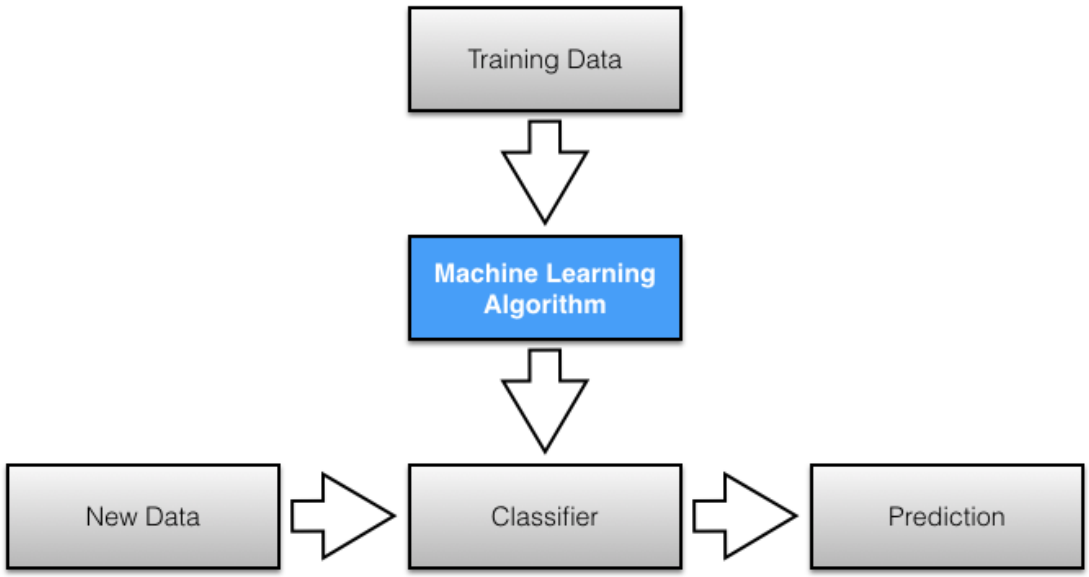
به این صورت که اگر گره عضو یک کلاس باشد با تمام اعضای آن کلاس یال خواهد داشت و اگر عضو ان کلاس نباشد با هیچ کدام از اعضا یال مشترک نخواهد داشت. چنانچه شباهت هر یک از اعضای گراف با هم محاسبه شود می تواند گراف را به حالت وزن دار ترسیم کردم که در این صورت برای گره جدید یا همان داده جدید هم میتواند بر حسب درصد تعیین کرد که چند درصد امکان متقلب بودن یا نبودن وجود دارد.

این عملیات می تواند نتایج را دقیق تر سازد اما مشکل اصلی این روش پیچیدگی آن و زیاد بودن تعداد محاسبات است.به این ترتیب ه هرگاه داده جدیدی به مجموعه اضافه شود باید شباهت آن با تمام اعضای کلاس خود و کلاس دیگر محاسبه شود. این امر در کلان داده ها مشکلات و معضلات خود را به همراه خواهد داشت. لذا در این پژوهش ما به باینری بودن یال ها اکتفا می کنیم.

این الگوریتم از در نهایت یک طبقه بند، فراهم میکند که قادر از برای هر ورودی ای کلاس خاصی که به آن شبیه تر از سایر کلاسها باشد را پیشنهاد دهد.

در ادامه با فرایند یادگیری و نحوه عملکرد الگوریتم بیشتر آشنا خواهیم شد شکل (۹) روند کلی یکی الگوریتم طبقه‌بندی کننده را نشان می‌دهد.

مجموعه داده آموزش که در این فصل برای اموزش الگوریتم مورد بررسی قرار گرفت ابتدا به عنوان ورودی



شکل ۹ فرایند یادگیری و نحوه عملکرد یک الگوریتم طبقه بند

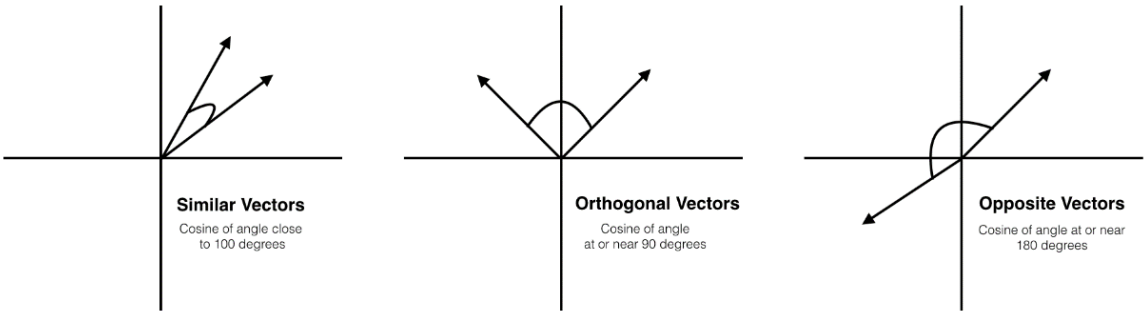
به الگوریتم داده می شود.ضرایب الگوریتم همگی در اساس این مجموعه داده تنظیم شده و مدل به بهترین دقت خود میرسد. برای داده جدیدی که به سیستم وارد شود، الگوریتم شروع می کند بردار مربوط به این داده جدید را با مرکز دسته هایی که داریم مقایسه می کند چنانچه داده جدید به یکی از این میانگین ها یا مرکز دسته ها شباهت بیشتری داشت، به آن کلاس تعلق می گیرد.

معیارهای شباهت، معیارهایی مانند معیارهای فاصله هستند که میزان دور و یا نزدیک بودن دو موجودیت را مشخص می‌کنند. بدیهی است که معیار شباهت با معیارهای فاصله رابطه عکس دارند و به عبارتی هر چه میزان شباهت بیشتر باشد می‌توان نتیجه گرفت فاصله‌ی دو شیئ کمتر است.

برای محاسبه شباهت از روشهای مختلفی می توان استفاده کرد که در زیر برخی از آنها بررسی می گردد.

۱-۳-۳ معیار شباهت کسینوسی[۳۸]:

برای تبدیل کسینوس وزن دهی شده با معکوس درجه ، بین یک ارائه دهنده و عضو مجموعه های مرجع مثبت یا منفی به ویژگی های قابل استفاده جهت تخمین (برآورد) ریسک ، ما میانگین اعضای هر مجموعه را گرفته و بر اساس آن عمل می‌کنیم. نمایی از این روش در شکل (۱۰) مشاهده می‌شود.



شکل ۱۰ نمودار معیار شباهت کسینوسی در حالت های مختلف

در صورت انطباق دو بردار (در این معیار نشانه شباهت کامل است) که زاویه‌ی بین دو بردار صفر می‌باشد مقدار آن برابر **۱** خواهد شد و در کمترین میزان شباهت دو بردار یعنی اگر زاویه بین دو بردار ۱۸۰ درجه باشد نتیجه این معیار **۱** –خواهد شد.

$$\cos(x,y) = \frac{x.y}{||x||.||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

این معیار یکی از پرکاربردترین معیارها است.

دلیل اصلی استفاده ازین معیار در این پایان نامه این است که در ویژگی کسینوسی صفر به معنای خالی و نامشخص بودن حالت ویژگی در مساله است.یعنی زمانی که صفر به معنای این است که برچسب موجود نیس نه اینکه لزوما فرد متخلف نیست.

[۳۸] cosine

۳-۴ طراحی روند روش پیشنهادی

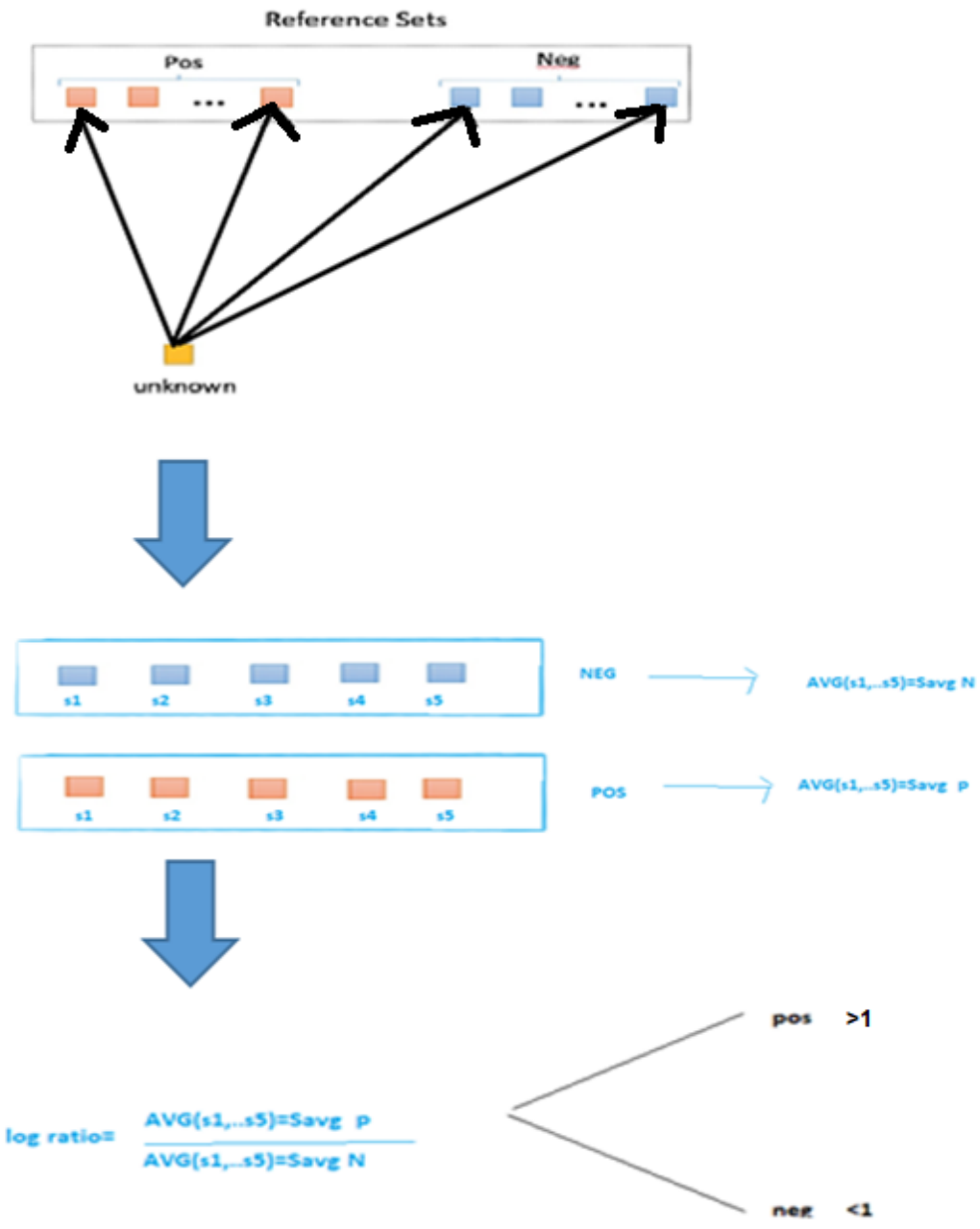
همانطور که گفته شد، متاسفانه داده‌های اولیه‌ی قابل استفاده حجم کمی داشته و نمی‌توان از روش‌های یادگیری ماشین پیچیده برای آموزش و ایجاد مدل استفاده کرد. از طرفی مدل نهایی مورد استفاده‌ی دستگاه‌های نظارتی پزشکی و حقوقی خواهد بود که الزاما تخصصی در هوش مصنوعی نداشته و از مدل‌های ساده و قابل تفسیر بیش‌تر استقبال می‌کنند. به دو دلیل مذکور ایجاد مدلی مرکزی که بتواند با حجم داده‌ی کم نیز آموزش ببینید و نیاز به آموزش مجدد نداشته لازم می‌باشد. روش‌های مبتنی بر گراف از مثال‌های بسیار خوبی برای این‌گونه مسائل می‌باشند.

در این بخش روند ایجاد یک مدل مبتنی بر گراف شرح داده می‌شود.

نمای شماتیکی از نحوه کار الگوریتم در شکل (۱۱) ارائه شده است. همان‌طور که دیده می‌شود داده‌ی آموزش مجموعه‌ی آموزش به دوگروه منفی (نمونه‌های سالم) و مثبت (نمونه‌های متقلب) تقسیم‌بندی می‌شود. گراف این مساله وزن دار نیست و لذا اعضای هر کلاس ارتباط کامل باهم و هیچ ارتباطی با کلاس دیگر ندارند.

برای هر نمونه داده‌ی جدید مقدار شباهت بردار ویژگی‌های آن با بردار ویژگی های هر گروه محاسبه می‌شود. در اینجا از شباهت کسینوسی استفاده شده‌است. اما نگارنده اصراری به استفاده از آن ندارد و هر شباهت دیگری قابل استفاده می‌باشد. بعد از محاسبه‌ی شباهت‌ها دو بردار شباهت برای نمونه‌ی جدید بدست می‌آید. بردار اول شباهت با گروه منفی و بردار دوم شباهت با گروه مثبت. این دو بردار نزولی مرتب می‌شوند. سپس میانگین ۵ عدد بزرگ‌تر در بردار شباهت محاسبه می‌شود تا دو امتیاز عضویت در گروه منفی و امتیاز عضویت در گروه مثبت محاسبه شود. استفاده از همسایگی برای جلوگیری از ایجاد مشکل در ارتباط با نمونه داده‌های پرت می‌باشد. لذا مدل همچون روش چند نزدیک‌ترین همسایگی[۳۹] (KNN) به صورت محلی عمل کرده و برای نمونه‌ی جدید تصمیم‌گیری می‌کند. تعداد همسایگی در این تحقیق ۵ درنظر گرفته‌شده است. این مقدار بهینه نشده‌است.

[۳۹] K- Nearest Neighbor



شکل ۱۱ نمای شماتیکی از نحوه‌ی کار الگوریتم

بعد از محاسبه‌ی امتیازهای میانگین محاسبه شده که عضویت به دو گروه را نشان می‌داد، مقدار log ratio یا نسبت بخت محاسبه می شود. در نهایت از روی مقدار log ratio و فرمول مربوط به ان طبق زیر گروه جدید به یکی از زیر گراف ها نسبت داده می‌شود.

$$\log ratio = \frac{avg(p_1, \ldots., p_n)}{avg(n_1, \ldots., n_{n'})} \begin{cases} pos & > 1 \\ neg & < 1 \end{cases}$$

۵-۳ جمع‌بندی فصل

در این فصل پس از بیان رویکرد مورد استفاده برای حل این مساله توسط این پایان نامه به بررسی نکات مهم در پیاده سازی الگوریتم پرداخته وروش کار آن تشریح گردید. در ادامه باتوجه به تحقیقات صورت گرفته الگوریتم پیاده سازی شده و نتایج و عملکرد آن در موقعیت های مختلف و همچنین در مقایسه با رقبا سنجیده می شود.

فصل ٤

۴ ارزیابی روش پیشنهادی و گزارش نتایج الگوریتم

۴-۱ مقدمه

تا این بخش با مباحث مربوط به داده‌های سلامت آشنا شده و تقلب در این داده‌ها تعریف شد. الگوریتم های مبتنی بر یادگیری ماشین که روی کلان داده‌های مربوط به این حوزه ازمایش و بررسی شدند مرور شد و روش‌هایی که تا کنون وجود دارد دسته بندی شدند.الگوریتم پیشنهادی این پایان نامه در فصل سوم مطرح و جزیات پیاده‌سازی آن و نحوه عملکردش بیان شد. در فصل پنجم پس از پیاده‌سازی این الگوریتم به بررسی نتایج آن پرداخته و عملکرد الگوریتم را از جنبه‌های مختلف مورد سنجش قرار می دهیم.

۴-۲ معیارهای ارزیابی

انتخاب معیار ارزیابی مناسب برای بررسی عملکرد الگوریتم این امکان را فراهم می کند که بتوانیم علاوه بر بررسی و ارزیابی عملکرد الگوریتم، امکان منقایسه نتایج خود را با سایر الگوریتم‌ها هم فراهم نماییم. از این رو در این تحقیق، ما از معروف ترین معیارهای مورد استفاده در مبحث تحلیل احساسات استفاده می‌کنیم. از آنجایی که مساله تحلیل احساسات، خود زیرمجموعه‌ای از روش‌های با ناظر و یا همان طبقه‌بندی است، معروف‌ترین معیارهای عرصه طبقه‌بندی بهترین گزینه برای سنجش عملکرد الگوریتم ما هستند.

در ارزیابی تقلب پزشکی دو حالت در نظر می‌گیریم; ارتکاب تقلب و یا عدم ارتکاب تقلب. در این تحقیق کلاس مثبت یا کلاس هدف، ارتکاب تقلب است و کلاس منفی، عدم ارتکاب تقلب است.

۱-۲-۴ ماتریس درهم ریختگی[۴۰]

این ماتریس از معروف‌ترین ابزارهای سنجش عملکرد روش‌های باناظر است و حتی گاها در روش‌های بدون ناظر هم از آن استفاده می شود. (Wan بدون تاریخ)

ماتریس پیچیدگی تعداد نمونه های واقعی را با تعداد نمونه های پیش بینی شده مقایسه میکند. با توجه به ماتریس نتایج، ما از AUC[۶۷,۶۸] برای ارزیابی کارایی تشخیص تقلب استفاده می کنیم AUC ناحیه زیرمنحنی(Receiver Operating Characteristic)ROC است و ROC مقایسه بینFalse Positive وTrue Positive است. Recall از TP/TP+FN بدست می آید. تعاریف برای TP و TN و FPوFN مستقیما از ماتریس پیچیدگی به صورت زیر بدست می آیند:

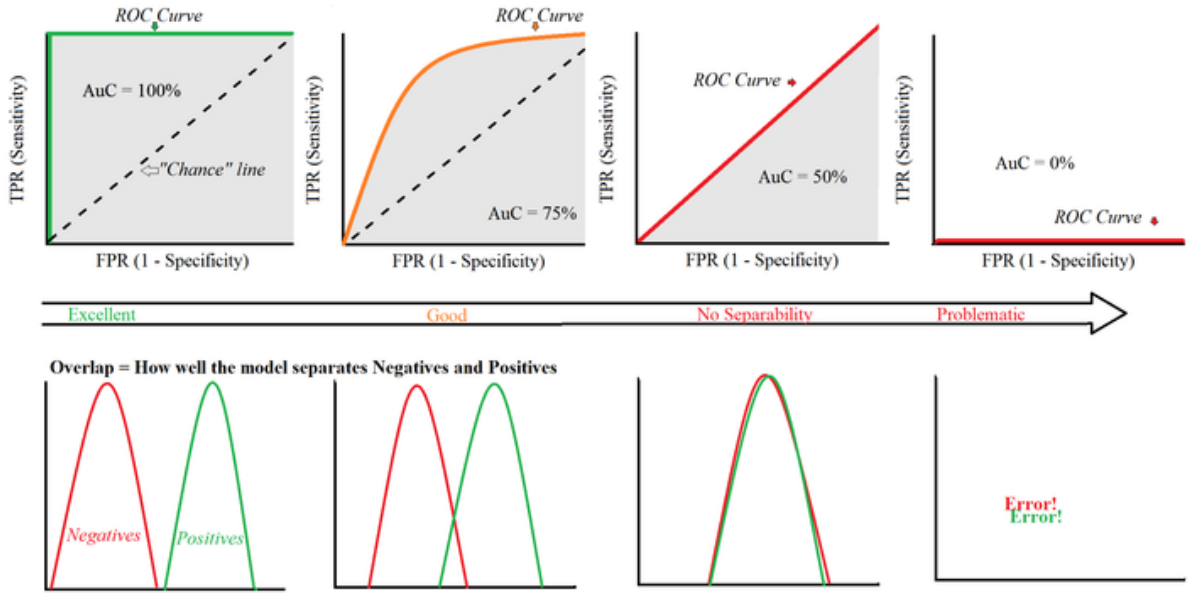
- True Positive(TP) : تعداد نمونه های مثبت واقعی که به درستی مثبت پیش بینی شده اند.
- True Negative(TN) :تعداد نمونه های منفی واقعی که به درستی منفی پیش بینی شده اند.
- False Positive(FP) :تعداد نمونه های منفی واقعی که به اشتباه مثبت پیش بینی شده اند.
- False Negative(FN) : تعداد نمونه های مثبت واقعی که به اشتباه منفی پیش بینی شده اند.

[40] Consusion Matrix

۲-۲-۴ منحنی AUC

منحنیAUC یک ارزیابی فراگیر از یادگیرنده است که ارزیابی را در سراسر آستانه های تصمیم گیری به تصویر می کشد. نتایج AUC شکل (۱۲) چند نمونه از آن را نشان می‌دهد. در یک بازه بین ۰ تا ۱ هستند که یک طبقه بند خوب دارای مقدار ۱ در AUC می باشد، طبقه بند تصادفی مقدار ۰.۵ دارد و مقادیر AUC کمتر از ۰.۵ نشان دهنده ی بایاس به سمت یک کلاس خاص است ثابت شده است که AUCبرای نامتوازن بودن کلاس موثر است.

سطح زیر منحنی راک Area under the ROC curve معمولاً به اختصار AUC نامیده می‌شود.



شکل ۱۲ منحنی AUC و معنی آن

شکل (۱۲) منحنی راک (ROC Curve) است که سطح زیر منحنی آن مشخص شده است. به این سطح (Area under the ROC Curve = AuC)گفته میشود. دیده میشودمقدار عددی AUC عددی بین صفر تا یک است و نشان می‌دهد قدرت تشخیص یا درستی نتایج یک آزمون چقدر می‌باشد. درستی نتایج آزمون به این بستگی دارد که روش آزمون چقدر توانایی تفاوت درست نشان دادن نتایج مثبت صحیح (TP) و منفی صحیح (TF)را دارد. اگر این عدد به یک نزدیک باشد، به معنای آن است که داده‌ها عموماً در بالای خط نیم‌ساز قرار

گرفته‌اند و میزان نرخ مثبت صحیح بالا است و روش آزمون از قدرت تشخیص یا درستی مناسبی برخوردار است. اعداد AUC نزدیک به ۰.۵ همان برابری نرخ مثبت صحیح و نرخ مثبت کاذب را نشان می‌دهد و اعداد کمتر از ۰.۵ بیانگر بالاتر بودن نرخ مثبت کاذب در مقایسه با نرخ مثبت صحیح است .[۹۹]

۳-۲-۴ حساسیت[۴۱]

این مفهوم از جنس احتمال و در نتیجه عددی بین صفر و یک می‌باشند و می‌توان آنها را بر حسب درصد (بین صفر و صد) بیان نمود. حساسیت (sensitivity)به احتمال مثبت شدن صحیح نتیجه آزمون وقتی که نمونه دارای آلودگی است، اشاره می‌کند که می‌توان آن را به صورت زیر به دست آورد .حساسیت را نرخ مثبت صحیح (TPR)نیز می‌نامند.

$$\text{حساسیت} = \frac{TP}{TP + FN}$$

۴-۲-۴ تشخیص‌پذیری [۴۲]

تشخیص‌پذیری یا ویژگی نیز به احتمال منفی شدن صحیح نتیجه آزمون وقتی که نمونه سالم (فاقد آلودگی) است، اشاره می‌کند. تشخیص‌پذیری یا ویژگی را می‌توان به صورت زیر به دست آورد. تشخیص‌پذیری یا ویژگی را نرخ منفی صحیح (TNR)نیز می‌گویند.

$$\text{تشخیص پذیری} = \frac{TN}{TN + FP}$$

---

۴۱ Sensitivity

۴۲ specificity

جدول ۶ دسته بندی معیارهای ارزیابی عملکرد یک طبقه بند

		Condition		
		Positive	Negative	
Test outcome	Positive	True Positive (TP)	False Positive (FP)	→ Positive predictive value
	Negative	False Negative (FN)	True Negative (TN)	→ Negative predictive value
		↓ Sensitivity	↓ Specificity	

دسته بندی معیارهای ارزیابی استفاده شده به صورت جدول (۶) است:

۴-۳ نتایج عملکرد الگوریتم

اگرچه حل یک مساله شخص از رویکرد های مختلف بسیار با ارزش و به خودی خود نواوری محسوب می شود. اما در نهایت عملکرد مناسب الگوریتم است که باعث می شود از آن در حل مسایل دنیای واقعی استفاده شد. راه حل خلاقانه ای که نتایج خوبی ارایه ندهد و دقت خوبی نداشته باشد خیلی زود محو خواهد شد. لذا ضروری است که نتایج الگوریتم را برای دسته بندی و روی داده هایی به عنوان تست بسنجیم. لذا در یک روال اعتبارسنجی ۱۰ برابری[۴۳] داده‌های به صورت تصادفی به ده بخش تقسیم و در یک روند تکراری با تعداد ۱۰، هربار با ۹ بخش مدل ساخته شده و روی یک بخش باقی مانده اعتبارسنجی انجام شد. نتایج این اعتبارسنجی‌ها به چند روش مختلف ضبط شده و در آخر مقادیر میانگین گرفته و همچنین تغییرات آن‌ها نیز گزارش شد.

۴-۳-۱ نتایج مقدارهای TP,TN,FP,FN

پس از اجرای الگوریتم جدول زیر برای نتایج مربوط به TP,TN,FP,FN بدست آمد که در جدول (۷) مشاهده می‌شود. از آنجایی که این اعدا درک و خوانایی سختی دارند لذا در ادامه با استفاده از سایر معیار های معرفی شده به درک بهتری از عملکرد الگوریتم خواهیم رسید.

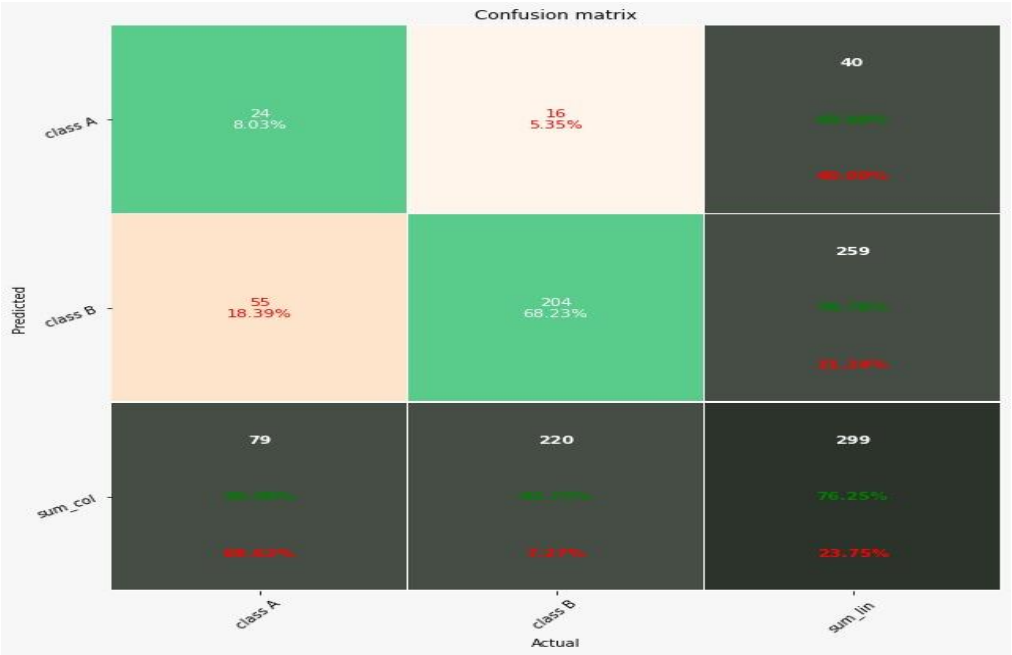
[۴۳] 10 – Fold Cross Validation

جدول۷ نتایج معیار های پایه برای الگوریتم

		Real	
		Positive	Negative
prediction	Positive	TP=24	FP=16
	Negative	FN=55	TN=204

۴-۳-۲ ماتریس درهم ریختگی

در شکل (۱۳) ماتریس در هم ریختگی حاصل از اجرای الگوریتم روی مجموعه داده‌ی انتخابی دیده می‌شود.



شکل ۱۳ ماتریس در همریختگی برای نتایج روش پیشنهادی

۴-۴ مقایسه با سایر الگوریتم های موجود

جدول۸ مقایسه نتایج مدل با سایر الگوریتم ها

Specificity	Sensivity	**AUC**	**الگوریتم**
0.676	0.536	**0.629**	LOF40
0.679	0.497	**0.613**	KNN1
0.645	0.527	**0.603**	URF100
0.650	0.463	**0.555**	AE50_Tanh
0.436	0.712	**0.554**	IF100
0.5	0.5	**0.64**	**Similarity-based Anomaly Detection**

در جدول (۸) مشاهده می‌شود الگوریتم ارائه شده با نام Similarity-based Anomaly Detection از نظر AUC نتایج بالاتری نسبت به الگوریتم‌های پیشین دارد اما از نظر حساسیت و تشخیص پذیری عملکرد مناسبی ندارد. این نتایج در حالی است که نوع روش شباهت‌های دیگر استفاده نشده و همین‌طور مقدار همسایگی بهینه نشده‌است.

۴-۵ بررسی بازه اطمینان نتایج الگوریتم

در مباحث آمار و داده‌کاوی، از عبارت فاصله اطمینان[۴۴] استفاده می‌کنند تا نشان دهند که تقریبا مطمئن هستیم یک فاصله یا محدوده‌ای عددی، شامل پارامتر مورد جامعه است. لذا فاصله اطمینان نوعی برآورد فاصله‌ای در نظرگرفته شده و هرقدر کوچکتر باشد نشان دهنده بهتر بودن نتایج است . اغلب فاصله اطمینان را با CI نشان می‌دهند. این بازه یک کران بالا و پایین دارد که حد اطمینان را مشخص کند. در این الگوریتم برای نتایج AUC

[۴۴] Confidence Interval

Sensitivity ,و Specifity بازه اطمینان محاسبه شد و همان‌طور که در جدول (۹) از نتایج مشخص است این بازه در حد قابل قبولی است.

جدول۹ بازه اطمینان برای نتایج الگوریتم

معیار	نتیجه	حدبالای بازه اطمینان	حدپایین بازه اطمینان
AUC	0.64	-0.0246	0.0246
Sensitivity	0.5	-0.239	0.239
Specificity	0.5	-0.239	0.239

۴-۶ جمع بندی فصل

بعد از پیاده سازی روش ارایه شده در فصل چهارم، در این فصل به انتخاب معیار مناسب پرداخته شد تا امکان ارزیابی دقیق الگوریتم را فراهم سازد. در ادامه با سنجش الگوریتم توسط معیار های مختلف و مقایسه آن با نتایج ارایه شد توسط بقیه الگوریتم های دسته بندی که در این موضوع محبوب هستند و مورد استفاده قرار می گیرند، عملکرد واقعی الگوریتم ارزیابی شد.

فصل ۵

۵ نتایج و کارهای آتی

۵-۱ مقدمه

اندازه بخش مراقبت سلامت و حجم زیاد پولی که شامل آن است، آن را برای اهداف تقلب جذاب می‌سازد. تقلب مراقبت سلامت بر اساس تعریف NHCAA[45] یک فریب عمدی یا ارائه اطلاعات نادرست است که توسط یک شخص یا یک موجودیت با علم به اینکه این فریب می‌تواند منجربه مقداری سود غیرمجاز برای آن فرد یا موجودیت شود انجام می‌شود. هزینه بهداشت و درمان با توجه به جمعیت، اقتصاد، جامعه، و تغییرات قانون به سرعت در حال افزایش است. این افزایش در هزینه‌های بهداشت و درمان بر دولت و سیستم‌های بیمه سلامت خصوصی تأثیر

[45] National HealthCare Anti-Fraud Association

می‌گذارد. رفتارهای متقلبانه‌ی ارائه‌دهندگان بهداشت و درمان و بیماران با تحمیل هزینه‌های غیرضروری به مشکلی جدی برای سیستم‌های بیمه تبدیل شده است. بنابراین، حوزه سلامت به یک منبع هزینه‌ای قابل توجه در بسیاری از کشورها تبدیل شده است. وسیع بودن حوزه سلامت و حجم زیاد مالی باعث شده تا حوزه سلامت یه یک هدف جذاب برای کلاهبرداری تبدیل شود. شرکت‌های بیمه روش‌هایی را برای تشخیص تقلب ایجاد می‌کنند که عمدتا برگرفته از تجارب خبرگان بوده و کمتر به روش‌های مبتنی بر تحلیل داده متکی است.

فقط خسارت مالی نگرانی عمده نیست بلکه تقلب به شدت مانع از ارائه مراقبت با کیفیت و امن سیستم مراقبت سلامت آمریکا از بیماران مشروع می‌شود. بنابراین تشخیص تقلب مؤثر برای بهبود کیفیت و کاهش هزینه‌ی خدمات مراقبت بهداشت مهم است. تقلب در حوزه سلامت یک جرم بزرگ است و هزینه‌های شخصی و بودجه‌ای قابل توجهی به افراد، دولت‌ها و جامعه وارد میکند. بنابراین، کشف موثر تقلب برای کاهش هزینه‌ها و بهبود کیفیت سیستم سلامت بسیار مهم است. به منظور دستیابی به کشف موثرتر تقلب، بسیاری از پژوهشگران رویکردهای ضد تقلب پیچیده‌ای بر پایه داده‌کاوی، یادگیری ماشین و دیگر روش‌های تحلیلی توسعه دادند. این رویکردهای جدید ارائه شده دارای مزیت‌هایی مانند یادگیری خودکار الگوهای تقلب از داده‌ها، مشخص کردن احتمال تقلب برای هر مورد و شناسایی گونه‌های جدید تقلب دارند.

۲-۵ نتیجه گیری

بدلیل حجم بسیار زیاد داده‌های این عرصه، نیروی انسانی به تنهایی قادر به تشخیص خطاهای مربوط به این حوزه نخواهد بود و این امر باعث اسیبهای جبران ناپذیر خواهد شد. از این رو لازم است با استفاده از سیستم های یادگیرنده و یا تشخیص الگو استفاده کرده و به کمک آن‌ها درصد خطا را کاهش دهیم.

رویکردهای کشف تقلب را میتوان به صورت کلی به سه دسته‌ی روش‌های آماری، روش‌های یادگیری ماشین با ناظر، روش‌های یادگیری بدون ناظر و روش‌های یادگیری ماشین ترکیبی تقسیم نمود. اگرچه روش‌های آماری میتواند عملکرد سریعی در شناسایی تقلب داشته باشد و عملکرد مطلوبی در شناسایی انواع جدید تقلب دارد، این روش‌ها ممکن است مطالبات قانونی را به عنوان جعلی شناسایی کند که نقطه ضعف بسیار مهمی در میان روش‌های موجود است. همچنین، این روش‌ها نیاز به تخصص و دانش زیادی در زمینه‌ی روش‌های آماری و تشخیص تقلب دارد.

انواع روش‌های داده‌کاوی همانطور که پیشتر بحث شد، دارای مزایا و برتری بیشتری نسبت به دیگر روش‌ها می‌باشد. از مزایای مهم آن می‌توان به سادگی و عدم نیاز به پردازشگر قوی، هزینه‌ی کم‌تر به دلیل عدم نیاز به داده‌ی برچسب‌دار، دارای نرخ کشف کاذب کمتر و کاهش هزینه‌ی برچسب داده‌ها اشاره کرد. درجدول (۱۰) این الگوریتم‌ها به همراه مزایا و معایب آن‌ها آماده‌است.

جدول ۱۰ دسته‌بندی به تفکیک رویکردهای کلی کشف تقلب

رویکردها	مزایا	معایب
روش‌های آماری	• به سرعت ارائه‌دهندگان مشکوک را شناسایی می‌کند. • می‌تواند نوع جدیدی از تقلب را شناسایی کند.	• ممکن است مطالبات قانونی را به عنوان جعلی شناسایی کند • نیاز به بررسی مطالبات پس از ارزیابی آماری. • نیاز به دانش از روش‌های آماری. • نیاز به تخصص مقدم بر تشخیص تقلب
روش‌های یادگیری ماشین با ناظر	• سادگی و عدم نیاز به پردازشگر قوی	• نیاز به متخصص برای داده‌ی برچسب دار • overfitting
روش‌های یادگیری ماشین بدون ناظر	• هزینه‌ی کم‌تر به دلیل عدم نیاز به داده‌ی برچسب‌دار	• تغییر مداوم وزن ها و پارامترها توسط متخصصان برای شناسایی انواع جدید تقلب • تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب • هزینه‌ی بالای داده‌ی برچسب‌دار
روش‌های یادگیری ماشین ترکیبی	. • در مقایسه با روش‌های کاوش فرایند دارای نرخ کشف کاذب[۴۶] کمتر است. • کاهش هزینه‌ی برچسب داده‌ها	• پوشش کم موارد تشخیص تقلب

۴۶ false discovery rate

روش پیشنهادی این پایان نامه روشی مبتنی بر گراف، با تاکید بر شباهت ورودی جدید به سایر داده‌های موجود است. همانطور که بحث شد در این روش داده‌های مجموعه آموزش به دو دسته تقسیم شده و ورودی جدید بر اساس میزان شباهتی که به هر کلاس دارد در یکی از دو گروه تقلب یا مشوک ویا گروه عادی قرار می گیرد.

نتایج این روش نشان ازین داد که علاوه بر سرعت بالا به دلیل عدم نیاز به آموزش مجدد شبکه، دقت خوبی داشته و باسایر الگوریتم های این حوزه رقابت می کند. بازه اطمینان برای نتایج محاسبه و در فصل پنجم بیان شد و اعداد نشان دهنده این مورد بودند که نتایج الگوریتم قابل اتکا است.

متاسفانه به دلیل در دسترس نبودن اطلاعات دیگری مانند زمان اجرای الگوریتم ها، امکان مقایسه الگوریتم پیشنهادی این پایان نامه با سایر روش‌ها فراهم نیست. که البته این مورد ازین جهت قابل درک است که در مباحث مربوط به تقلب موضوع دقت اهمیت به مراتب بیشتری از سرعت و یا میزان حافظه مصرفی و ... دارد.

در نهایت با بیان این نکته روشهای مبتنی بر شباهت در گرافها همگی به نحوی از دسته روشهای شناخت ناهنجاری محسوب میشوند، بیان می شود که این کار دست محقق رابرای کارهای اتی و طبقه بندی حتی نوع ناهنجاری ها، یا دسته بندی گراف به صورت کلی به انجمن و سپس بررسی اینکه امکان تخلف در کدام انجمن ها بیشتر و محتمل تر است را فراهم می کند.

در ادامه بعضی از کارهای آتی که به دلیل کمبود وقت وسایر محدویت ها در این بازه زمانی قابل پیاده سازی در این پایان نامه نبود مطرح می گردد.

۴-۵ کارهای آتی

استفاده از الگوریتم های یادگیری ماشین به دلیل پویا بودن و گستردگی همیشه این امکان را فراهم می سازد که با کشف و معرفی متدها و پارامترهای جدیدی که مطرح می شود، مدل را بهبود بخشید و یا حداقل آن را پیاده سازی و نتایج را بررسی کرد. در این پایان نامه امکان ادامه کار در هر کدام از حوزه های زیر فراهم است و ممکن است منجر به بهبود عملکرد الگوریتم گردد:

- استفاده از متریک های جدید برای بررسی و اندازه گیری شباهت
- استفاده از معیار های ارزیابی دیگر

- بررسی تاثیر نویز بر روی عملکرد الگوریتم و اینکه تا چه اندازه الگوریتم نسبت به اختلالات ورودی ثبات دارد.
- امکان تقسیم کلاس ها به چند کلاس بجای حالت دوتایی(احتمال بالای۷۵٪ تقلب- احتمال بالای ۵۰٪تقلب- احتمال زیر۵۰٪ تقلب- احتمال زیر ۲۵٪ تقلب) و ...

مراجع

[۱] M. E. Johnson and N. Nagarur, "Multi-stage methodology to detect health insurance claim fraud," *Health care management science,* vol. 19, no. 3, pp. 249-260, 2016.

[۲] H. Sadeghian N, "Assessment and recognition the trueness of the assurance claims using data mining techniques based on the supervised learn," Industrial Management, Shahrood University, 2016 .

[۳] E. A. Duman and Ş. Sağıroğlu, "Heath care fraud detection methods and new approaches," in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017: IEEE, pp. 839-844 .

[٤] ح. ع. م. ج. تارخ, "کشف تقلب در بیمه سلامت بر اساس رویکرد داده کاوی," کنفرانس بین المللی پژوهش های نوین در مدیریت ، اقتصاد ، توانمندی صنعت جهانگردی در توسعه, ۲۰۱۷.

[٥] M. K. Wynia, D. S. Cummins, J. B. VanGeest, and I. B. Wilson, "Physician manipulation of reimbursement rules for patients: between a rock and a hard place," *Jama,* vol. 283, no. 14, pp. 1858-1865, 2000.

[٦] د. س. ح. ه. د. ع. ز. د. ع. ربیعي, "ارزش نسبی خدمات و مراقبتهای سلامت در جمهوری اسلامی ایران," ۱۳۹۳.

[۷] L. K. Branting, F. Reeder, J. Gold, and T. Champney, "Graph analytics for healthcare fraud risk estimation," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016: IEEE, pp. 845-851 .

[۸] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Systems with Applications,* vol. 31, no. 1, pp. 56-68, 2006.

[۹] H. Shin, H. Park, J. Lee, and W. C .Jhee, "A scoring model to detect abusive billing patterns in health insurance claims," *Expert Systems with Applications,* vol. 39, no. 8, pp. 7441-7450, 2012.

[۱۰] J. Li, K.-Y. Huang, J. Jin, and J. Shi, "A survey on statistical methods for health care fraud detection," *Health care management science,* vol. 11, no. 3, pp. 275-287, 2008.

[۱۱] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119,* 2010.

[۱۲] J.-H. Wang, Y.-L. Liao, T.-m. Tsai, and G. Hung, "Technology-based financial frauds in Taiwan: issues and approaches," in *2006 IEEE International Conference on Systems, Man and Cybernetics*, 2006, vol. 2: IEEE, pp. 1120-1124 .

[۱۳] I. Bose and R. K. Mahapatra, "Business data mining—a machine learning perspective," *Information & management,* vol. 39, no. 3, pp. 211-225, 2001.

[۱٤] E. Turban, R. Sharda, and D. Delen, "Decision support and business intelligence systems (required)," *Google Scholar,* 2010.

[١٥] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," *AI magazine,* vol. 13, no. 3, pp. 57-57, 1992.
[١٦] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *IEEE International Conference on Networking, Sensing and Control, 2004*, 2004, vol. 2: IEEE, pp. 749-754 .
[١٧] W. D. Savedoff and K. Hussmann, "The causes of corruption in the health sector: a focus on health care systems," *Transparency International. Global Corruption Report,* 2006.
[١٨] B. Manjula, S. Sarma, A. Govardhan, and L. Naik, "DFFS: Detecting Fraud in Finance Sector," *Int. J. Adv. Eng. Sci. Technol,* vol. 9, no. 2, pp. 178-182, 2011.
[١٩] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *Journal of Network and Computer Applications,* vol. 68, pp. 90-113, 2016.
[٢٠] H. A. Abbass, J. Bacardit, M. V. Butz, and X. Llora, "Online adaptation in learning classifier systems: stream data mining," *Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign, IlliGAL Report,* no. 2004031, 2004.
[٢١] D. Malekian and M. R. Hashemi, "An adaptive profile based fraud detection framework for handling concept drift," in *2013 10th International ISC Conference on Information Security and Cryptology (ISCISC)*, 2013: IEEE, pp. 1-6 .
[٢٢] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR),* vol. 46, no. 4, pp. 1-37, 2014.
[٢٣] R. J. Bolton and D. J .Hand, "Unsupervised profiling methods for fraud detection," *Credit scoring and credit control VII,* pp. 235-255, 2001.
[٢٤] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning,* vol. 3, no. 1, pp. 1-130, 2009.
[٢٥] Q. Liu and M. Vasarhelyi, "Healthcare fraud detection: A survey and a clustering model incorporating Geo-location information," in *29th world continuous auditing and reporting symposium (29WCARS), Brisbane, Australia*, 2013 .
[٢٦] V. López, A. Fernández, J. G. Moreno-Torres, and F. Herrera, "Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics," *Expert Systems with Applications,* vol. 39, no. 7, pp. 6585-6608, 2012.
[٢٧] C. S. Hilas and J. N. Sahalos, "An application of decision trees for rule extraction towards telecommunications fraud detection," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, 2007: Springer, pp. 1112-1121 .
[٢٨] S. Viaene, R. A. Derrig, and G. Dedene, "A case study of applying boosting Naive Bayes to claim fraud diagnosis," *IEEE Transactions on Knowledge and Data Engineering,* vol. 16, no. 5, pp. 612-620, 2004.
[٢٩] T. Lane and C. E. Brodley, "Temporal sequence learning and data reduction for anomaly detection," *ACM Transactions on Information and System Security (TISSEC),* vol. 2, no. 3, pp. 295-331, 1999.
[٣٠] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[٣١] P. L. Brockett, L. L. Golden, J. Jang, and C. Yang, "A comparison of neural network, statistical methods, and variable choice for life insurers' financial distress prediction," *Journal of Risk and Insurance,* vol. 73 ,no. 3, pp. 397-419, 2006.
[٣٢] J. Ai, P. L. Brockett, and L. L. Golden, "Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables," *North American Actuarial Journal,* vol. 13 ,no. 4, pp. 438-458, 2009.
[٣٣] V. Almendra and D. Enachescu, "A supervised learning process to elicit fraud cases in online auction sites," in *2011 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, 2011: IEEE, pp ١٦٨-١٧٤ .
[٣٤] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in *International Conference on Intelligent Data Engineering and Automated Learning*, 2007: Springer, pp. 178-187 .
[٣٥] M. S. Islam, M. M. Hasan, X. Wang, and H. D. Germack, "A systematic review on healthcare analytics: application and theoretical perspective of data mining," in *Healthcare*, 2018, vol. 6, no. 2: Multidisciplinary Digital Publishing Institute, p. 54 .
[٣٦] A. Kusiak, C. A. Caldarone, M. D. Kelleher, F. S. Lamb, T. J. Persoon, and A. Burns, "Hypoplastic left heart syndrome: knowledge discovery with a data mining approach," *Computers in Biology and Medicine,* vol. 36, no. 1, pp. 21-40, 2006.
[٣٧] N. N. Taleb, *The black swan: The impact of the highly improbable*. Random house, 2007.
[٣٨] N. Cercone, X. An, J. Li, Z. Gu, and A. An, "Finding best evidence for evidence-based best practice recommendations in health care: the initial decision support system design," *Knowledge and information systems,* vol. 29, no. 1, p. 159, 2011.
[٣٩] Y. Huang, P. McCullagh, N. Black, and R. Harper, "Feature selection and classification model construction on type 2 diabetic patients' data," *Artificial intelligence in medicine,* vol. 41, no. 3, pp. 251-262, 2007.
[٤٠] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare informatics research,* vol. 19, no. 2, pp. 121-129, 2013.
[٤١] R. S. Santos, S. M. Malheiros, S. Cavalheiro, and J. P. De Oliveira, "A data mining system for providing analytical information on brain tumors to public health decision makers," *Computer methods and programs in biomedicine,* vol. 109, no. 3, pp. ٢٨٢-٢٠١٣.269 ,
[٤٢] C.-P. Shen *et al.*, "A data-mining framework for transnational healthcare system," *Journal of medical systems,* vol. 36, no. 4, pp. 2565-2575, 2012.
[٤٣] L. Duan, W. N. Street, and E. Xu, "Healthcare information systems: data mining methods in the creation of a clinical recommender system," *Enterprise Information Systems,* vol. 5, no. 2, pp. 169-181, 2011.
[٤٤] D. Toshniwal and S. Yadav, "Adaptive outlier detection in streaming time series," in *Proceedings of International Conference on Asia Agriculture and Animal, ICAAA, Hong Kong*, 2011, vol. 13, pp. 186-192 .

[٤٥] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR),* vol. 41, no. 3, pp. 1-58, 2009.
[٤٦] S. Bendre, "Outliers in Statistical Data," ed: JSTOR, 1994.
[٤٧] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review,* vol. 22, no. 2, pp. 85-126, 2004.
[٤٨] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egyptian informatics journal,* vol. 17, no. 2, pp. 199-216, 2016.
[٤٩] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," *Social Networks,* vol. 39, pp. 62-70, 2014.
[٥٠] R. Hassanzadeh ,R. Nayak, and D. Stebila, "Analyzing the effectiveness of graph metrics for anomaly detection in online social networks," in *International Conference on Web Information Systems Engineering*, 2012: Springer, pp. 624-630 .
[٥١] L. Mookiah, W. Eberle, and L. Holder, "Discovering Suspicious Behavior Using Graph-Based Approach," in *The Twenty-Eighth International Flairs Conference*, 2015 .
[٥٢] F. Moradi, T. Olovsson, and P. Tsigas, "Overlapping communities for identifying misbehavior in network communications," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2014: Springer, pp. 398-409 .
[٥٣] B. Perozzi, L. Akoglu, P. Iglesias Sánchez, and E. Müller, "Focused clustering and outlier detection in large attributed graphs," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 1346-1355 .
[٥٤] A. Chaudhary, H. Mittal, and A. Arora, "Anomaly Detection Using Graph Neural Networks," in *2019 International Conference on Machine Learning ,Big Data, Cloud and Parallel Computing (COMITCon)*, 2019: IEEE, pp. 346-350 .
[٥٥] J. Sun, Y. Xie, H. Zhang, and C. Faloutsos, "Less is more: Sparse graph mining with compact matrix decomposition," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 1, no. 1, pp. 6-22, 2008.
[٥٦] C. C. Aggarwal, Y. Zhao, and S. Y. Philip, "Outlier detection in graph streams," in *2011 IEEE 27th International Conference on Data Engineering*, 2011: IEEE, pp. 399-409 .
[٥٧] N. A. Heard, D. J. Weston, K. Platanioti, and D. J. Hand, "Bayesian anomaly detection methods for social networks," *The Annals of Applied Statistics,* vol. 4, no. 2, pp. 645-662, 2010.
[٥٨] W. Eberle and L. Holder, "A partitioning approach to scaling anomaly detection in graph streams ",in *2014 IEEE International Conference on Big Data (Big Data)*, 2014: IEEE, pp. 17-24 .
[٥٩] D. Y. Perwej, "An Experiential Study of the Big Data," *International Transaction of Electrical and Computer Engineers System, ISSN (Print): 2373-1273, ISSN (Online): 2373-1281,USA,* vol. Volume 4, pp. Page 14-25, 03/24 2017, doi: 10.12691/iteces-4-1-3.
[٦٠] B. Ruhnau, "Eigenvector-centrality—a node-centrality?," *Social networks,* vol. 22, no. 4, pp. 357-365, 2000.

[٦١] M. Barthelemy, "Betweenness centrality in large complex networks," *The European physical journal B,* vol. 38, no. 2, pp. 163-168, 2004.
[٦٢] F. C. Cunningham, G. Ranmuthugala, J. Plumb, A. Georgiou, J. I. Westbrook, and J. Braithwaite, "Health professional networks as a vector for improving healthcare quality and safety: a systematic review," *BMJ quality & safety,* vol. 21, no. 3, pp. 239-249, 2012.
[٦٣] F.-M. Liou, Y.-C. Tang, and J.-Y. Chen, "Detecting hospital fraud and claim abuse through diabetic outpatient services," *Health care management science* , vol. 11, no. 4, pp. 353-358, 2008.
[٦٤] P. A. Ortega, C. J. Figueroa, and G. A. Ruz, "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile," *DMIN,* vol. 6, pp. 26-29, 2006.
[٦٥] T. M. Padmaja, N. Dhulipalla, R. S. Bapi, and P. R. Krishna, "Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection," in *15th International Conference on Advanced Computing and Communications (ADCOM 2007)*, 2007: IEEE, pp. 511-516 .
[٦٦] J .A. Major and D. R. Riedinger, "EFD: A hybrid knowledge/statistical-based system for the detection of fraud," *International Journal of Intelligent Systems,* vol. 7, no. 7, pp. 687-703, 1992.
[٦٧] H. He, W. Graco, and X. Yao, "Application of genetic algorithm and k-nearest neighbour method in medical fraud detection," in *Asia-Pacific Conference on Simulated Evolution and Learning*, 1998: Springer, pp. 74-81 .
[٦٨] M. Kumar, R. Ghani, and Z.-S. Mei, "Data mining to predict and prevent errors in health insurance claims processing," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 65-74 .
[٦٩] T. Ekin, F. Ieva, F. Ruggeri, and R. Soyer, "Statistical medical fraud assessment: exposition to an emerging field," *International Statistical Review,* vol. 86, no. 3, pp. 379-402, 2018.
[٧٠] C. Lin, C.-M. Lin, S.-T. Li, and S.-C. Kuo, "Intelligent physician segmentation and management based on KDD approach," *Expert Systems with Applications,* vol. 34, no. 3 ,pp. 1963-1973, 2008.
[٧١] R. M. Musal, "Two models to investigate Medicare fraud within unsupervised databases," *Expert Systems with Applications,* vol. 37, no. 12, pp. 8628-8633, 2010.
[٧٢] T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of bayesian methods in detection of healthcare fraud," *chemical engineering Transaction,* vol. 33, 2013.
[٧٣] G. C. Capelleveen, ""Outlier based predictors for health insurance fraud detection within US Medicaid"," *MS thesis. University of Twente,* 2013.
[٧٤] Y .Shan, D. W. Murray, and A. Sutinen, "Discovering inappropriate billings with local density based outlier detection method," in *Proceedings of the Eighth Australasian Data Mining Conference-Volume 101*, 2009, pp. 93-98 .
[٧٥] M. Tang, B. S. U. Mendis, D .W. Murray, Y. Hu, and A. Sutinen, "Unsupervised fraud detection in Medicare Australia," in *Proceedings of the Ninth Australasian Data Mining Conference-Volume 121*, 2011, pp. 103-110 .

[٧٦] L. F. Carvalho, C. H. Teixeira, W. Meira, M. Ester, O. Carvalho ,and M. H. Brandao, "Provider-consumer anomaly detection for healthcare systems," in *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 2017: IEEE, pp. 229-238 .

[٧٧] V. S. Iyengar, K. B. Hermiz, and R. Natarajan, "Computer-aided auditing of prescription drug claims," *Health care management science,* vol. 17, no. 3, pp. 203-214, 2014.

[٧٨] R. A. Bauder and T. M. Khoshgoftaar, "A probabilistic programming approach for outlier detection in healthcare claims," in *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, 2016: IEEE, pp. 347-354 .

[٧٩] R. M. Musal and T. Ekin, "Medical overpayment estimation: A Bayesian approach," *Statistical Modelling,* vol. 17, no. 3, pp. 196-222, 2017.

[٨٠] S. Rao and P. Gupta, "Implementing improved algorithm over apriori data mining association rule algorithm 1," 2012.

[٨١] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, 1994, vol. 1215, pp. 487-49 .٩

[٨٢] Y. Ji, H. Ying, J. Tran, P. Dews, A. Mansour, and R. M. Massanari, "Mining infrequent causal associations in electronic health databases," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011: IEEE, pp. 421-428 .

[٨٣] B. Patil, R. Joshi, and D. Toshniwal, "Association rule for classification of type-2 diabetic patients," in *2010 second international conference on machine learning and computing*, 2010: IEEE, pp. 330-334 .

[٨٤] U. Abdullah, J. Ahmad, and A. Ahmed, "Analysis of effectiveness of apriori algorithm in medical billing data mining," in *2008 4th International Conference on Emerging Technologies*, 2008: IEEE, pp. 327-331 .

[٨٥] M. Ilayaraja and T. Meyyappan, "Mining medical data to identify frequent diseases using Apriori algorithm," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013: IEEE, pp. 194-199 .

[٨٦] R. J. Bolton and D. J. Hand, "Peer group analysis–local anomaly detection in longitudinal data," *Technical Report*.٢٠٠١ ,

[٨٧] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.

[٨٨] P. Travaille, "Electronic fraud detection in the US Medicaid Health Care Program," University of Twente, 2011 .

[٨٩] P. Travaille, R. M. Müller, D. Thornton, and J. Van Hillegersberg, "Electronic Fraud Detection in the US Medicaid Healthcare Program: Lessons Learned from other Industries," in *AMCIS*, 2011 .

[٩٠] G. van Capelleveen, M. Poel, R. M. Mueller, D. Thornton, and J. van Hillegersberg" ,Outlier detection in healthcare fraud: A case study in the Medicaid dental domain," *International journal of accounting information systems,* vol. 21, pp. 18-31, 2016.

[٩١] J. Seo and O. Mendelevitch, "Identifying frauds and anomalies in Medicare-B dataset," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017: IEEE, pp. 3664-3667 .

[٩٢] J. WU, R. ZHANG, X. SHANG, and F. CHU, "Medical insurance fraud recognition based on improved outlier detection algorithm," *DEStech Transactions on Computer Science and Engineering,* no. aiea, 2017.

[٩٣] A. Gangopadhyay and S. Chen, "Health care fraud detection with community detection algorithms," in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2016: IEEE, pp. 1-5 .

[٩٤] H. Joudaki *et al.*, "Improving fraud and abuse detection in general physician claims: a data mining study," *International journal of health policy and management,* vol. 5, no. 3, p. 165, 2016.

[٩٥] P. Ferreira, R. Alves ,O. Belo, and L. Cortesão, "Establishing fraud detection patterns based on signatures," in *Industrial Conference on Data Mining*, 2006: Springer, pp. 526-538 .

[٩٦] C. Sun, Q. Li, L. Cui, Z. Yan, H. Li, and W. Wei, "An effective hybrid fraud detection method," in *International Conference on Knowledge Science, Engineering and Management*, 2015: Springer, pp. 563-574 .

[٩٧] I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Applied Soft Computing,* vol. 36, pp. 283-299, 2015.

[٩٨] K. D. Aral, H. A. Güvenir, İ. Sabuncuoğlu, and A. R. Akar, "A prescription fraud detection model," *Computer methods and programs in biomedicine,* vol. 106, no. 1, pp. 37-46, 2012.

[٩٩] Jeni, László A., Jeffrey F. Cohn, and Fernando De La Torre. "Facing imbalanced data--recommendations for the use of performance metrics ".2013 Humaine ,association conference on affective computing and intelligent interaction. IEEE 2013.

واژه‌نامه‌ی فارسی به انگلیسی

specificity	اختصاصی
Facility	امکانات
National HealthCare Anti-Fraud Association	انجمن ملی بهداشت و درمان ضد تقلب
Adaptive	انطباقی
Actor	بازیگر
betweenness	بینیت
Compact Matrix Decomposition (CMD)	تجزیه ماتریس فشرده
peak analysis	تجزیه و تحلیل اوج
data aggregation	تجمع داده
Peer Group Analysis	تحلیل گروه همتا
Service Provider’s Fraud	تقلب ارائه دهنده خدمات
variety	تنوع
Skewed class distribution	توزیع کلاس کج
conspiracy	توطیه
conspiratorial	توطیه آمیز
Early stopping	توقف زودهنگام
K – Nearest Neighbor	چند نزدیک‌ترین همسایگی
breadth-first search	جستجو عرض اول
volume	جلد

Maximum Common Sub graph	حداکثر زیر نمودار مشترک
Maximum Description Length	حداکثر طول توضیحات
Sensitivity	حساسیت
Error correcting graph matching distance	خطا در تصحیح فاصله تطبیق نمودار
Bayesian Bernoulli co-clustering	خوشه بندی مشترک بیزین برنولی
University of California, San Diego	دانشگاه کالیفرنیا ، سن دیگو
Office of inspector general	دفتر بازرسی کل
Drift	رانش
time evolving	زمان در حال تکامل
velocity	سرعت
Health Common Procedure Coding System	سیستم کدگذاری رویه مشترک بهداشتی
National Provider Identifier	شناسه ارائه دهنده ملی
concentration function	عملکرد غلظت
Non-facility	غیر تسهیلات
Confidence Interval	فاصله اطمینان
width binning interval	فاصله باند عرض
Graph Edit Distance	فاصله ویرایش نمودار
boxplots	قطعه جعبه
cosine	کسینوس
Consusion Matrix	ماتریس در هم ریختگی
distance of adjacency matrices	ماتریس مجاورت فاصله ها

Categorical variable	متغیرهای دسته ای
potential overestimation issues	مسائل بالقوه برآورد
eigenvalue	مقدار خاص
false discovery rate	میزان کشف کاذب
closeness	نزدیکی
Overfitting	نصب بیش از حد
static attributed graph	نمودار منسجم استاتیک
static plain graphs	نمودارهای ثابت استاتیک
Pattern Learning and Anomaly Detection on Streams	یادگیری الگو و تشخیص ناهنجاری در جریان ها

واژه نامه انگلیسی به فارسی

Actor	بازیگر
Adaptive	انطباقی
Bayesian Bernoulli co-clustering	خوشه بندی مشترک بیزین برنولی
betweenness	بینیت
boxplots	قطعه جعبه
breadth-first search	جستجو عرض اول
Categorical variable	متغیرهای دسته ای
closeness	نزدیکی
Compact Matrix Decomposition	تجزیه ماتریس فشرده
concentration function	عملکرد غلطت
Confidence Interval	فاصله اطمینان
conspiracy	توطیه
conspiratorial	توطیه آمیز
Confusion Matrix	ماتریس در هم ریختگی
cosine	کسینوس
data aggregation	تجمع داده
distance of adjacency matrices	ماتریس مجاورت فاصله ها
Drift	رانش
Early stopping	توقف زودهنگام
eigenvalue	مقدار خاص

Error correcting graph matching distance	خطا در تصحیح فاصله تطبیق نمودار
Facility	امکانات
false discovery rate	میزان کشف کاذب
Graph Edit Distance	فاصله ویرایش نمودار
Health Common Procedure Coding System	سیستم کدگذاری رویه مشترک بهداشتی
K – Nearest Neighbor	چند نزدیک‌ترین همسایگی
Maximum Common Sub graph	حداکثر زیر نمودار مشترک
Maximum Description Length	حداکثر طول توضیحات
National HealthCare Anti-Fraud Association	انجمن ملی بهداشت و درمان ضد تقلب
National Provider Identifier	شناسه ارائه دهنده ملی
Non-facility	غیر تسهیلات
Office of inspector general	دفتر بازرسی کل
Overfitting	نصب بیش از حد
Pattern Learning and Anomaly Detection on Streams	یادگیری الگو و تشخیص ناهنجاری در جریان ها
peak analysis	تجزیه و تحلیل اوج
Peer Group Analysis	تحلیل گروه همتا
potential overestimation issues	مسائل بالقوه برآورد
Sensitivity	حساسیت
Service Provider’s Fraud	تقلب ارائه دهنده خدمات

Skewed class distribution	توزیع کلاس کج
specificity	اختصاصی
static attributed graph	نمودار منسجم استاتیک
static plain graphs	نمودارهای ثابت استاتیک
time evolving	زمان در حال تکامل
University of California, San Diego	دانشگاه کالیفرنیا ، سن دیگو
variety	تنوع
velocity	سرعت
volume	جلد
width binning interval	فاصله باند عرض

Abstract

Health insurance is an urgent problem and increases the costs of health insurance programs significantly; HCF is a multibillion-dollar scam. The vastness of the field of health and a large amount of money has made this field a target for fraud. Thus, the field of health has become a significant source of costs in many countries. One of the sources of high costs of the Health and Medical Organization is the payment of insurance share of prescribed drugs for covered patients. In general, the purpose of detecting fraud is to maximize the right predictions and keep the wrong predictions at an acceptable level of cost. Despite the constant changes in fraudsters' behavior, models based on the analysis of past data may not be able to detect new forms of fraud. Also, none of the fraud detection systems alone can detect and cover all forms of fraud. In this dissertation, a new approach to estimating the possibility of fraud in medical records by the graph analysis method is investigated. A set of algorithms calculates behavioral similarities between the two categories of fraudulent and non-fraudulent health care providers according to measurable criteria for health care activities such as medical procedures and prescribing medications. Another set of algorithms estimates the extent to which healthcare providers are at risk of fraud through geographic location, duplication of shared locations, or other addresses. These algorithms have been evaluated for their ability to predict a provider's presence on the Inspector General's list of providers (disqualification from participating in geriatric health insurance and other federal health care programs).

Keywords— Drug prescription, abnormality detection, health care systems, graph analysis

Al-Zahra University - Faculty of Engineering

Thesis for master's degree

Computer Engineering - Artificial Intelligence

Title

Detection of fraud in health care systems with graph analysis approach

Supervisor

Dr.Mohammad Reza Keyvan pour

Student

Roonak Namaki

Summar of 2020