

فهرست مطالب

Contents

۱	مقدمه و کلیات	۲
۱,۱	مقدمه	۲
۱,۲	تعریف مساله	۳
۱,۳	اهداف پژوهش	۴
۱,۴	تعریف تقلب	۵
۱,۵	انواع تقلب	۵
1.5.1	Upcoding	۶
۱,۵,۲	دستکاری صورتحسابها	۷
۱,۵,۳	خدمات غیر ضروری	۷
۱,۵,۴	تشخیص اشتباه	۷
۱,۶	بازیگران نظام سلامت	۸
۱,۶,۱	تقلب ارائه دهنده خدمات	۸
۱,۶,۲	تقلب مشترکان بیمه	۸
۱,۷	ساختار پژوهش	۱۰
۱,۸	جمع بندی	۱۰
۲	مفاهیم و پیشینه پژوهش	۱۲
2.1	مقدمه ۱۲	
۲,۲	چالش های کشف تقلب	۱۲
2.2.1	پدیده Drift	۱۴
۲,۲,۲	توزیع اریب کلاس ها	۱۴
۲,۲,۳	خطای پیش بینی و اثر "Black Swan"	۱۴
۲,۲,۴	از دست دادن اطلاعات در پیش پردازش	۱۵
۲,۲,۵	خودکارسازی فرآیند داده کاوی برای کاربران غیر متخصص	۱۵
۲,۲,۶	ماهیت بین رشته ای تحقیق و دانش متخصص حوزه	۱۵
۲,۳	انواع ناهنجاری در تشخیص تقلب	۱۶
۲,۳,۱	تشخیص ناهنجاری گراف ایستا	۱۶
۲,۳,۲	تشخیص ناهنجاری مبتنی بر ساختار	۱۶
۲,۳,۳	تشخیص ناهنجاری مبتنی بر اجتماع	۱۷
۲,۳,۴	تشخیص ناهنجاری گراف پویا	۱۷
۲,۳,۵	تشخیص ناهنجاری مبتنی بر فاصله	۱۸
۲,۳,۶	تشخیص ناهنجاری مبتنی بر فشردم سازی	۱۸

تشخیص ناهنجاری مبتنی بر تجزیه	2.3.7	۱۸
تشخیص ناهنجاری مبتنی بر خوشه یا اجتماع	۲,۳,۸	۱۸
تشخیص ناهنجاری مبتنی بر مدل‌های احتمالاتی	۲,۳,۹	۱۸
تشخیص ناهنجاری مبتنی بر پنجره	۲,۳,۱۰	۱۹
BigData در کشف تقلب	۲,۴	۱۹
معیارهای تحلیل شبکه پزشکان	۲,۵	۲۱
رویکردهای کلی کشف تقلب	۲,۶	۲۱
الگوریتم‌های خوشه‌بندی	۲,۶,۱	۲۱
Apriori الگوریتم	۲,۶,۲	۲۲
روش‌های کلاس بندی	۲,۶,۳	۲۴
روش‌های یادگیری ماشین ترکیبی	۲,۶,۴	۲۶
جمع بندی فصل	۲,۷	۲۹
روش پیشنهادی تشخیص تقلب	۳	۳۲
مقدمه	۳,۱	۳۲
مجموعه داده استفاده شده	۳,۲	۳۲
مجموعه داده LEIE	۳,۲,۱	۳۲
مجموعه داده Medicare Provider Utilization and Payment	۳,۲,۲	۳۴
ویژگی‌های مورد استفاده	۳,۲,۳	۳۴
آماده‌سازی داده	۳,۳	۳۶
نیازمندی‌های روش پیشنهادی	۳,۴	۳۶
معیار شباهت کسینوسی	۳,۴,۱	۳۷
توصیف روش پیشنهادی	۳,۵	۳۸
جمع‌بندی فصل	3.6	۴۰
ارزیابی روش پیشنهادی و گزارش نتایج الگوریتم	۴	۴۲
مقدمه	۴,۱	۴۲
معیارهای ارزیابی و روال اعتبار سنجی متقابل	۴,۲	۴۲
نتایج مقدارهای TP, TN, FP, FN	۴,۲,۱	۴۳
حساسیت	4.2.2	۴۳
تشخیص‌پذیری	4.2.3	۴۴
منحنی AUC	4.2.4	۴۴
نتایج حاصل از اجرا روی ماتریس درهم ریختگی	۴,۲,۵	۴۴
مقایسه با سایر الگوریتم‌های موجود	۴,۳	۴۵
بررسی بازه اطمینان نتایج الگوریتم	۴,۴	۴۶
جمع بندی فصل	۴,۵	۴۶

۴۸	نتایج و توسعه‌های آتی	۵
۴۸	مقدمه	۵,۱
۴۹	نتیجه‌گیری	۵,۲
۵۰	توسعه‌های آتی	۵,۳
۵۲	مراجع	

فهرست شکل‌ها

شکل ۱-۱ نمودار کلی گردش اطلاعات در سامانه	۳
شکل ۲-۱ درخت دسته‌بندی بازیگران نظام سلامت	۹
شکل ۱-۲ توزیع مقالات FDS بر اساس مسائل و چالش‌های بین سال‌های ۱۹۹۴ تا ۲۰۱۴	۱۳
شکل ۱-۳ معماری کلی سیستم پیشنهادی	۳۷
شکل ۲-۳ فرمول محاسبه معیار شباهت کسینوسی	۳۷
شکل ۳-۳ معماری جزئیات سیستم پیشنهادی	۳۹
شکل ۴-۳ فرمول محاسبه log ratio	۳۹
شکل ۱-۴ نتایج معیارهای پایه برای الگوریتم	۴۳
شکل ۲-۴ فرمول sensitivity	۴۳
شکل ۳-۴ فرمول specificity	۴۴
شکل ۴-۴ ماتریس در هم‌ریختگی برای نتایج روش پیشنهادی	۴۵

فهرست جدول‌ها

۶	جدول ۱-۲ انواع تقلب در بیمه سلامت
۲۴	جدول ۲ مثالی از میانگین گروهی برای PGA [۸۸]
۲۶	جدول ۳ انواع رویکردها و روش‌های موجود در کشف تقلب سیستم سلامت
۳۳	جدول ۱-۳ قانون‌های مربوط به مجموعه داده LEIE
۳۵	جدول ۲-۳ ستونهای مجموعه داده
۴۵	جدول ۱-۴ مقایسه نتایج مدل با سایر الگوریتم‌ها
۴۶	جدول ۲-۴ بازه اطمینان برای نتایج الگوریتم
۴۹	جدول ۱-۵ دستهبندی به تفکیک رویکردهای کلی کشف تقلب

فصل ۱

مقدمه

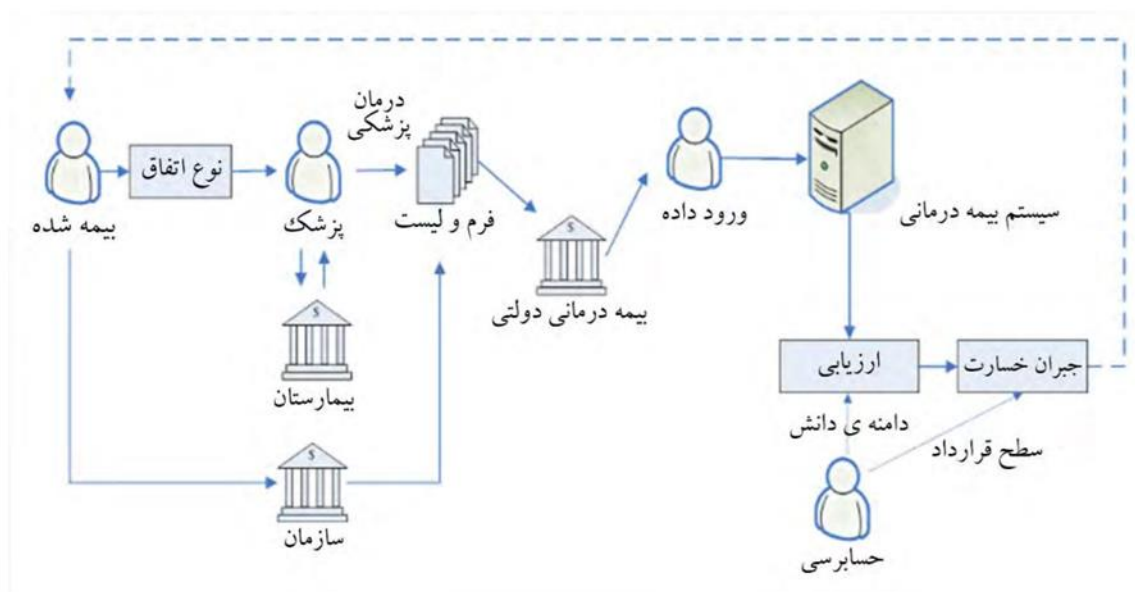
۱. مقدمه و کلیات

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 سیستم های پزشکی، اهمیت بررسی آنها و نکات مهم در مورد این مساله و همچنین روش های 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 چالش های 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 هزینه های درمانی به علت بیماری یا آسیب دیدگی تعریف می کند. این پوشش می تواند توسط بعضی سازمان های مرکزی، برای مثال شرکت های خصوصی یا دولتی، ارائه شود. منبع این پوشش در بسیاری از کشورها صرف نظر از سیستم های بهداشت و درمانشان، متفاوت است. بررسی سالیانه انجام شده توسط صندوق مشترک المنافع، سیستم های بهداشت و درمان استرالیا، نیوزیلند، بریتانیا، آلمان، کانادا و ایالات متحده را مقایسه می کند. این بررسی تاکید می کند که ایالات متحده تنها کشور بدون پوشش بیمه سلامت سراسری است. اداره آمار ایالات متحده بیان می کند که ۳۱ درصد از آمریکایی ها طرح بیمه سلامت عمومی دارند، در حالی که ۵۵ درصد از آنها پوشش خود را از طریق کارفرمایانشان می گیرند. اگرچه، تحت پوشش بودن تضمین نمی کند که شخص بیمه شده هیچ هزینه پزشکی پر داخت نکند. میزانی که بیمه شده باید بپردازد، قبل از اینکه بیمه گر برای یک ویزیت یا خدمت خاص بپردازد، پرداخت مشترک نامیده می شود. جدای از پرداخت مشترک، ممکن است خدماتی باشد که بیمه گر بر اساس حق بیمه ای که بیمه شونده می خرد، بازپرداخت می کند. مانند خدماتی که به عنوان بیمه تکمیلی شناخته می شوند که در آن درصد بیشتری از هزینه ها در قبال دریافت حق بیمه بیشتر پرداخت می شود. روند گردش اسناد در سیستم بیمه در شکل (۱) نشان داده شده است [4].



شکل ۱-۱ نمودار کلی گردش اطلاعات در سامانه

تقریباً 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 دستورالعمل‌هایی 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.

۱,۲ تعریف مساله

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. متأسفانه، 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 آنها را که به آن رسیدگی به اسناد می‌گویند را آغاز مینمایند. با توجه به حجم اسناد قابل رسیدگی و کند بودن روش‌های 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 کوتاهی 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 انبار 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.

۱،۳ اهداف پژوهش

کلاهبرداری در حوزه 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 هزینه‌های مراقبت‌های 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 تقریباً 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 برنامه‌های مراقبت‌های 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.

۱.۴ تعریف 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 لزوماً 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.

۱.۵ انواع تقلب

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، بیمه گر از ارائه‌کننده خدمات سلامت یا بیمه شده می‌خواهد 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 ۱-۱ انواع تقلب در بیمه سلامت

ردیف	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

۱.۵.۱ Upcoding

کدگذاری اشتباه فعالیت‌ها، می‌تواند سرخ‌هایی 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 بهداشتی 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 کدینگ 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 جزئی تر به چندین کد جزئی تر، 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 فعالیت‌های 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.

۱.۵.۲ دستکاری صورتحساب‌ها

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

۱.۵.۳ خدمات غیرضروری

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.

۱.۵.۴ تشخیص اشتباه

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 می‌توانسته 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 پزشکی 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 آمریکا 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 خلاصه، 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.

۱,۶ بازیگران نظام سلامت

ارتباط میان بازیگران^۱ مختلف سیستم سلامت به واضح‌ترین شکل ممکن در شکل ۱-۲ 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].

کلاهدرداری مراقبت‌های بهداشتی موضوعی خاص برای هر کشور است و رفتارهای کلاهدردانه، متفاوتی به همین نسبت تغییر می‌کند. با این وجود، انواع کلاهدرداری‌هایی که در زمینه مراقبت‌های بهداشتی انجام می‌شود، تقریباً برای همه کشورها رخ می‌دهد. همانطور که در شکل ۱-۲ 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 می‌شود)، و شرکت‌های بیمه‌ای که ادارات درمان و بهداشت دولتی و شرکت‌های بیمه خصوصی را در بر می‌گیرند. براساس اینکه کدام طرف مرتکب تقلب می‌شود، رفتارهای تقلب‌آمیز در ادامه دسته‌بندی و توضیح داده می‌شود. بر اساس اینکه چه کسی مرتکب تقلب می‌شود، رفتارهای تقلب‌آمیز به صورتی که در ادامه آمده‌اند دسته‌بندی می‌شوند [۳]:

۱,۶,۱ تقلب ارائه‌دهنده‌ی خدمات^۲

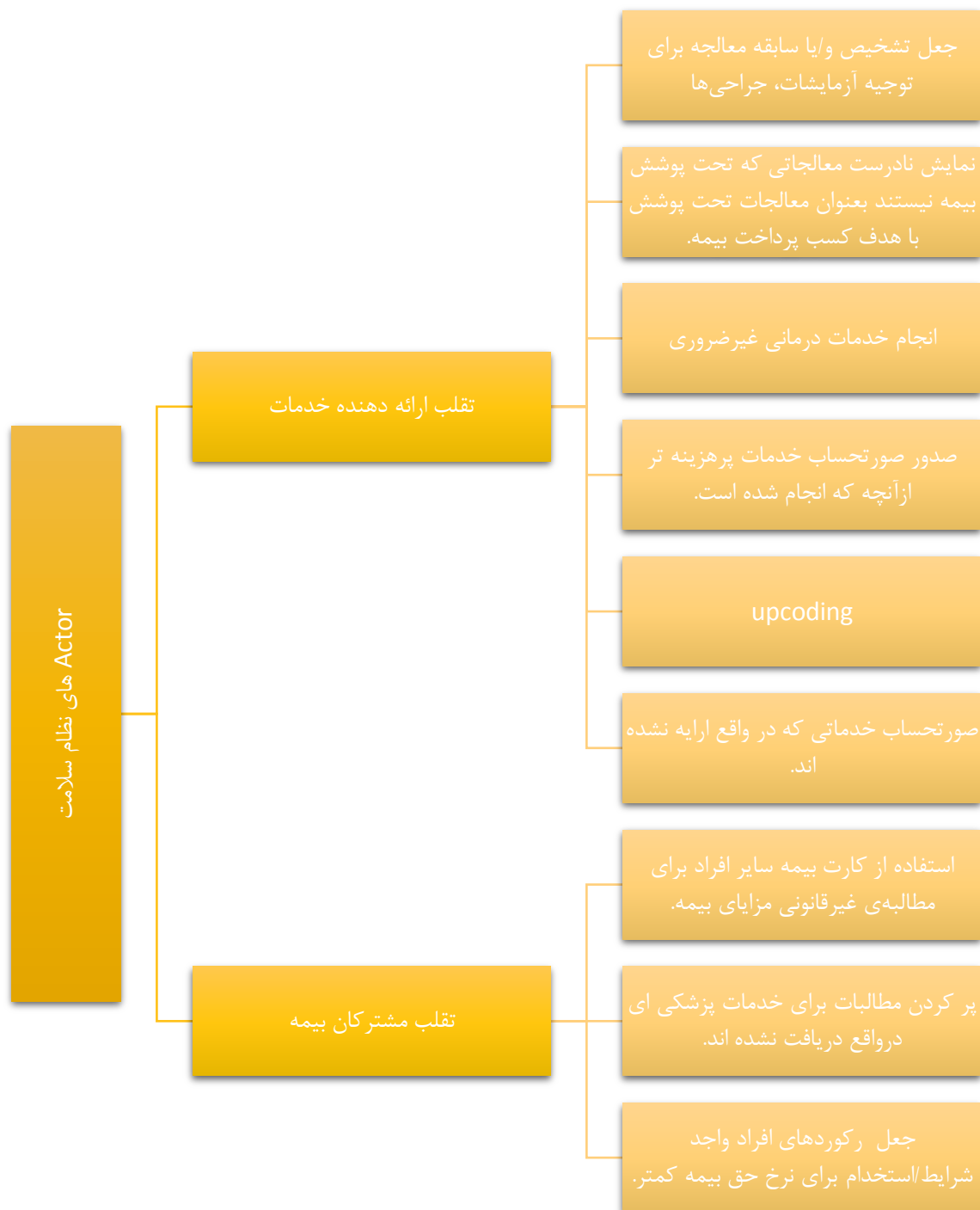
شامل تخلفات ارائه‌دهندگان خدمات می‌شود که می‌تواند به عنوان مثال، جعل تشخیص و یا سابقه معالجه برای توجیه آزمایشات، جراحی‌ها یا سایر فرآیندهایی که از نظر پزشکی غیرضروری هستند، باشد. با توجه به انواع تقلب، اکثر مطالعات تا کنون برای کشف تقلب ارائه‌دهندگان خدمات مورد استفاده قرار گرفته است. از آنجا که کشف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

۱,۶,۲ 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.

^۱ Actor

^۲ Service Provider's Fraud



شکل ۲-۱ درخت دسته‌بندی بازیگران نظام سلامت

۱,۷ ساختار پژوهش

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 می‌شود.

در فصل سوم مروری به تقلب در این سیستم ها داشته و مطالب مرتبط با آن، عواقب این موضوع، راه حل های پیشین برای حل این مساله و چالش های پیش رو مطرح می‌گردد.

در فصل چهارم این پایان نامه به بررسی روشی پرداخته خواهد شد که به عنوان راه حلی برای پیش بینی این تقلب ها یا ناهماهنگی ها در سیستم ارایه گردیده است. این روش مبتنی بر نظریه گراف است و سعی دارد علاوه بر حل چالش های موجود، عملکرد روش های 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. بخش های مختلف مدل مطرح شده و شرح آن ها نیز بیان می‌شود.

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

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

۱,۸ جمع بندی

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

فصل ۲

پیشینه پژوهش

۲. مفاهیم و پیشینه پژوهش

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 انواع مختلف تقلب را شناسایی کنند، می‌پردازیم.

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 به یک هدف جذاب برای کلاهبرداری تبدیل شود. شرکت‌های بیمه روش‌هایی را برای تشخیص تقلب ایجاد می‌کنند که عمدتاً برگرفته از تجارب خبرگان بوده و کمتر به روش‌های مبتنی بر تحلیل داده متکی است. صنعت بیمه و در راس آن بیمه سلامت با هزینه‌ای 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 به صورت دستی و توسط افراد خبره صورت می‌گیرد. استفاده از روش‌های تحلیل داده‌های بزرگ نظیر داده کاوی به دینفعان کمک می‌کند تا بتوانند ضمن تعمیم و بهره‌برداری از الگوهای شناخته شده جهت بکارگیری الگوریتم‌های با ناظر به کشف الگوهای ناشناخته از طریق به کارگیری الگوریتم‌ها و مدل‌های نظارت نشده، پردازند.

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

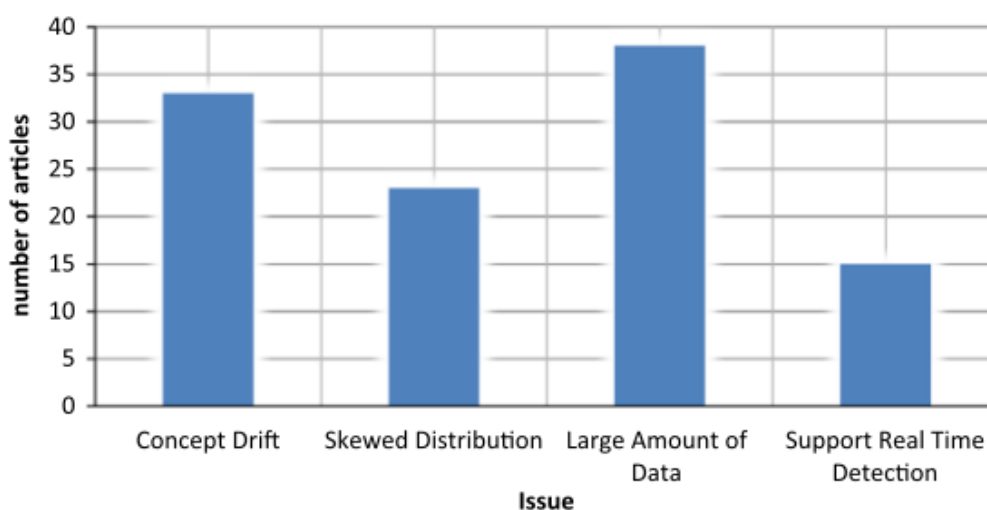
۲,۲ چالش‌های کشف تقلب

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 تفصیل پرداخته می‌شود.

شرکت های بیمه می توانند با آگاهی از انواع تقلبات و فرآیندهایی که احتمال بروز تقلب در آن ها وجود دارد سیستم هشدار دهنده و پیشگیرانه ای را طراحی کنند و با آگاهی از میزان آسیب پذیری خود استراتژی های موثرتری را به کار گیرند، اما برای تحقق این امور شرکت های بیمه با محدودیت ها و پیچیدگی های زیادی مواجه اند [۱۸]:

- پنهان بودن ماهیت تقلب
- پویایی و حساسیت به تغییر در تقلب (هنگام شناسایی یک سبک تقلب، کلاهبرداری با سبک دیگری در حال شکل گیری است)
- عدم توافق اجماع در برخی موارد بر آنچه که واقعاً به منزله ی تقلب در بیمه استاندارد.
- نگاه جامعه به شرکت های بیمه ای و ارائه دهندگان خدمات درمانی بعنوان نهادهای حمایتی
- ضعف سیستم های کنترلی مبتنی بر فناوری اطلاعات
- از چالش های موجود در کشف تقلب عدم وجود داده های با برچسب سالم و وجود تعداد بسیار کم داده های تقلبی جهت یادگیری است. بصورت طبیعی تعداد تقلب کم است ولی تعداد دفعاتی که ثبت شده و کدگذاری شده است کم تر است. لذا داده هایی آماده نشده با تعداد کمی برچسب و تعداد زیادی نمونه های نامشخص وجود دارد.

شکل (۳) توزیع مطالعات سیستم های تشخیص تقلب را بر اساس چالش ها بر اساس تعداد مقالات منتشر شده در سال های ۱۹۹۴ تا ۲۰۱۴ نشان می دهد و به رایج ترین انواع تقلب های الکترونیکی مانند کارت های اعتباری، بیمه مراقبت سلامت، مخابرات، بیمه اتومبیل متمرکز است [۱۹].



شکل ۱-۲ توزیع مقالات FDS براساس مسائل و چالش های بین سال های ۱۹۹۴ تا ۲۰۱۴

تعاریف مختلفی برای مفهوم مسئله رانش وجود دارد. در داده‌کاوی رانش به پدیده‌ای که مدل پایه‌ی آن در طول زمان در حال 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 متغیر هدف در طول زمان تغییر می‌کند، ارجاع داده می‌شد. اگرچه در یادگیری با ناظر هدف پیش‌بینی یک متغیر هدف y با استفاده از مجموعه‌ای از ویژگی‌های ورودی x است. در نمونه یادگیری که برای ساخت مدل استفاده می‌شود هر دوی x و y در زمان پیشگویی ناشناخته‌اند و رابطه بین داده ورودی و متغیر 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 الگوریتم‌های یادگیری تطبیقی^۳ برای مقابله با مفهوم رانش لازم است. الگوریتم‌های یادگیری تطبیقی می‌توانند بعنوان الگوریتم‌های یادگیری افزایشی توسعه یافته دیده شوند که قادر به بروزرسانی مدل تشخیص برای داده‌ی در جریان تکامل، در طول زمان می‌باشند [۲۲]، [۲۳].

۲,۲,۲ توزیع اریب کلاس‌ها^۴

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

۲,۲,۳ 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 پیش‌بینی بالا برای سرطان یا هر بیماری دیگر، برنامه صحیحی را برای تصمیم‌گیری فراهم نمی‌کند.

^۳ Adaptive

^۴ Skewed class distribution

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 پیش بینی "The Black Swan" توسعه دهند [۳۷].

یک مطالعه [۳۸] یک مسئله مشابه را در توصیه های 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. پژوهش های آینده باید به چالش های پیاده سازی مدل های پیش بینی کننده بپردازد، به ویژه اینکه چگونه فرآیند تصمیم گیری باید در صورت اشتباهات و حوادث غیرقابل پیش بینی سازگار شود.

۲,۲,۴ از دست دادن اطلاعات در پیش پردازش

پیش پردازش داده ها، از جمله دستکاری داده های از دست رفته، پرهزینه ترین و مهم ترین 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 داده های پرت، ما مقدار قابل توجهی از اطلاعات را از دست می دهیم. پژوهش های آینده باید بر روی یافتن یک روش بهتر تخمین مقادیر از دست رفته نسبت به حذف تمرکز نمایند. علاوه بر این، تکنیک های جمع آوری داده ها باید به منظور جلوگیری از این موضوع توسعه یابند یا اصلاح شوند.

۲,۲,۵ خودکارسازی فرآیند داده کاوی برای کاربران غیرمتخصص

کاربران نهایی داده کاوی در مراقبت های بهداشتی، پزشکان، پرستاران و متخصصین مراقبت های بهداشتی هستند که آموزش های 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 از خطاهای پزشکی نیز می تواند توسعه یابد [۴۲]؛ اما این کار چالش برانگیز خواهد بود زیرا در آن زمینه های کاربردی مختلف وجود دارد و یک الگوریتم دقت مشابهی برای تمام برنامه های کاربردی ندارد [۴۱].

۲,۲,۶ ماهیت بین رشته ای تحقیق و دانش متخصص حوزه

تجزیه و تحلیل بهداشت و درمان یک زمینه تحقیقاتی بین رشته ای است [۴۱]. به عنوان یک روش تجزیه و تحلیل، داده کاوی 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]. تقریباً 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].

۲,۳ 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 میدانند که تا حدی زیادی از دیگر مشاهدات متفاوت است [۴۶]. منشأ ناهنجاری‌ها میتواند رفتار کلاهبردانه، 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].

۲,۳,۱ 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.

۲,۳,۲ تشخیص ناهنجاری مبتنی بر ساختار

دو نوع روش تشخیص ناهنجاری مبتنی بر ساختار به نام‌های ناهنجاری در گراف‌های ساده ایستا ۵ و ناهنجاری در گراف ویژگی ایستا ۶ وجود دارد که بصورت زیر تشریح می‌شوند:

در طرح تشخیص ناهنجاری در گراف ساده 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 و community cohesiveness برای تمایز رفتارهای آنلاین افراد توسط الگوهای مصرف آن‌ها استفاده نمودند. به علاوه 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

static plain graphs^۵

static attributed graph^۶

node (ego) و one-level neighborhood (an egonet) و two-level neighborhood (a super egonet) و average betweenness user's egonet و betweenness centrality 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^y) پیدا می‌کند و زیرساخت‌های 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 کشف بهترین زیرساختار در گراف استفاده می‌کند اما به جای 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 به عنوان یک ابزار برای پشتیبانی تحقیقات جرم استفاده نموده است.

۲,۳,۳ تشخیص ناهنجاری مبتنی بر اجتماع

در [۵۲] یک روش تشخیص ناهنجاری مبتنی بر اجتماع ارائه دادند، با شناسایی اجتماعاتی که برای مرزهای اجتماعی اهمیتی قابل نیستند. این کار بر اساس یک نظریه تعلق گره‌های دارای سورتار 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 به عنوان 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.

۲,۳,۴ 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 فشردگی، پویای مبتنی بر برانگیز، پویای مبتنی بر خوشه یا اجتماع، پویای مبتنی بر مدل‌های احتمالاتی و پویای مبتنی بر پنجره تقسیم می‌شوند.

۲,۳,۵ تشخیص ناهنجاری مبتنی بر فاصله

معیار مبتنی بر فاصله می‌تواند برای اندازه‌گیری تغییر بین دو شی به کار 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 ۸، بیشینه زیرگراف مشترک ۹، فاصله ماتریس همسایگی ۱۰، فاصله ویرایش گراف ۱۱، فاصله همینگ برای ماتریس‌های همسایگی گراف‌ها و ... [۵۴].

۲,۳,۶ تشخیص ناهنجاری مبتنی بر فشردسازی

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

۲,۳,۷ تشخیص ناهنجاری مبتنی بر تجزیه

این روش ناهنجاری‌های موقتی را با نمایش مجموعه‌ای از گراف‌های تکامل زمانی ۱۲ بعنوان یک تنسور یا آرایه‌ی چند بعدی تشخیص می‌دهد و factorization یا کاهش بعد انجام می‌دهد. یک روش جدید تجزیه ماتریس فشرده ۱۳ 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.

۲,۳,۸ 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 غیرعادی‌ای 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.

۲,۳,۹ 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

Error correcting graph matching distance ۸

Maximum Common Sub graph (MCS) ۹

distance of adjacency matrices ۱۰

Graph Edit Distance (GED) ۱۱

time evolving ۱۲

Compact Matrix Decomposition (CMD) ۱۳

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.

۲,۳,۱۰ تشخیص ناهنجاری مبتنی بر پنجره

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 جریان‌ها ۱۴ که 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.

۲,۴ BigData در کشف تقلب

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 پرونده‌های 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 شده 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.

از نظر اقتصادی تقلب در بیمه به یک مساله جدی در حال افزایش تبدیل شده است. براساس گزارش خبر BBC در سال ۲۰۰۷ مطالبه‌های بیمه تقلبی 1.6 بلیون 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 به کار می‌رود. داده‌کاوی درباره‌ی پیدا کردن روش‌هایی است که قابل اعتمادند، قبلاً ناشناخته بودند و از داده‌ها قابل اجرا هستند. این داده باید در دسترس و مرتبط و کافی و تمیز باشد. همچنین مسئله

داده کاوی باید به خوبی قابل تعریف باشد و با ابزارهای پرس و جو قابل حل نباشد و با یک مدل پردازش داده کاوی تعلیم داده شود [۱۳]. داده کاوی را به این صورت تعریف می کنند: پروسه شناسایی الگوهای مورد علاقه در پایگاه داده بطوریکه بعداً بتوانند در تصمیم گیری استفاده شوند [۱۴].

[۱۴] داده کاوی را پروسه ای تعریف می کند که از ریاضیات آماری و هوش مصنوعی و تکنیک های یادگیری ماشین برای استخراج و شناسایی اطلاعات مفید استفاده می کند و بطور پیوسته از یک پایگاه داده بزرگ تجربه کسب می کند. [۱۵] بیان می کند که هدف داده کاوی به دست آوردن اطلاعات مفید و غیر آشکار از داده های ذخیره شده در مخازن بزرگ است. [۱۶] مشخص می کند که یکی از مزایای مهم داده کاوی این است که می تواند برای توسعه یک کلاس جدید از مدل ها برای شناسایی جملات، قبل از اینکه توسط متخصصان شان تشخیص داده شوند، استفاده شود. [۱۱] اشاره دارد که تشخیص قلب یکی از بهترین کاربردهای داده کاوی در صنعت و دولت است. تکنیک های داده کاوی گوناگونی در تشخیص قلب بیمه به کار گرفته شده اند مانند شبکه های عصبی، مدل های رگرسیون منطقی، روش های نایوبیز و درخت تصمیم.

در داده های کلان، فقط موضوع حجم^{۱۵} مطرح نیست و باید سایر موارد از قبیل تنوع^{۱۶} داده ها و سرعت^{۱۷}، برقرار باشد تا بتوان به دنیای یادگیری عمیق وارد شد. در موضوع کشف قلب حوزه درمان نه تنها با حجم انبوهی از داده های متنوع مواجه ایم، بلکه این داده ها و الگوهای مربوط روزه روز در حال تغییرند.

در سال های اخیر علاقه رو به افزایش در کاوش داده های مراقبت سلامت برای تشخیص قلب شکل گرفته است. سیستم ها برای پردازش مطالبات الکترونیک پیاده سازی شده اند تا بصورت اتوماتیک بازرسی و مرور از داده های مطالبات را انجام دهند. این سیستم ها برای تشخیص اعمال قلبی، صورتحساب اشتباه، مطالبات تکراری و سرویس هایی که تحت پوشش درمانی نیستند، طراحی شده اند. قابلیت های تشخیص قلب این سیستم ها معمولاً محدود است زیرا تشخیص بطور عمده مبتنی بر قوانین ساده از پیش تعریف شده توسط متخصصان امر است. برای رسیدن به تشخیص موثرتر، بسیاری از محققین روش های پیچیده تر مقابله با قلب را توسعه داده اند که بر اساس داده کاوی، یادگیری ماشین و سایر روش های تحلیلی است. روش های جدید ارائه شده برخی مزیت های اصلی مانند یادگیری خودکار الگوهای قلب از داده ها و تعیین احتمال قلب برای هر مورد و شناسایی انواع جدید قلب که قبلاً ثبت نشده اند را دارا هستند [۱۰]، [۲۵].

volume	۱۵
variety	۱۶
velocity	۱۷

۲.۵ معیارهای تحلیل شبکه 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:
- درجه: اهمیت پزشک متناسب با تعداد بیماران مشترک با پزشکان دیگر است.
- مقدار خاص ۱۸: هرچه تعداد بیماران مشترک با سایر پزشکان مهم بیش تر باشد، پزشک مهم تر در نظر گرفته می شود. اگر پزشک تعداد زیادی از بیماران را به اشتراک بگذارد، اما با پزشکانی که در شبکه مهم نیستند، پزشک مهم در نظر گرفته نمی شود [۶۰].
- بینابینی ۱۹: با توجه به اینکه پزشکان می توانند برای دیگر پزشکانی که نزدیک ترند تاثیر گذار باشند، مثلاً زمانی که پزشک یک پزشک دیگر را به بیمار خود نشان می دهد، تعریف معیاری که این مجاورت را نشان دهد امکان پذیر است [۶۱].
- نزدیکی ۲۰: پزشکان می توانند توسط مقدار سایر پزشکان که در میان یکدیگر شناخته شده اند، غیرمستقیم و بدون اطلاع آن ها، مرتبط باشند. بنابراین، با توجه به تعداد پزشکان مورد نیاز برای اتصال دو پزشک، می توان میزان جدایی پزشکان را محاسبه کرد [۶۲].

۲.۶ رویکردهای کلی کشف تقلب

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

۲.۶.۱ الگوریتم های خوشه بندی

خوشه بندی اولین بار روی داده های پزشکی برای بخش بندی درمان پزشکان عمومی توسط [۷۰] اعمال شد. در [۷۱] و [۲۵] از داده های جغرافیایی در یک رویکرد مبتنی بر خوشه بندی استفاده نمودند. الگوریتم گروه خوشه ای بیزین برنولی^{۲۱} [۷۲] با تمرکز بر وقوع ویزیت میان ارائه دهندگان و ذینفعان، داده های دوتایی را مدل می کند. این به طور بالقوه می تواند یک نوع تقلب در حال ظهور به نام "تقلب توطئه" را نشان دهد که شامل ویژگی های بیش از یک عضو از سیستم پزشکی است. این الگوریتم های خوشه بندی به بازرسان کمک می کند که صورتحساب و متغیر مطلوبشان را گروه بندی کنند. [۷۳]

eigenvalue ۱۸

betweenness ۱۹

closeness ۲۰

Bayesian Bernoulli co-clustering ۲۱

یک مرور کلی از روش های تشخیص داده پرتدر برخی از آزمایش ها برای ارزیابی اثربخشی آن ارائه می کند. این روش های تجزیه تحلیل شامل مدل های خطی، طرح جعبه ای^{۲۲}، تحلیل قله^{۲۳}، خوشه بندی چند متغیره و ارزیابی متخصص می باشد. [۷۴] یک روش تشخیص داده پرت مبتنی بر چگالی محلی برای شناسایی الگوهای پرداخت نامناسب در سیستم پزشکی استرالیا ارائه می دهد. [۷۵] یک رویکرد یکپارچه که ترکیبی از انتخاب ویژگی، خوشه بندی، تشخیص الگو و تشخیص بیرونی است برای شناسایی تقلب در سیستم پزشکی استرالیا ارائه نمودند. [۷۶] یک روش تشخیص ناهنجاری دو مرحله ای برای شناسایی بیمارستان های جعلی در سیستم مراقبت بهداشت عمومی برزیل ارائه می کند. همچنین شامل مطالعات تشخیص داده پرت با داده های تجویزی است. [۷۷] یک مدل رفتاری پایه نرمال را برای شناسایی ناهنجاری ها برای شناسایی ناهنجاری های مربوط به هر نسخه ایجاد می کند. [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 می دهد. [۷۹] استفاده از یک تابع غلظت^{۲۴} را به عنوان یک ابزار تشخیص پیش نمایش داده پرت برای کمک به ارزیابی تقلب پزشکی ارائه می دهد. علاوه بر این، ابزارهای صنعتی مبتنی بر تجزیه و تحلیل گراف، تجزیه و تحلیل ارتباطات و انجمن ها، 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 آن ها، مکان، ارائه دهندگان خدمات، دارایی ها و وابستگی ها 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 می تواند برای آشکارسازی شبکه های سازمان یافته، پیچیده و هماهنگ ارائه دهندگان و بیماران مفید باشد. رویکردهای بدون ناظر به طور کلی مورد استفاده قرار می گیرد تا قبل از اینکه متخصصان حوزه را به تحقیق بفرستند فعالیت های جعلی را به طور بالقوه برچسب بزنند. بنابراین، یک همکاری نزدیک بین پزشکان، آمارگیران و افرادی که در تصمیم گیری شرکت دارند، در مراحل تعیین و تنظیم مدل و تجزیه و تحلیل و تفسیر نتایج سودمند خواهد بود [۶۹].

۲.۶.۲ الگوریتم Apriori

الگوریتم Apriori یکی دیگر از تکنیک هایی است که در تشخیص تقلب استفاده می گردد. این الگوریتم (Agrawal و همکاران ۱۹۹۳)، مهم ترین الگوریتم کلاسیک برای کاوش اقلام مکرر است. Apriori برای یافتن همه اقلام مکرر در پایگاه داده شده DB استفاده می شود. بر اساس اصل Apriori هر زیر مجموعه ای 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^{۲۵} (جستجوی سطح هوشمندانه) شناخته می شود که در آن k-آیتم برای کشف (k + ۱) آیتم بکار می روند. در ابتدا، مجموعه اقلام ۱-تکراری یافت می شود که آستانه پشتیبانی را برآورده

۲۲ boxplots

۲۳ peak analysis

۲۴ concentration function

۲۵ breadth-first search

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^{۲۶} ارائه نمودند که براساس نظر متخصصین پزشکی انتخاب شده است. یک تحقیق دیگر [۸۴]، صورت حساب 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 صورت حساب 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^{۲۷} 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. به طور 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. به عنوان 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 به طور 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 به طور قابل ملاحظه ای بالاتر از سایر معالجات است و بنابراین محتمل است که دندانپزشک با قصد منحرف از این معالجه برای صورت حساب بیشتر استفاده کرده باشد.

۲۶ width binning interval

۲۷ 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

۲,۶,۳ 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. زمانی که 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 این گونه موارد زمانی که 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. Profiling و 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 می‌توان 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 به هم 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 مجموعه داده‌های یکپارچه است. همچنین ترکیب چنین سیستمی در مراجع بهداشتی، سازمان‌های تأمین اجتماعی و شرکت‌های بیمه می‌تواند کارایی را برای اطمینان از رعایت قانون بهبود بخشد، و هزینه‌های حساس‌تری 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 مجموعه داده 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.

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 می‌بخشد و SimLOF که LOF را بهبود می‌بخشد 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 می‌دهد که روش آن‌ها دقت بیشتری نسبت به روش‌های موجود در تشخیص تقلب بیمه‌های 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 سیستم 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 تکنیک‌های تشخیص شامل Outlier انحراف از مدل خطی، انحراف خوشه، انحراف از خوشه تکی، 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 نیاز به outlier همکاری متخصصان امر برای طراحی معیارها و به خصوص تفسیر نتایج دارد outlier تکنولوژی هنوز در مرحله آزمایشگاهی است و خود را در عمل و در	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	اجرای طولانی ثابت نکرده		گرایشی، حداکثر انحراف		
Medical procedure code national provider identifier specialty تعداد ها در procedure هر سال	وقوع برخی ها به دلیل FP اشتراک مشخصه‌های CPT کلی در (نسخه‌ی) پزشکان با تخصص‌های مختلف	که Page rank سابقاً در حوزه تحلیل شبکه به کار رفته بود در حوزه قلب در مطالبات بیمه نیز مؤثر عمل می‌کند	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	A Novel Page Rank-Based Algorithm to Identify Anomalies [۹۱]	۲
بازپرداخت مربوط به برونشیت مزمن بازپرداخت مربوط به بیماری‌های قلبی- ریوی بازپرداخت مربوط به ذات الریه	ویژگی‌های استفاده شده فقط مربوط به بازپرداخت بیمه سلامت هستند و شامل اطلاعات بیمار و نسخه نمی‌باشند	استفاده از مجموعه داده واقعی کاهش زمان اجرای الگوریتم با یافتن مقدار با پیچیدگی k بهینه زمانی از مرتبه $O(k^2 \cdot m \cdot \log(n))$ و پیچیدگی مکانی از مرتبه $O((n-o) \cdot k \cdot m)$ که نسبت به CBLOF الگوریتم بهبود داشت.	استفاده از یک الگوریتم Outlier تشخیص بهبود یافته بر اساس k -خوشه‌بندی means	Medical Insurance Fraud Recognition Based on Improved Outlier Detection Algorithm [۹۲]	۳
procedure code صورتحساب هر procedure code	استفاده از داده‌های ساختگی استفاده از برخی پکیج‌های نرم افزاری برای محاسبه ماتریس روابط، که برای مجموعه	تخصیص احتمال Likelihood / احتمال تشکیل شبکه انحصاری) به هر پزشک سرعت تشخیص بالا	ساخت ماتریس ارتباط میان دو پزشک و شناسایی پزشکانی که در یک شبکه بهم متصلند و با سایر پزشکان ارتباطی ندارند،	Community Detection Algorithm to Find Suspicious Group of Provider Community	۴

health claim data	داده‌های بزرگ بهینه نیست		بعنوان شبکه قلب آمیز، و همچنین استفاده از اپراتورهای DB به جای برای loop بیش از دو پزشک	[۹۳]	
تاریخ ارایه خدمات کد پزشک تعداد ویزیت‌های هر بیمار به ازای هر پزشک	عدم تفکیک پزشکان طبق تخصص آن‌ها post payment				
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	حذف داده‌های ناشناس عدم استفاده از روش‌های آماری برای پرکردن داده‌های از دست رفته	انجام تحقیق روی پزشکان هردو بخش عمومی و خصوصی استفاده از مجموعه داده واقعی	انجام عمل خوشه بندی بر اساس hierachical clustering method و محاسبه Euclidian distance measures با استفاده از شاخص اعتباری بیشینه مقدار ضریب همبستگی سیلویت	Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study [۹۴]	۵
بیمار ID نوع نسخه اطلاعات زمانی مرتبط به حادثه مشاغل درگیر در ID فرآیند هزینه نسخه تخصص پزشک	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	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	به کارگیری تکنیک‌های برای تحلیل مطالبات بیمه سلامت از طریق نگاشت پزشکان با استفاده از بیماران مشترک بعنوان یک پروکسی برای ارتباط میان آن‌ها	A Social Network Analysis Framework for Modeling Health Insurance Claims Data [۹۵]	۶
نسبت تعداد نسخه‌ها به تعداد مشخصی از بیمه شدگان	تغییر مداوم وزن‌ها توسط متخصصان	سازگار در یک محیط پویا	تولید سناریو توسط متخصصان و پزشکان برای رفتارهای	An Interactive Machine Learning Based Electronic	۷

نسبت تعداد نسخه‌ها به تعداد مشخصی از پزشکان	برای شناسایی انواع جدید تقلب	قابل استفاده برای proactive , reactive تحلیل	غیرطبیعی و سپس ها با actor وزن‌دهی روش وزن‌دهی binary pairwise comparison و محاسبه امتیاز خطای ها و مطالبات، actor استفاده از visualization بزار توسعه یافته تحت که برای QlikView تحلیل بکار می‌رود	Fraud and Abuse Detection System in HealthCare Insurance [۹۷]	
تعداد کل نسخه‌ها	تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب	کاهش زمان تحلیل نتایج خروجی توسط کاربران به دلیل استفاده از ابزار visualization استفاده از بیش از actor یک			
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	عدم آزمایش روش پیشنهادی با تعداد متخصصان بیشتر	دقت بالا در مقایسه با روش دقت بالا روی همه ۴ تخصص	محاسبه یک معیار ریسک بر اساس فاصله مهالنویس و چگالی‌ها و محاسبه ریسک و ساخت درخت تصمیم آن	Multi Stage Method to Detect Provider and Patient Fraud [۹۸]	۸

۲,۷ 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 داده‌های 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 عمدتاً 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 استاندارد 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^{۲۸} 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 در نظر گرفته می‌شود. این شرایط مفهوم کلان داده را در پی دارد.

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

فصل ۳

روش پیشنهادی

۳. روش پیشنهادی تشخیص تقلب

۳,۱ مقدمه

در فصل‌های گذشته با اهمیت موضوع تقلب در داده‌های سلامت آشنا شده و انواع مختلف تقلب در این داده‌ها بررسی گردید. پژوهش‌های پیشین صورت گرفته در این موضوع همگی دسته‌بندی و مورد بحث قرار گرفتند. در این فصل با ارایه روشی مبتنی بر گراف سعی در کلاس‌بندی داده‌هایی داریم که برچسب متقلب بودن یا نبودن مربوط به آنها وجود ندارد. به عبارت دیگر ابتدا مدل مجموعه داده‌ای شامل دو کلاس متقلب و عادی را دریافت می‌کند. سپس سعی میکند شباهت‌ها و الگوهای را پیدا کند که در کلاس متقلب و غیر متقلب تکرار می‌شود. در نهایت برای یک ورودی جدید که برچسب یا کلاس آن وجود ندارد، بعد از محاسبه میزان شباهت این ورودی جدید به هر کدام از کلاس‌های متقلب یا غیر متقلب، الگوریتم عملیات دسته‌بندی را انجام می‌دهد.

۳,۲ مجموعه داده استفاده شده

در این تحقیق از دو مجموعه داده استفاده شده است 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 موجودیت‌های 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.

مجموعه داده LEIE ، NPI هر ارائه دهنده را نشان می‌دهد ، که برای برچسب زدن ادعاهای جعلی استفاده می‌شود. مجموعه داده‌های موجود از CMS ، ۲۰۱۲ تا ۲۰۱۵ ، با از پایگاه داده LEIE با در نظر گرفتن دوره‌های 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 [13] LEIE برای بدست آوردن برچسب‌های تقلبها به مجموعه داده اضافه شدند.

این مجموعه داده ایجاد شده و ماهیانه توسط اداره بازرسی عمومی (OIG^{۲۹}) مطابق با بخش ۱۱۲۸ و ۱۱۵۶ قانون امنیت اجتماعی نگه داری می‌شود. در این پژوهش از آخرین نسخه‌ی سپتامبر ۲۰۲۰ استفاده شده‌است. OIG اختیار محرومیت افراد از برنامه‌های مراقبت سلامت فدرالی مانند بیمه پزشکی را داراست. متأسفانه LEIE فراگیر نیست و ۳۸٪ رایه دهندگان با محرومیت تقلب به جراحی ادامه می‌دهند و ۲۱٪ با وجود محرومیت از عمل جراحی تعلیق نشده‌اند. 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 شده‌اند در جدول زیر نشان داده شده‌است، که هر پزشک بدون NPI مقدار ۰ را در داده‌ی LEIE دارد.

در سطح رایه دهنده تجميع شده و اطلاعات خاصی با توجه به روندها، داروها یا تجهیزات مرتبط با فعالیت‌های کلاهبردارانه ندارد. دسته‌های مختلفی از محرومیت/ اخراج بر اساس شدت گناه وجود دارد که توسط شماره قوانین توصیف شده‌اند. ما از همه‌ی محرومیت‌ها استفاده نمی‌کنیم، بلکه رایه دهندگان محروم شده را با قوانین انتخاب شده‌ی نشان دهنده‌ی ارتکاب تقلب فیلتر می‌کنیم. جدول ۴ این قوانین را که منطبق بر محرومیت رایه 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.

جدول ۱-۳ قانون‌های مربوط به مجموعه داده LEI E

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

محکومیت سه جرم استثنا تخلفات نامعین	1128(C)(3)(G)(II)
-------------------------------------	-------------------

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

مجموعه داده‌ی Medicare Provider Utilization and Payment اطلاعات مطالبات برای هر روند(نسخه) که یک پزشک در یک سال انجام داده را فراهم میکند. اخیراً این مجموعه داده در وبسایت CMS از سال ۲۰۱۲ تا ۲۰۱۵ در دسترس است. پزشکان با استفاده از NPI^{۳۰} یکتا شناسایی شده اند، در حالی که روندها بر اساس کد HCPCS^{۳۱} 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 داده‌ی Medicare Provider Utilization and Payment را تجمیع کند با:

(۱) NPI مربوط به ارایه دهنده‌ی خدمات

(۲) کد HCPCS برای روند یا سرویس انجام شده

(۳) محل خدماتی که یک تسهیلات است(F^{۳۲}) یا غیر تسهیلات(O^{۳۳})، 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 داده‌ی Medicare Provider Utilization and Payment میتواند برای فراهم نمودن اطلاعات سطح روند در نظر گرفته شود.

۳,۲,۳ ویژگی‌های مورد استفاده

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

^{۳۰} National Provider Identifier

^{۳۱} Health Common Procedure Coding System

^{۳۲} Facility

^{۳۳} Non-facility

[7] OIG در نظر گرفتند ، همانطور که در جدول ۱-۳ ذکر شده است برای ساخت و آزمایش مدلهای ما ، تصور می کنیم تعدادی از پزشکان LEIE کلاهبردار تلقی می شوند و کسانی که شامل آنها نمی شوند تقلبی نیستند. داده های 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 ویژگی های عددی در این الگوریتم برای محاسبه درصد شباهت مورد استفاده قرار می گیرد لذا جدولی از مجموعه داده ی نهایی که در این پایان نامه استفاده شد در جدول ۲-۳ ارائه می شود.

جدول ۲-۳ ستونهای مجموعه داده

ردیف	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	HCPCS
۱۲	provider_type

همچنین در این مجموعه داده ویژگی های `nppes_provider_gender` و `provider_type` به دلیل اسمی بودن قابل استفاده در مدل به صورت مستقل نیستند لذا آن‌ها را به متغیر دسته ای^{۳۴} 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 همچنین یک فرد می‌تواند با اخذ تخصص‌های بیش‌تر شماره‌های نظام پزشکی بیشتری دریافت کند. لذا مشخصاتی همچون اسم نمی‌تواند کلید اولیه و منحصر بفرد 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 متغیر دسته‌ای تبدیل شد. سپس نمونه داده‌هایی 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 (در حدود چند گیابایت) کوچک می‌باشد. متأسفانه داده‌های اولیه‌ی قابل استفاده حجم کمی داشته و نمی‌توان از روش‌های یادگیری ماشین پیچیده برای آموزش و ایجاد مدل استفاده کرد.

۳,۴ نیازمندی‌های روش پیشنهادی

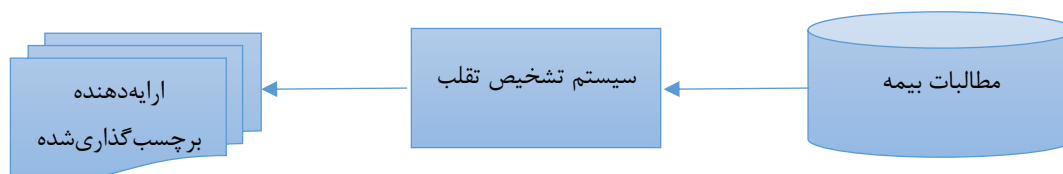
برای پیش بینی اینکه آیا یک ورودی جدید در دسته متقلب‌ها قرار می‌گیرد یا خیر، الگوریتم زیر طراحی گردیده است. این الگوریتم در ابتدا مجموعه داده‌ای دریافت کرده و آن را بر اساس برچسب‌هایش به دو دسته تقسیم می‌کند. در داده یک گره در گراف محسوب می‌شود. و داده‌هایی که در یک دسته قرار دارند توسط یال به هم وصل می‌شوند. 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 آن کلاس نباشد با هیچ کدام از اعضا یال مشترک نخواهد داشت. چنانچه شباهت هر یک از اعضای گراف با هم محاسبه شود می‌تواند گراف را به حالت وزن دار ترسیم کرد 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 می‌تواند 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 داده‌ها مشکلات و معضلات خود را به همراه خواهد داشت. لذا در این پژوهش ما به باینری بودن یال‌ها اکتفا می‌کنیم.

^{۳۴} Categorical variable

در ادامه با فرایند یادگیری و نحوه عملکرد الگوریتم بیشتر آشنا خواهیم شد شکل ۳-۱ معماری کلی سیستم پیشنهادی را نشان می‌دهد.



شکل ۳-۱ معماری کلی سیستم پیشنهادی

مجموعه داده آموزش که در این فصل برای آموزش الگوریتم مورد بررسی قرار گرفت ابتدا به عنوان ورودی به الگوریتم داده می‌شود. 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 می‌کند بردار مربوط به این داده جدید را با مرکز دسته‌هایی که داریم مقایسه می‌کند چنانچه داده جدید به یکی از این میانگین‌ها یا مرکز دسته‌ها شباهت بیشتری داشت، به آن کلاس تعلق می‌گیرد.

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 شیئی کمتر است. برای محاسبه شباهت از روشهای مختلفی می‌توان استفاده کرد که در زیر برخی از آنها بررسی می‌گردد.

۳،۴،۱ معیار شباهت کسینوسی

برای تبدیل کسینوس وزن دهی شده با معکوس درجه، بین یک ارائه دهنده و عضو مجموعه‌های مرجع مثبت یا منفی به ویژگی‌های قابل استفاده جهت تخمین (برآورد) ریسک، ما میانگین اعضای هر مجموعه را گرفته و بر اساس آن عمل می‌کنیم.

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

شکل ۳-۲ فرمول محاسبه معیار شباهت کسینوسی

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. دلیل اصلی استفاده از این معیار در این پایان‌نامه این است که این معیار یکی از پرکاربردترین معیارها است و 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 نیست نه اینکه لزوماً فرد متخلف نیست.

۳,۵ توصیف روش پیشنهادی

همانطور که گفته شد، متأسفانه 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 الزاماً 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.

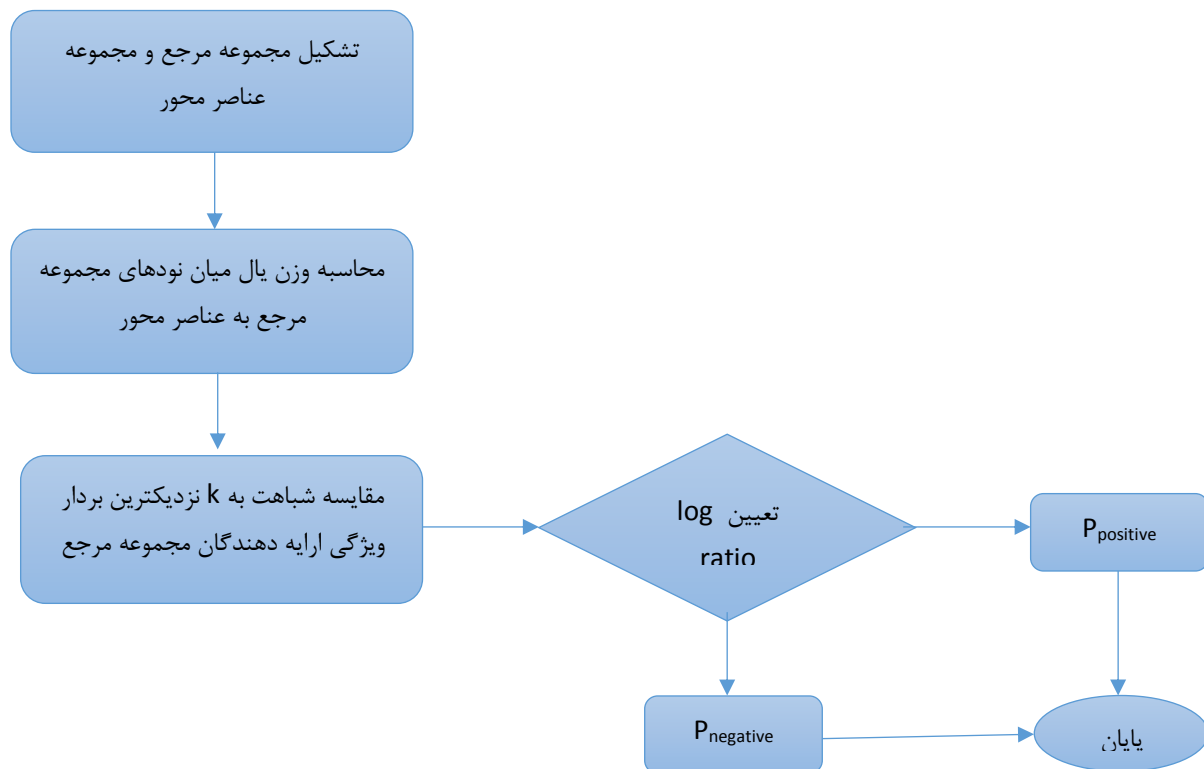
نمای شماتیکی از نحوه کار الگوریتم در شکل ۳-۳ ارائه شده است. همان‌طور که دیده می‌شود، یک مجموعه مرجع^{۳۵} از ارایه دهندگان تشکیل و به دو بخش مثبت و منفی تقسیم می‌شوند، مجموعه ویژگی‌های HCPCS نیز در دسته‌ی عناصر محور^{۳۶} قرار می‌گیرند و چنانچه هر موجودیت از ارایه دهندگان، خدمتی ارایه نموده باشد یالی میان اعضای مجموعه مرجع و pivot ها متصل می‌شود؛ وزن هر یال از حاصلضرب bene_unique_cnt در average_Medicare_payment_amt محاسبه می‌شود. برای هر ارایه دهنده یک بردار رفتار به این صورت تعریف می‌کنیم که دربرگیرنده برداری از وزن یال‌های رابط به هر نود محوری مجاور است. گراف این مساله وزن دار نیست و لذا اعضای هر کلاس ارتباط شود کامل باهم و هیچ ارتباطی با کلاس دیگر ندارند.

برای هر نمونه داده‌ی جدید مقدار شباهت بردار ویژگی‌های آن با بردار ویژگی‌های هر گروه محاسبه می‌شود. در اینجا از شباهت کسینوسی استفاده شده است. 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^{۳۷} (KNN) به صورت محلی عمل کرده و برای نمونه‌ی جدید تصمیم‌گیری می‌کند. تعداد همسایگی در این تحقیق ۵ در نظر گرفته شده است. این مقدار بهینه نشده است.

^{۳۵} Reference set

^{۳۶} Pivot

^{۳۷} K- Nearest Neighbor



شکل ۳-۳ معماری جزئیات سیستم پیشنهادی

بعد از محاسبه‌ی امتیازهای میانگین محاسبه شده که عضویت به دو گروه را نشان می‌داد، مقدار $\log ratio$ یا نسبت بخت محاسبه می‌شود. در نهایت از روی مقدار $\log ratio$ و فرمول مربوط به آن طبق زیر گروه جدید به یکی از زیر گراف‌ها نسبت داده می‌شود.

$$\log ratio = \frac{avg(p_1, \dots, p_n)}{avg(n_1, \dots, n_n)} \begin{cases} pos & > 1 \\ neg & < 1 \end{cases}$$

شکل ۴-۳ فرمول محاسبه $\log ratio$

۳,۶ 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 و روش کار آن تشریح گردید. در ادامه با توجه 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 ۴

پیاده‌سازی و ارزیابی

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. الگوریتم‌های مبتنی بر یادگیری ماشین که روی کلان داده‌های مربوط به این حوزه آزمایش 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 جزئیات 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 و روال اعتبار سنجی متقابل

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 می کنیم. از آنجایی که مساله تحلیل احساسات، خود زیرمجموعه‌ای از روش‌های با ناظر و یا همان طبقه‌بندی است، معروف ترین معیارهای عرصه طبقه‌بندی بهترین گزینه برای سنجش عملکرد الگوریتم ما هستند.

در ارزیابی تقلب پزشکی دو حالت در نظر می گیریم؛ 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.

نتایج روش پیشنهادی با نتایج پژوهش [24] که از مجموعه داده ای مشابه استفاده نموده‌اند مقایسه می گردد.

اگرچه حل یک مساله شخص از رویکرد های مختلف بسیار با ارزش و به خودی خود نوآوری 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^{۳۸} داده‌های به صورت تصادفی به ده بخش تقسیم

^{۳۸} 10 – Fold Cross Validation

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.

۴,۲,۱ 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 اعداد درک و خوانایی سختی دارند لذا در ادامه با استفاده از سایر معیارهای معرفی شده به درک بهتری از عملکرد الگوریتم خواهیم رسید.

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

شکل ۱-۴ نتایج معیارهای پایه برای الگوریتم

۴,۲,۲ حساسیت^{۳۹}

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 و به شکل زیر محاسبه و مقدار آن در جدول ۴-۱ گزارش شده است که میزان احتمال مثبت شدن صحیح نتیجه آزمون وقتی که نمونه مرتکب تقلب شده است، را نشان می‌دهد.

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

شکل ۴-۲ فرمول sensitivity

^{۳۹} Sensitivity

۴,۲,۳ تشخیص پذیری^{۴۰}

تشخیص پذیری یا ویژگی نیز به احتمال منفی شدن صحیح نتیجه آزمون وقتی که نمونه سالم (فاقد آلودگی) است، اشاره می کند. که مقدار آن بصورت زیر محاسبه و نتیجه در جدول گزارش شده است.

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

شکل ۳-۴ فرمول specificity

۴,۲,۴ منحنی AUC^{۴۱}

مقدار AUC در این تحقیق نسبت به روش LOF^{۴۲} که بهترین نتیجه را در میان سایر روش ها در تحقیق [24] در پی داشته است، به میزان ۰,۰۱۱ بهبود را نشان می دهد که بیانگر این می باشد که با وجود سادگی روش و عدم نیاز به محاسبات پیچیده، روش آزمون از قدرت تشخیص بهتری برخوردار است.

۴,۲,۵ نتایج حاصل از اجرا روی ماتریس درهم ریختگی

این ماتریس از معروف ترین ابزارهای سنجش عملکرد روش های باناظر است و حتی گاهی در روش های بدون ناظر هم از آن استفاده می شود. (Wan بدون تاریخ)

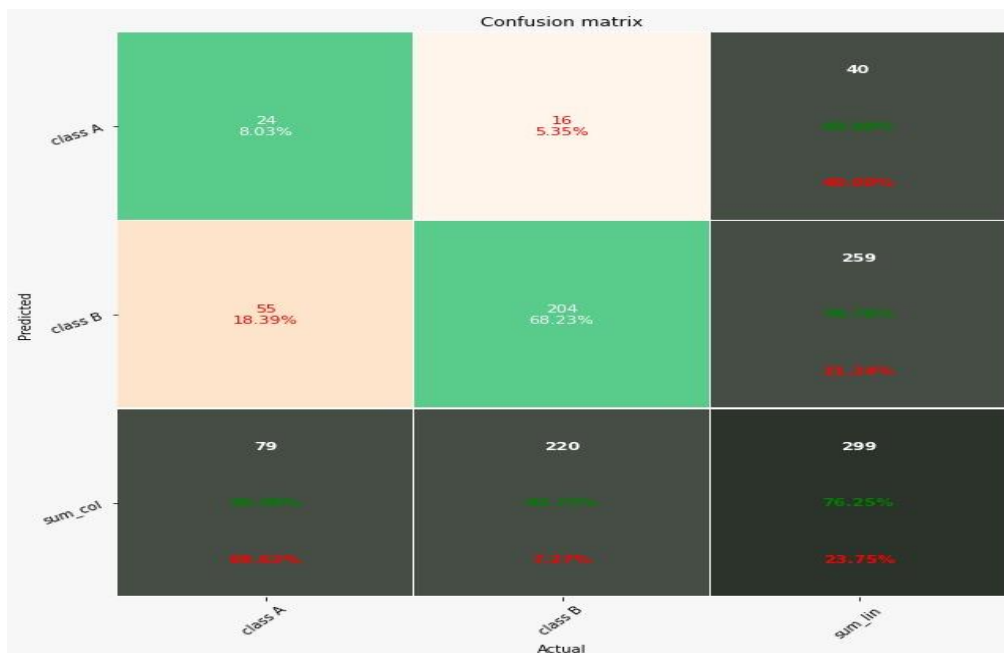
ماتریس پیچیدگی تعداد نمونه های واقعی را با تعداد نمونه های پیش بینی شده مقایسه میکند. با توجه به ماتریس نتایج، ما از AUC [۶۷,۶۸] برای ارزیابی کارایی تشخیص ثقل استفاده می کنیم AUC ناحیه زیرمنحنی ROC (Receiver Operating Characteristic) است و ROC مقایسه بین True Positive و False Positive است. Recall از TP/TP+FN بدست می آید. تعاریف برای TP و TN و FN و FP مستقیماً از ماتریس پیچیدگی بدست می آیند.

در شکل ۴-۵ ماتریس در هم ریختگی حاصل از اجرای الگوریتم روی مجموعه داده ی انتخابی دیده می شود.

^{۴۰} specificity

^{۴۱} Area under the ROC curve

^{۴۲} Local Outlier Factor



شکل ۴-۴ ماتریس در هم‌ریختگی برای نتایج روش پیشنهادی

۴,۳ مقایسه با سایر الگوریتم‌های موجود

در جدول ۱-۴ مشاهده می‌شود الگوریتم ارائه شده 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 نشده است.

جدول ۱-۴ مقایسه نتایج مدل با سایر الگوریتم‌ها

Specificity	Sensitivity	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	PROPOSED

۴,۴ بررسی بازه اطمینان نتایج الگوریتم

در مباحث آمار و داده‌کاوی، از عبارت فاصله اطمینان^{۴۳} استفاده می‌کنند تا نشان دهند که تقریباً مطمئن هستیم یک فاصله یا محدوده‌ای عددی، شامل پارامتر مورد جامعه است. لذا فاصله اطمینان نوعی برآورد فاصله‌ای در نظر گرفته شده و هر قدر 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 Sensitivity, AUC و Specifity بازه اطمینان محاسبه شد و همان‌طور که در جدول (۹) از نتایج مشخص است این بازه در حد قابل قبولی است.

جدول ۴-۲ بازه اطمینان برای نتایج الگوریتم

حد پایین بازه اطمینان	حد بالای بازه اطمینان	نتیجه	معیار
0.64	-0.0246	0.0246	AUC
0.5	-0.239	0.239	Sensitivity
0.5	-0.239	0.239	Specificity

۴,۵ 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.

^{۴۳} Confidence Interval

فصل ۵

نتیجه‌گیری و توسعه‌های آتی

۵. نتایج و توسعه‌های 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^{۴۴} یک فریب عمدی یا ارائه اطلاعات نادرست است که توسط یک شخص یا یک موجودیت با علم به اینکه این فریب می‌تواند منجر به 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 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 به یک هدف جذاب برای کلاهبرداری تبدیل شود. شرکت‌های بیمه روش‌هایی را برای تشخیص تقلب ایجاد می‌کنند که عمدتاً 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 مؤثر تقلب برای کاهش هزینه‌ها و بهبود کیفیت سیستم سلامت بسیار مهم است. به منظور دستیابی به کشف مؤثرتر 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.

National HealthCare Anti-Fraud Association^{۴۴}

۵,۲ نتیجه‌گیری

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 آسیبهای جبران ناپذیر خواهد شد. از این رو لازم است با استفاده از سیستم‌های 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 برجسب‌دار، دارای نرخ کشف کاذب کمتر و کاهش هزینه‌ی برجسب داده‌ها اشاره کرد. در جدول (۱۰) این الگوریتم‌ها به همراه مزایا و معایب آن‌ها آماده‌است.

جدول ۵-۱ دسته‌بندی به تفکیک رویکردهای کلی کشف تقلب

رویکردها	مزایا	معایب
روش‌های آماری	<ul style="list-style-type: none"> • به سرعت ارائه‌دهندگان مشکوک را شناسایی می‌کند. • می‌تواند نوع جدیدی از تقلب را شناسایی کند. 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • سادگی و عدم نیاز به پردازشگر قوی 	<ul style="list-style-type: none"> • نیاز به متخصص برای داده‌ی برجسب دار • overfitting
روش‌های یادگیری ماشین بدون ناظر	<ul style="list-style-type: none"> • هزینه‌ی کمتر به دلیل عدم نیاز به داده‌ی برجسب‌دار 	<ul style="list-style-type: none"> • تغییر مداوم وزن‌ها و پارامترها توسط متخصصان برای شناسایی انواع جدید تقلب

<ul style="list-style-type: none"> تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب هزینه‌ی بالای داده‌ی برچسب‌دار 		
<ul style="list-style-type: none"> پوشش کم موارد تشخیص تقلب 	<ul style="list-style-type: none"> در مقایسه با روش‌های کاوش فرایند دارای نرخ کشف کاذب^{۴۵} کمتر است. کاهش هزینه‌ی برچسب داده‌ها 	روش‌های یادگیری ماشین ترکیبی

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 مشکوک و یا 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 با سایر 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.

متأسفانه 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 را برای 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 محدودیت ها در این بازه زمانی قابل پیاده سازی در این پایان نامه نبود مطرح می گردد.

۵.۳ توسعه‌های 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 بالای ۷۵٪ تقلب- احتمال بالای ۵۰٪ تقلب- احتمال زیر ۵۰٪ تقلب- احتمال زیر ۲۵٪ تقلب) و ...

- [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 .
- [۴] ح. ع. م. ج. تاریخ, "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.
- [۵] 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.
- [۶] د. س. ح. ه. د. ع. ز. د. ع. ربیعی, "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.
- [۱۴] E. Turban, R. Sharda, and D. Delen, "Decision support and business intelligence systems (required)," *Google Scholar*, 2010.
- [۱۵] 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.
- [۱۶] 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 .

- [17] 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.
- [18] 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.
- [19] 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.
- [20] H. A. Abbass, J. Bacardit, M. V. Butz, and X. Llorca, "Online adaptation in learning classifier systems: stream data mining," *Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign, IlliGAL Report*, no. 2004031, 2004.
- [21] 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 .
- [22] 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.
- [23] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," *Credit scoring and credit control VII*, pp. 235-255, 2001.
- [24] Da Rosa, Raquel C. "An Evaluation of Unsupervised Machine Learning Algorithms for Detecting Fraud and Abuse in the US Medicare Insurance Program." Florida Atlantic University, Boca Raton, FL (2018): 9-27.
- [25] 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 .
- [26] 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.
- [27] C. S. Hilaris 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 .
- [28] 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.
- [29] 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.
- [30] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
- [31] 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.
- [32] 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. ١٧٤-١٦٨ .
- [٣٤] N. Sánchez-Maróño, 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. 269.٢٠١٣ ,٢٨٢-
- [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. Ekin, F. Ieva, 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 ".2013Humaine ,association conference on affective computing and intelligent interaction. IEEE 2013.