

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه الزهرا - دانشکده فنی و مهندسی

پایان نامه جهت اخذ درجه کارشناسی ارشد

رشته مهندسی کامپیوتر - گرایش هوش مصنوعی

عنوان

کشف تقلب در سیستم های مراقبت سلامت با رویکرد تحلیل گراف

استاد راهنما

دکتر محمد رضا کیوان پور

دانشجو

روناک نمکی

تابستان ۱۳۹۹

کد: EM-FR-۰۷-۰۱ مدرسه	صورت جلسه دفاع از پایان نامه دوره کارشناسی ارشد ورودی ۹۶ به بعد																														
شماره: تاریخ:	بازگرمی: ۱ تاریخ بازگرمی: ۱۳۹۶/۱۲/۲۱																														
<input type="checkbox"/> نسخه پرونده دانشجو <input type="checkbox"/> نسخه تحصیلات تکمیلی <input type="checkbox"/> نسخه مالی (به تعداد مورد نیاز)																															
جلسه دفاع از پایان نامه تحصیلی خاتم به شماره دانشجویی دانشجوی کارشناسی ارشد رشته / گرایش یا کد درس به ارزش واحد یا عنوان: در تاریخ / / از ساعت تا یا حضور هیأت داوران تشکیل شد و پس از ارزیابی، اختیار پایان نامه برای اخذ مدرک کارشناسی ارشد																															
<input type="checkbox"/> مورد تایید قرار گرفت <input type="checkbox"/> مورد تایید قرار نگرفت																															
اختیار پایان نامه برای اخذ مدرک کارشناسی ارشد به شرح زیر مورد تأیید است: الف) قبول با درجه <input type="checkbox"/> عالی (نمره ۱۹ تا ۲۰) <input type="checkbox"/> خیلی خوب (نمره ۱۸ تا ۱۸/۹۹) <input type="checkbox"/> خوب (نمره ۱۶ تا ۱۷/۹۹) <input type="checkbox"/> متوسط (نمره ۱۴ تا ۱۵/۹۹) ب) <input type="checkbox"/> مردود (نمره کمتر از ۱۴)																															
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 25%;">هیأت داوران</th> <th style="width: 25%;">نام و نام خانوادگی</th> <th style="width: 25%;">امضا</th> <th style="width: 25%;">تاریخ</th> </tr> </thead> <tbody> <tr> <td>استاد راهنمای اول</td> <td></td> <td></td> <td></td> </tr> <tr> <td>استاد راهنمای دوم</td> <td></td> <td></td> <td></td> </tr> <tr> <td>استاد مشاور اول</td> <td></td> <td></td> <td></td> </tr> <tr> <td>استاد مشاور دوم</td> <td></td> <td></td> <td></td> </tr> <tr> <td>داور داخلی ++</td> <td></td> <td></td> <td></td> </tr> <tr> <td>داور خارجی ++</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>				هیأت داوران	نام و نام خانوادگی	امضا	تاریخ	استاد راهنمای اول				استاد راهنمای دوم				استاد مشاور اول				استاد مشاور دوم				داور داخلی ++				داور خارجی ++			
هیأت داوران	نام و نام خانوادگی	امضا	تاریخ																												
استاد راهنمای اول																															
استاد راهنمای دوم																															
استاد مشاور اول																															
استاد مشاور دوم																															
داور داخلی ++																															
داور خارجی ++																															
نام و نام خانوادگی استاد ناظر تحصیلات تکمیلی: امضا: تاریخ: نام و نام خانوادگی مدیر گروه: امضا: تاریخ: نام و نام خانوادگی معاون آموزشی / ریس دانشکده: امضا: تاریخ:																															

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

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

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

کلیه دستاوردهای این تحقیق
متعلق به دانشگاه الزهراء^(س) است.

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

یارب دل ما را تو به رحمت جان ده درد	همه را به صابری درمان ده
این بنده چه داند که چه می باید جست	داننده تویی هر آنچه دانی آن ده

چکیده:

بیمه مراقبت‌های بهداشتی یک مشکل مبرم است و موجب افزایش هزینه‌های قابل توجهی در برنامه‌های بیمه درمانی می‌شود؛ بطوریکه کلاهبرداری در حوزه بهداشت و درمان (HCF) یک کلاهبرداری چند میلیارد دلاری است. وسیع بودن حوزه سلامت و حجم زیاد مالی باعث شده تا این حوزه برای کلاهبرداری مورد هدف قرار بگیرد. بنابراین حوزه سلامت به یک منبع هزینه‌ای قابل توجه در بسیاری از کشورها تبدیل شده است. یکی از منابع هزینه‌های قابل توجه سازمان بهداشت و سلامت، پرداخت سهم بیمه داروهای تجویز شده برای بیماران تحت پوشش است. بطور کلی، هدف از تشخیص تقلب، به حداکثر رساندن پیشبینی‌های درست و حفظ پیشبینی‌های نادرست در یک سطح قابل قبول از هزینه می‌باشد. با وجود تغییرات پیوسته رفتار متقلبین، مدل‌هایی که براساس تحلیل داده‌های گذشته ساخته میشوند ممکن است نتوانند شکل‌های جدید تقلب را شناسایی کنند. همچنین، هیچ یک از سیستم‌های شناسایی تقلب به تنهایی نمیتواند تمام شکل‌های تقلب را شناسایی و پوشش دهد و همچنین بیان گردید که روش ترکیبی مدل‌های تشخیص ناهنجاری و روش سوءاستفاده می‌تواند تشخیص تقلب و امنیت سیستم را بهبود ببخشد. در این پایان‌نامه یک رویکرد نوین برای تخمین احتمال تقلب در اسناد درمانی با روش تحلیل گراف مورد بررسی قرار گرفت. در این پژوهش، الگوریتم‌های شبکه، به گراف‌های حاصل از مجموعه داده اعمال می‌شود. یک گروه از الگوریتم‌ها، شباهت‌های رفتاری را در دو دسته‌ی ارائه‌دهندگان مراقبت‌های بهداشتی کلاهبردار و غیرکلاهبردار با توجه به معیارهای قابل اندازه‌گیری فعالیت‌های مراقبت‌های بهداشتی مانند روش‌های پزشکی و تجویز داروها، محاسبه می‌کنند. مجموعه دیگری از الگوریتم‌ها، میزان انتشار خطر ناشی از تقلب ارائه‌دهندگان مراقبت‌های بهداشتی را از طریق همایند جغرافیایی، تکرار مکان‌های مشترک یا آدرس‌های دیگر، تخمین می‌زنند. این الگوریتم‌ها با توجه به توانایی آن‌ها در پیشبینی حضور یک ارائه‌دهنده در لیست ارائه‌دهندگان دفتر بازرسی کل (محرومیت از مشارکت در بیمه پزشکی سالمندان و سایر برنامه های مراقبت های بهداشتی فدرال)، ارزیابی شده‌اند. در یک ارزیابی تجربی، ترکیبی از ۱۱ ویژگی در پیش‌بینی محرومیت (حضور در لیست) از $F1\text{-score}$ (*) و ROC (*) برخوردار بوده است. تجزیه و تحلیل ابلیشن نشان داده است که بیشتر دقت پیش‌بینی نتیجه ویژگی‌هایی است که میزان انتشار خطر را از طریق همایند جغرافیایی اندازه‌گیری کردند.

واژگان کلیدی: نسخه دارویی، کشف ناهنجاری، سیستم‌های مراقبت سلامت، تحلیل گراف

فصل اول

مقدمه	۱.....
۱-۱ مقدمه	۲.....
۲-۱ سیستم‌های بیمه سلامت	۴.....
۱-۲-۱ سندسازی، تقلب و سوءاستفاده از بیمه	۵.....
۲-۲-۱ سند سازی توسط ارائه‌دهندگان خدمات بهداشتی و درمانی	۵.....
۳-۲-۱ شاخص‌های بالقوه سندسازی، تقلب و سوءاستفاده در بیمه	۶.....
۳-۱ طرح مسئله	۸.....
۴-۱ اهداف و نوآوریها	۹.....
۵-۱ محدودیت های تحقیق	۹.....
۶-۱ ساختار پایان نامه	۱۰.....
۷-۱ جمع بندی	۱۱.....

فصل دوم

۱-۲ مقدمه	۱۳.....
۲-۲ تعریف تقلب	۱۳.....
۳-۲ نقش داده‌کاوی در تشخیص تقلب	۱۳.....
۴-۲ بازیگران نظام سلامت	۱۴.....
۵-۲ دسته بندی چالش‌های کشف تقلب	۱۷.....
۱-۵-۲ چالش‌های کشف تقلب از منظر سازمان بیمه	۱۷.....
۲-۵-۲ چالش‌های کشف تقلب از منظر داده	۱۷.....
۳-۵-۲ چالش‌های کشف تقلب از منظر مقالات	۱۷.....
۶-۲ مفاهیم پایه	۱۸.....
۱-۶-۲ مفهوم رانش	۱۸.....

۱۹.....	۲-۶-۲ مفهوم توزیع اریب کلاس‌ها.....
۲۱.....	۳-۶-۲ مفهوم کاهش داده‌ها.....
۲۲.....	۴-۶-۲ مفهوم پشتیبانی تشخیص آنلاین.....
۲۳.....	۵-۶-۲ ادغام در نظام مراقبت‌های بهداشتی.....
۲۳.....	۶-۶-۲ خطای پیش‌بینی و اثر "Black Swan".....
۲۳.....	۷-۶-۲ از دست دادن اطلاعات در پیش پردازش.....
۲۴.....	۸-۶-۲ خودکارسازی فرآیند داده‌کاوی برای کاربران غیرمتخصص.....
۲۴.....	۹-۶-۲ ماهیت بین رشته ای تحقیق و دانش متخصص حوزه.....
۲۴.....	۷-۲ انواع ناهنجاری در تشخیص تقلب.....
۲۵.....	۱-۷-۲ تشخیص ناهنجاری گراف ایستا.....
۲۵.....	۱-۱-۷-۲ تشخیص ناهنجاری مبتنی بر ساختار.....
۲۶.....	۲-۱-۷-۲ تشخیص ناهنجاری مبتنی بر اجتماع.....
۲۶.....	۲-۷-۲ تشخیص ناهنجاری گراف پویا.....
۲۶.....	۱-۲-۷-۲ تشخیص ناهنجاری مبتنی بر فاصله.....
۲۷.....	۲-۲-۷-۲ تشخیص ناهنجاری مبتنی بر فشرده سازی.....
۲۷.....	۳-۲-۷-۲ تشخیص ناهنجاری مبتنی بر تجزیه.....
۲۷.....	۴-۲-۷-۲ تشخیص ناهنجاری مبتنی بر خوشه یا اجتماع.....
۲۷.....	۵-۲-۷-۲ تشخیص ناهنجاری مبتنی بر مدل‌های احتمالاتی.....
۲۸.....	۶-۲-۷-۲ تشخیص ناهنجاری مبتنی بر پنجره.....
۲۸.....	۸-۲ تعریف Big Data و کاربرد آن در تقلب.....
۲۹.....	۹-۲ برخی معیارهای تحلیل شبکه‌ی پزشکان.....
۳۰.....	۱۰-۲ کلان داده‌ها در تشخیص تقلب.....
۳۰.....	۱۱-۲ رویکردهای کلی روش‌های کشف تقلب.....
۳۱.....	۱-۱۱-۲ روش‌های با ناظر.....

۳۲.....	۲-۱۱-۲ روش‌های بدون ناظر.....
۳۷.....	۲-۱۱-۳ روش‌های یادگیری ماشین ترکیبی.....
۴۰.....	۲-۱۱-۴ جمع‌بندی.....
۴۲.....	مراجع.....

- شکل (۱-۱) نمودار کلی گردش اطلاعات در سامانه‌های بیمه ۴
- شکل (۲-۱) درخت دسته‌بندی بازیگران نظام سلامت ۱۶
- شکل (۲-۲) توزیع مقالات FDS براساس مسائل و چالش‌های بین سال‌های ۱۹۹۴ تا ۲۰۱۴ ۱۸
- شکل (۲-۳) فرآیند یادگیری افزایشی در هر زمان t ۱۹
- شکل (۲-۴) مجموعه داده آموزشی ناهمگون UCSD ۲۰
- شکل (۲-۵) روش‌های رسیدگی به داده‌های ناهمگون ۲۱
- شکل (۲-۶) روش‌های کاهش داده ۲۲
- شکل (۲-۷) : پنج ۷ داده‌های بزرگ ۲۸

- جدول (۱-۱): انواع تقلب در بیمه سلامت..... ۶
- جدول (۲-۱): دسته‌بندی به تفکیک رویکردهای کلی کشف تقلب..... ۳۱
- جدول (۲-۲): مثالی از میانگین گروهی برای PGA ۳۵
- جدول (۲-۳): انواع رویکردها و روش‌های موجود در کشف تقلب سیستم سلامت ۳۷

فصل اول

مقدمه

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 متقلبان‌های 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 هزینه ای بالغ بر ۵۰ هزار میلیارد تومان یکی از کلیدی‌ترین هزینه‌های تحت مدیریت و نظارت دولت است که با تخمینی معادل ۳ تا ۱۰ درصد یعنی ۱۵ تا ۵۰ هزار میلیارد ریال تقلب مواجه است.

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

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

مراقبت سلامت در آمریکا از سال ۱۹۸۰ یک هزینه‌ی بزرگ ایجاد نموده است. بر اساس گزارشی از GAO^۱ به کنگره در سال ۲۰۰۴، هزینه‌های سالانه مراقبت سلامت به دو بیلیون دلار نزدیک شد که ۳/۱۵٪ تولید ناخالص داخلی برآورد شد. 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^۲ 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. NHCAA محتاطانه تخمین می‌زند که حداقل ۳٪ یا بیش از ۶۰ بیلیون دلار از هزینه‌های مراقبت سلامت سالانه‌ی آمریکا به دلیل تقلب آشکار از بین رفته است. دولت و سازمان‌های اجرای قانون این خسارت را ۱۰٪ یا ۱۷۰ بیلیون دلار تخمین

^۱ General Accounting Office

^۲ National HealthCare Anti-Fraud Association

می‌زند. 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. در سال‌های اخیر سیستم‌ها برای پردازش مطالبات الکترونیکی بطور فزاینده‌ای پیاده‌سازی شده‌اند تا بازرسی و بازبینی داده‌ی مطالبات، خودکار انجام شود. این سیستم‌ها برای شناسایی مناطقی که نیاز به توجه خاص دارند مانند داده‌ی ورودی اشتباه یا ناقص، تکرار مطالبات و خدمات پزشکی که تحت پوشش نیستند، طراحی شده‌اند. اگرچه این سیستم‌ها ممکن است برای تشخیص انواع خاصی از تقلب بکار روند، قابلیت تشخیص تقلب آن‌ها معمولاً محدود است از آنجایی که تشخیص عمدتاً برپایه‌ی قوانین ساده‌ی از پیش تعیین شده است که توسط متخصصان حوزه مشخص می‌شوند [۲].

سهم قالب هزینه صنعت بیمه سلامت در ایران توسط دو نهاد اصلی سازمان بیمه سلامت ایران و سازمان تأمین اجتماعی، که هر دو توسط دولت مدیریت می‌شوند، پرداخت می‌شود. برآورد به دست آمده از قانون بودجه سال ۱۳۹۶ حاکی از ۱۴.۳۹۳ میلیارد تومان هزینه در حوزه وزارت بهداشت است که به طور مشخص ۸.۹۵۸ میلیارد تومان به صورت مستقیم توسط سازمان بیمه سلامت ایران هزینه خواهد شد. به این اعداد باید پرداخت‌های مستقیم مردم در قالب فرانشیز اضافه گردد. با توجه به اینکه سالانه ۳٪ تا ۱۰٪ از هزینه‌های بیمه سلامت به صورت متقلبانه دریافت می‌شود و عدد مورد نظر به ۱۰٪ نزدیکتر است. این عدد برای سال ۹۶ به ۸۹۵ میلیارد تومان رسید.

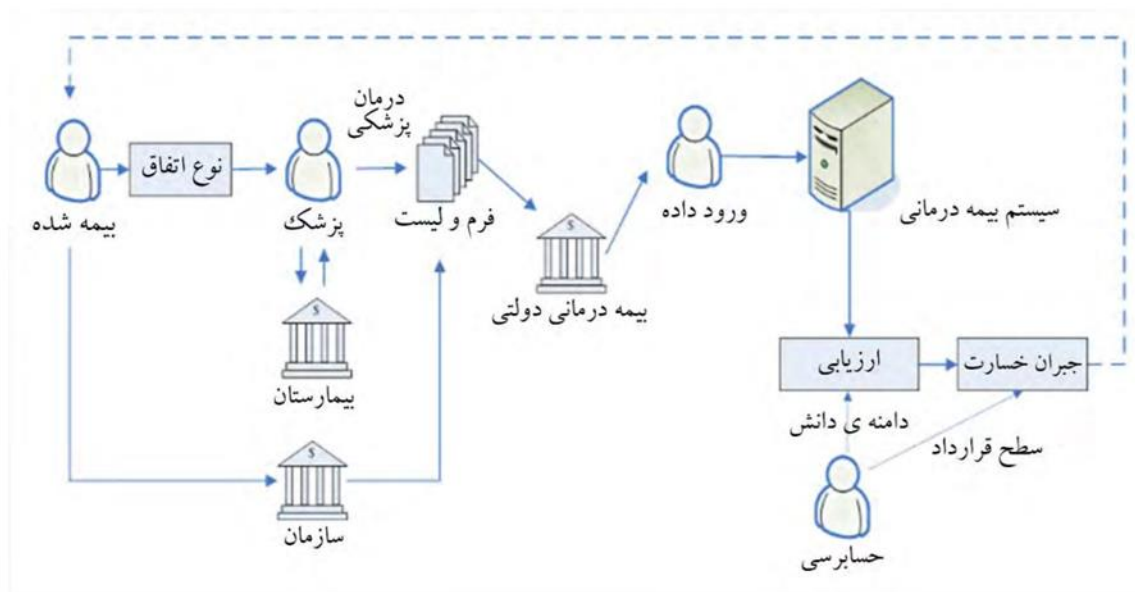
تقلب در حوزه سلامت یک جرم بزرگ است و هزینه‌های شخصی و بودجه‌ای قابل توجهی به افراد، دولت‌ها و جامعه وارد میکند. بنابراین، کشف مؤثر تقلب برای کاهش هزینه‌ها و بهبود کیفیت سیستم سلامت بسیار مهم است. به منظور دستیابی به کشف مؤثرتر تقلب، بسیاری از پژوهشگران رویکردهای ضد تقلب پیچیده‌ای بر پایه داده‌کاوی، یادگیری 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 [۳].

سازمان‌های بیمه‌گر پس از دریافت اسناد هزینه از ارائه‌دهندگان خدمات سلامت اعم از مراکز درمانی، پزشکان، داروخانه‌ها، آزمایشگاه‌ها فرآیند بررسی هزینه‌ها و تطبیق آن‌ها با معیارها و جداول هزینه شده توسط آن‌را 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 کوتاهتری از سبد هزینه سلامت حذف می‌گردد. همچنین با وجود حجم زیاد داده‌ها و متنوع بودن داده‌ها در حوزه سلامت، روش‌های سنتی یادگیری ماشین برای کشف تقلب در حوزه‌ی سلامت کافی نیستند. به همین دلیل استفاده از روش‌های تحلیل دادگان انبوه می‌تواند به فهم دقیقتر داده‌ها کمک کند. علاوه بر این، تحلیل دادگان انبوه نه تنها قادر به پردازش حجم انبوه داده هستند بلکه به خوبی از پردازش موازی داده‌ها پشتیبانی میکند [۳].

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 استانداردپذیر کردن داده‌های موجود در پایگاه داده‌ای 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.

۲-۱ سیستم‌های بیمه سلامت

انجمن بیمه سلامت آمریکا، بیمه سلامت را به عنوان پوششی علیه ریسک هزینه‌های درمانی به علت بیماری یا آسیب دیدگی تعریف می‌کند. این پوشش می‌تواند توسط بعضی سازمان‌های مرکزی، برای مثال شرکت‌های خصوصی یا دولتی، ارائه شود. منبع این پوشش در بسیاری از کشورها صرف نظر از سیستم‌های بهداشت و درمانشان، متفاوت است. بررسی سالیانه انجام شده توسط صندوق مشترک المنافع، سیستم‌های بهداشت و درمان استرالیا، نیوزیلند، بریتانیا، آلمان، کانادا و ایالات متحده را مقایسه می‌کند. این بررسی تاکید می‌کند که ایالات متحده تنها کشور بدون پوشش بیمه سلامت سراسری است. اداره آمار ایالات متحده بیان می‌کند که ۳۱ درصد از آمریکایی‌ها طرح بیمه سلامت عمومی دارند، در حالی که ۵۵ درصد از آنها پوشش خود را از طریق کارفرمایانشان می‌گیرند. اگرچه، تحت پوشش بودن تضمین نمی‌کند که شخص بیمه شده هیچ هزینه پزشکی پرداخت نکند. میزانی که بیمه شده باید بپردازد، قبل از اینکه بیمه‌گر برای یک ویزیت یا خدمت خاص بپردازد، پرداخت مشترک نامیده می‌شود. جدای از پرداخت مشترک، ممکن است خدماتی باشد که بیمه‌گر بر اساس حق بیمه‌ای که بیمه شونده می‌خرد، بازپرداخت می‌کند. مانند خدماتی که به عنوان بیمه تکمیلی شناخته می‌شوند که در آن درصد بیشتری از هزینه‌ها در قبال دریافت حق بیمه بیشتر پرداخت می‌شود. روند گردش اسناد در سیستم بیمه در شکل (۱-۱) نشان داده شده است.



شکل (۱-۱): نمودار کلی گردش اطلاعات در سامانه‌های بیمه

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.

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.

۱-۲-۱ سندسازی، تقلب و سوءاستفاده از بیمه

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

۱-۲-۲ سند سازی توسط ارائه‌دهندگان خدمات بهداشتی و درمانی

هشتاد درصد هزینه بهداشت و درمان مربوط به تصمیم پزشکان درباره خدماتی است که بیماران نیاز دارند. بنابراین، تقلب و سوء استفاده رخ داده توسط پزشکان می‌تواند خیلی قابل توجه باشد [۴]. البته دلایل و انگیزه‌هایی وجود دارد که چرا پزشکان، قانون مربوط به تقلب و سوءاستفاده را زیر پا می‌گذارند. دیدی که پزشکان از فعالیت خود به عنوان کسب و

کار دارند، میتواند نقشی حیاتی در ارتکاب به تقلب یا سوء استفاده ایفا کند. برای مثال، هزینه صدور صورتحساب میتواند انگیزه بزرگی برای پزشکانی باشد که خودشان را به عنوان یک فروشنده می بینند. پزشکان می توانند اقدامات غیرضروری برای افزایش هزینه ها انجام دهند. اگرچه این روش ها بر سابقه پزشکی بیمار اثر میگذارد و آن را تحریف می کند و ممکن است منجر به درمان اشتباه در آینده شود. از طرف دیگر، پزشکان ممکن است در شرایط دشواری بین انتخاب تعهد حرفه‌ای در مقابل بیماران یا قوانین پوشش مندرج در قرارداد شان قرار گیرند. برای مثال، برخی پزشکان ممکن است در شرایط بیمار اغراق کنند یا درخواست آزمایشی را بکنند که نشان دهد این دارو یا درمان برای بیمار ضروری است، تا در بدست آوردن پوشش اضافه به آنها کمک کنند[۴].

۱-۲-۳ شاخص‌های بالقوه سندسازی، تقلب و سوءاستفاده در بیمه

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

جدول (۱-۱): انواع تقلب در بیمه سلامت

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

کدگذاری اشتباه فعالیت‌ها، می‌تواند سرنخ‌هایی از تقلب و سوءاستفاده داشته باشد. کدگذاری فعالیت‌ها زمانی رخ می‌دهد که ارائه‌کنندگان خدمات بهداشتی و درمانی از کدی استفاده می‌کنند که گران‌تر از خدمات بهداشت و درمان، تست‌ها، یا آیت‌هایی است که واقعاً 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 آمریکا اقتباس شده است. هر چند در این کتاب 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 جزئی‌تر به چندین کد جزئی‌تر، روشی دیگر برای افزایش هزینه و بدست آوردن منفعت غیر مجاز است. درمان‌ها یا آزمایش‌هایی وجود دارند که شامل بیش از یک خدمت است. وقتی این خدمات با هم انجام شوند، تامین کننده خدمات بهداشتی و درمانی نیاز به استفاده از کدهای مشخصی دارد که دو خدمت یا بیشتر را گروه بندی کند. اگر تامین کننده خدمات بهداشتی و درمانی از این کدهای صورتحساب مشخص، برای تمام خدمات اختصاص یافته استفاده نکند و به صورت مجزا آنها را صورتحساب کند، ممکن است پولی بیشتر از خدماتی که واقعاً انجام داده دریافت کند. برای مثال، تست کامل خون شامل آزمایش‌های زیادی مانند اندازه‌گیری آنزیم‌ها و مواد معدنی مختلف است. زمانی که این آزمایش‌ها جداگانه صورتحساب شود، نرخ پرداخت ممکن است دو برابر شود. ارائه صورتحساب برای مواردی که تحت پوشش بیمه نیست به جای موارد تحت پوشش نیز یکی از 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. پزشکان اغلب قوانین بازپرداخت را دستکاری می‌کنند تا به بیمارانشان کمک کنند تا برای خدمات ضروری در طرح درمان، پوشش لازم را بگیرند[۴]. 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 سوء استفاده‌گرانه بالقوه را نشان دهد. گذشته از شرایط پیش نیاز، یک

عدم تطابق بین تشخیص و طرح درمان می‌تواند نشانه یک رفتار مشکوک باشد. برای مثال، تشخیصی که نیاز به داروی خاص برای بیمار ندارد ممکن است نشان دهنده تقلب یا سوء استفاده بالقوه باشد.

نسبت برخورد غیر معمول با بیمار، پارامتر دیگری برای تخمین ریسک تقلب و سوء استفاده است. برای مثال، اگر پزشکی هر روز تعداد زیادی از بیماران را ببیند که بیشتر از میزانی است که او 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 پزشکانی که استخدام کرده‌اند را بیشتر از تعداد واقعی گزارش می‌دهند، تقلب کرده‌اند، زیرا ارائه اطلاعات نادرست نیز تقلب است. ارجاع منفعت طلبانه، معرفی بیماران به پزشکی خاص یا ارائه‌دهنده خدمات بهداشتی و درمانی خاص است. برای مثال، اگر یک پزشک منفعتی شخصی از یک کلینیک داشته باشد، نمی‌تواند هیچ بیماری را به آن کلینیک ارجاع دهد. در بعضی از کشورها از جمله امریکا قانونی برای مقابله با این امر وجود دارد. در ایران اشتراک منافع پزشکان با داروخانه‌ها و آزمایشگاه‌ها و بیمارستان‌ها به تناسب قرارداد سازمان‌های بیمه‌گر ممکن است با جرایمی همراه باشد. به صورت 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.

۱-۳ طرح مسئله

کلاهبرداری در حوضه بهداشت و درمان (HCF) با احتساب ۹۸ بیلون دلار از هزینه‌های سالانه که به بیمه پزشکی سالمندان (Medicare) و بیمه بهداشت مستمندان (Medicaid) در ایالات متحده مصرف می‌شود، یک تخلیه چند بیلون دلاری در 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 بازرسی 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 اهداف و نوآوری‌ها

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

از ابزارهای قدرتمند برای تحلیل دادگان انبوه می‌توان به تکنیک‌های تحلیل گراف اشاره کرد. تحلیل گراف‌های بدست آمده از داده‌های خام می‌تواند، اطلاعات و الگوهای مفیدی را کشف کند. اطلاعات مفیدی که بسیاری از سازمان‌ها به راحتی قادر به کشف آنها نیستند. به طور کلی، روش‌های تحلیل شبکه ابزار قدرتمندی برای تحلیل داده‌های با حجم و پیچیدگی زیاد است. به این ترتیب، در این پژوهش سعی داریم با شناسایی الگوهای موجود در تجویز نسخ دارویی و کشف موارد مشکوک به تقلب، همزمان با حفظ و حتی بهبود خدمات، هزینه‌های سیستم بیمه سلامت را به صورت قابل توجهی کاهش دهیم.

۱-۵ محدودیت‌های تحقیق

اگرچه تشخیص سندسازی و تقلب در بیمه حیاتی و به شدت مورد نیاز است، چالش‌ها و محدودیت‌های زیادی هستند که این کار را سخت می‌کنند. اول، تشخیص تقلب و سوء استفاده از بیمه سلامت نیازمند کارشناسانی است که از دانش پزشکی در سطح بالایی برخوردار باشند [۶]. بیشتر شرکت‌های بیمه از روش‌هایی استفاده می‌کنند که برای تشخیص فعالیت‌های متقلبانه یا سوء استفاده‌گرانه بالقوه، نیازمند نیروی انسانی جهت ارزیابی مدارک است. این روش‌ها که مبتنی بر

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

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

- اکثر فرم‌ها و پرونده‌ها بصورت غیرسیستمی و دستی ذخیره شده‌اند و تعداد کمی از آن‌ها بصورت الکترونیکی ذخیره و نگه داری می‌شوند و جستجوی دستی در این اسناد بسیار زمان‌بر است.
- شرکت‌های بیمه اطلاعات و پرونده‌های مشتریان خود را به آسانی در اختیار افراد خارج از سازمان قرار نمی‌دهند و اخذ مجوز برای دسترسی به این اطلاعات فرآیندی زمان‌بر و مشکل است.
- دسترسی به تعداد اندکی پرونده که وقوع قلب در آن‌ها محرز شده است بسیار مشکل است زیرا این پرونده‌ها اغلب محرمانه هستند.
- کسب دانش مورد نیاز از افراد خبره و کارشناسان کشف تقلبات بیمه‌ای مشکل است زیرا اکثر آن‌ها مدیران بخش بیمه هستند و به ندرت وقت آزادی برای قبول و انجام مصاحبه دارند [۸].

۱-۶ ساختار پایان نامه

این پایان‌نامه در پنج فصل و قسمت مراجع نوشته شده است. در این فصل (فصل ۱) معرفی و اهداف این مسئله ارائه شد. در فصل دوم مفاهیم مرتبط با این تحقیق معرفی می‌شوند و پژوهش‌های مرتبط پیشین بررسی می‌شوند. در فصل سوم روشی که در این پژوهش پیشنهاد شده است ارائه می‌گردد. در فصل چهارم مستندات نتایج و تحلیل نتایج آورده شده

و نتایج بدست آمده ارزیابی و با سایر روش‌ها مقایسه می‌شوند و در فصل پنجم جمع‌بندی و پیشنهاد پژوهش‌های آتی ارائه می‌شود.

۱-۷ جمع بندی

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

فصل دوم

پیشینه پژوهش

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 تاثیر می‌گذارد. رفتارهای متقلبانیه ارائه‌دهندگان بهداشت و درمان و بیماران با تحمیل هزینه‌های غیرضروری به مشکلی جدی برای سیستم‌های بیمه تبدیل شده است. بنابراین 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.

با توجه به حجم اسناد و انبوه بیمه شدگان و ارائه‌دهندگان خدمت، کشف تقلب به صورت سنتی غیرممکن است. در نتیجه، روش‌های داده‌کاوی و تحلیل شبکه با شناسایی الگوهای موجود در تجویز نسخه‌دارویی و کشف موارد مشکوک به تقلب، همزمان با حفظ و حتی بهبود خدمات، هزینه‌های سیستم بیمه سلامت را به صورت قابل توجهی کاهش می‌دهد. در این فصل، به مفاهیم و پژوهش‌هایی که در زمینه‌ی کشف تقلب و سست‌هنجاری در سیستم بیمه سلامت انجام شده، پرداخته شده است.

۲-۲ 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 کرده است. 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. تقلب را این‌گونه توصیف می‌کند که منجر به سوءاستفاده از منافع یکساز ما نمی‌شود و نه لزوماً منجر به یک پیامد قانونی مستقیم! [۹]

اگرچه که یک تعریف پذیرفته شده جهانی از تقلب مالی وجود ندارد، [۱۰] 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.

۲-۳ نقش داده‌کاوی در تشخیص تقلب

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 ۱.۶ بیلیون پوند در سال برای بیمه‌گذاران انگلستان هزینه دارد. خسارات کلی ناشی از تقلب توسط بیمه غیرقابل محاسبه است. تشخیص تقلب بیمه برای جلوگیری از نتایج مخرب تقلب بیمه مهم است. تشخیص تقلب بیمه شامل بررسی مطالبه‌های جعلی از مطالبه‌های اصلی است. به این ترتیب، افشای رفتار یا فعالیت جعلی،

تصمیم‌گیرندگان را برای توسعه استراتژی‌های مناسب برای کاهش اثر تقلب قادر می‌سازد. داده‌کاوی یک نقش مهم در تشخیص تقلب بیمه دارد، همانطور که اغلب برای استخراج و پرده‌برداری از حقایق پنهان، مقدار زیادی داده 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 داده‌کاوی باید به خوبی قابل تعریف باشد و با ابزارهای پرس‌وجو قابل حل نباشد و با یک مدل پردازش داده‌کاوی تعلیم داده شود [۱۱]. داده‌کاوی را به این صورت تعریف می‌کنند: پروسه شناسایی الگوهای مورد علاقه در پایگاه داده بطوریکه بعداً بتوانند در تصمیم‌گیری استفاده شوند [۱۲].

داده‌کاوی را پروسه‌ای تعریف می‌کند که از ریاضیات آماری و هوش مصنوعی و تکنیک‌های یادگیری ماشین برای استخراج و شناسایی اطلاعات مفید استفاده می‌کند و بطور پیوسته از یک پایگاه‌داده بزرگ تجربه کسب می‌کند. [۱۳] بیان می‌کند که هدف داده‌کاوی به دست آوردن اطلاعات مفید و غیر آشکار از داده‌های ذخیره شده در مخازن بزرگ است. [۱۴] مشخص می‌کند که یکی از مزایای مهم داده‌کاوی این است که می‌تواند برای توسعه یک کلاس جدید از مدل‌ها برای شناسایی جملات، قبل از اینکه توسط متخصصان شان تشخیص داده شوند، استفاده شود. [۹] اشاره دارد که تشخیص تقلب یکی از بهترین کاربردهای داده‌کاوی در صنعت و دولت است. تکنیک‌های داده‌کاوی گوناگونی در تشخیص تقلب بیمه به کار گرفته شده‌اند مانند شبکه‌های عصبی، مدل‌های رگرسیون منطقی، روش‌های نایوبیز و درخت تصمیم.

۲-۴ 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 [۴].

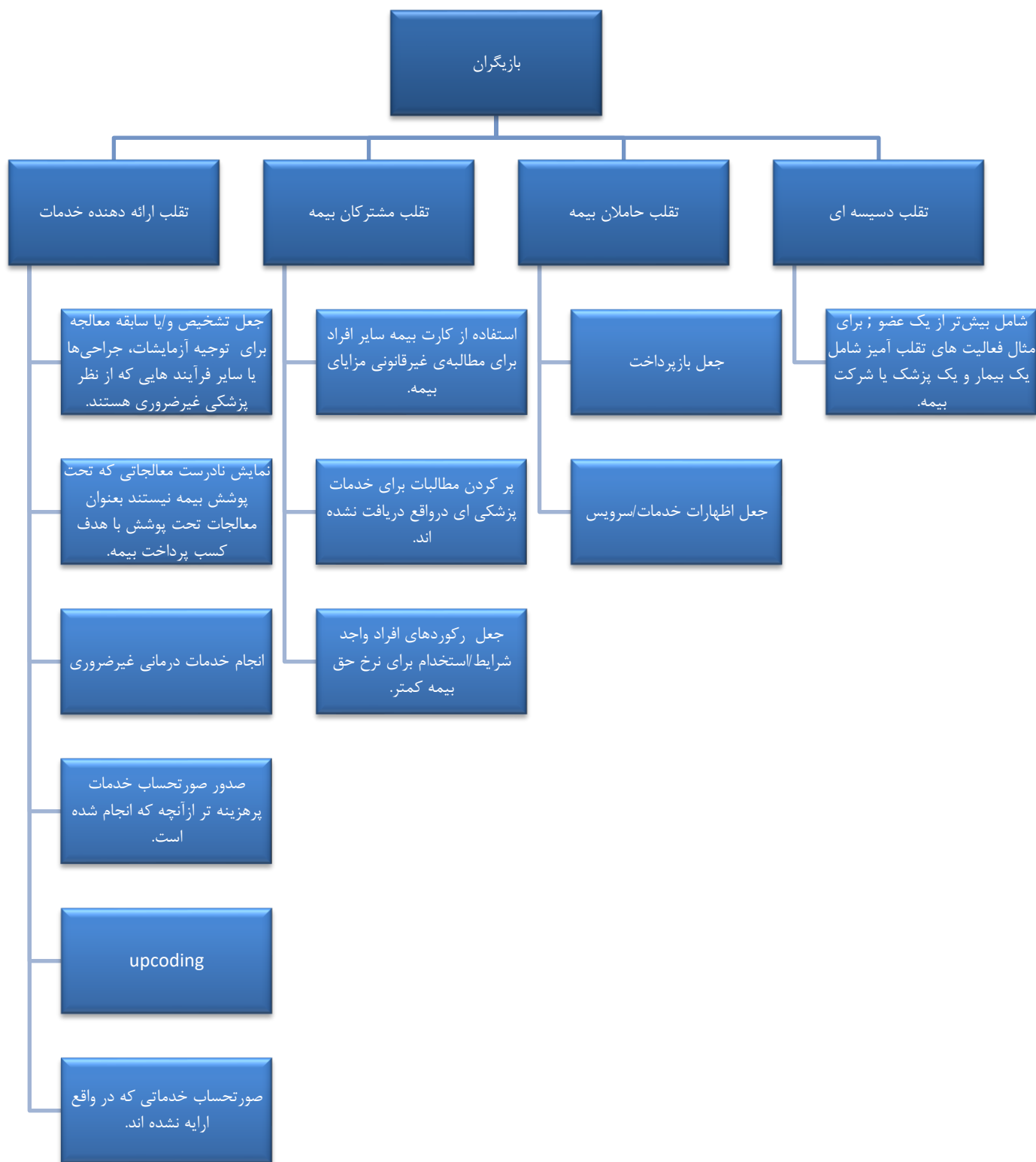
کلاهبرداری مراقبت‌های بهداشتی موضوعی خاص برای هر کشور است و رفتارهای کلاهبردارانه، متفاوتی به همین نسبت تغییر می‌کند. با این وجود، انواع کلاهبرداری‌هایی که در زمینه مراقبت‌های بهداشتی انجام می‌شود، تقریباً برای همه کشورها رخ می‌دهد. همانطور که در شکل (۱-۲) دیده می‌شود، چهار دسته اصلی در تقلب در مراقبت‌های 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 تقلب آمیز در ادامه دسته‌بندی و توضیح داده می‌شود. بر اساس اینکه چه کسی مرتکب تقلب می‌شود، رفتارهای تقلب آمیز به صورتی که در ادامه آمده‌اند دسته‌بندی می‌شوند [۱۲]:

^۱ Actor

۱. **تقلب ارائه‌دهنده‌ی خدمات^۱:** 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.
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.
۳. **تقلب حاملان بیمه:** شامل افرادی می‌شود که جعل در بازپرداخت یا جعل در اظهاراتشان نسبت به سود یا زیان‌دهی کرده‌اند.
۴. **تقلب دسیسه‌ای^۲:** شامل تخلفاتی می‌شود که در بیشتر از یکی از سه دسته دیگر درگیر آن هستند (برای مثال: یک دکتر و یک بیمار).

^۱ Service Provider's Fraud

^۲ conspiracy



شکل (۱-۲): درخت دسته‌بندی بازیگران نظام سلامت

۲-۵-۲ دسته بندی 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.

۲-۵-۲-۱ چالش‌های کشف تقلب از منظر سازمان بیمه

شرکت‌های بیمه می‌توانند با آگاهی از انواع تقلبات و فرآیندهایی که احتمال بروز تقلب در آن‌ها وجود دارد سیستم هشدار دهنده و پیشگیرانه‌ای را طراحی کنند و با آگاهی از میزان آسیب‌پذیری خود استراتژی‌های موثرتری را به کار گیرند، اما برای تحقق این امور شرکت‌های بیمه با محدودیت‌ها و پیچیدگی‌های زیادی مواجه‌اند [۱۵]:

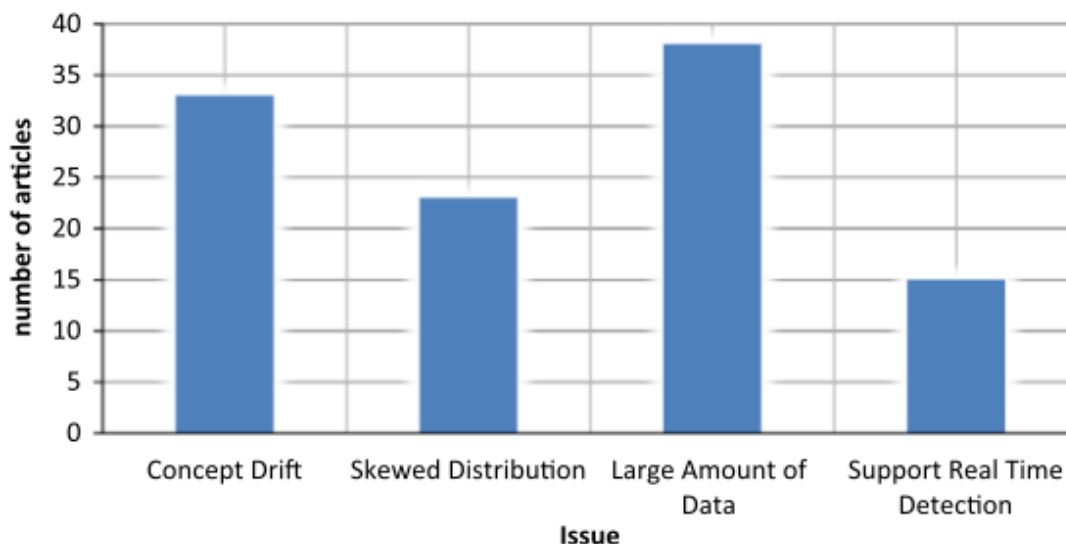
- پنهان بودن ماهیت تقلب
- پویایی و حساسیت به تغییر در تقلب (هنگام شناسایی یک سبک تقلب، کلاهبرداری با سبک دیگری در حال شکل‌گیری است)
- عدم توافق اجماع در برخی موارد بر آنچه که واقعاً به منزله‌ی تقلب در بیمه استاندارد.
- نگاه جامعه به شرکت‌های بیمه‌ای و ارائه‌دهندگان خدمات درمانی بعنوان نهادهای حمایتی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

۲-۵-۲-۲ چالش‌های کشف تقلب از منظر داده

از چالش‌های موجود در کشف تقلب عدم وجود داده‌های با برچسب سالم و وجود تعداد بسیار کم داده‌های تقلبی جهت یادگیری است.

۲-۵-۲-۳ چالش‌های کشف تقلب از منظر مقالات

شکل (۲-۲) توزیع مطالعات سیستم‌های تشخیص تقلب را بر اساس چالش‌ها بر اساس تعداد مقالات منتشر شده در سال‌های ۱۹۹۴ تا ۲۰۱۴ نشان می‌دهد و به رایج‌ترین انواع تقلب‌های الکترونیکی مانند کارت‌های اعتباری، بیمه مراقبت سلامت، مخابرات، بیمه اتومبیل متمرکز است [۱۶].



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

۲-۲ مفاهیم پایه

در این بخش قصد داریم به معرفی مفاهیم پایه در پژوهش های کشف تقلب در سیستم نظام سلامت بپردازیم.

۲-۲-۱ مفهوم رانش^۱

تعاریف مختلفی برای مفهوم مسئله رانش وجود دارد. در داده کاوی رانش به پدیده ای که مدل پایه ای آن در طول زمان در حال تغییر است اشاره دارد. کار سیستم های تشخیص تقلب در محیط پویا که رفتار کاربران قانونی/غیرقانونی بطور پیوسته در حال تغییر است مفهوم پدیده رانش گفته می شود [۱۷].

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 می تواند تغییر کند [۱۸].

به علاوه مفهوم رانش سابقاً به یک سناریوی یادگیری با ناظر زمانی که رابطه بین داده ورودی و متغیر هدف در طول زمان تغییر می کند، ارجاع داده می شد. اگرچه در یادگیری با ناظر هدف پیش بینی یک متغیر هدف y با استفاده از مجموعه ای از ویژگی های ورودی x است. در نمونه یادگیری که برای ساخت مدل استفاده می شود هر دو x و y در زمان پیشگویی ناشناخته اند و رابطه بین داده ورودی و متغیر هدف ممکن است تغییر کند [۱۹].

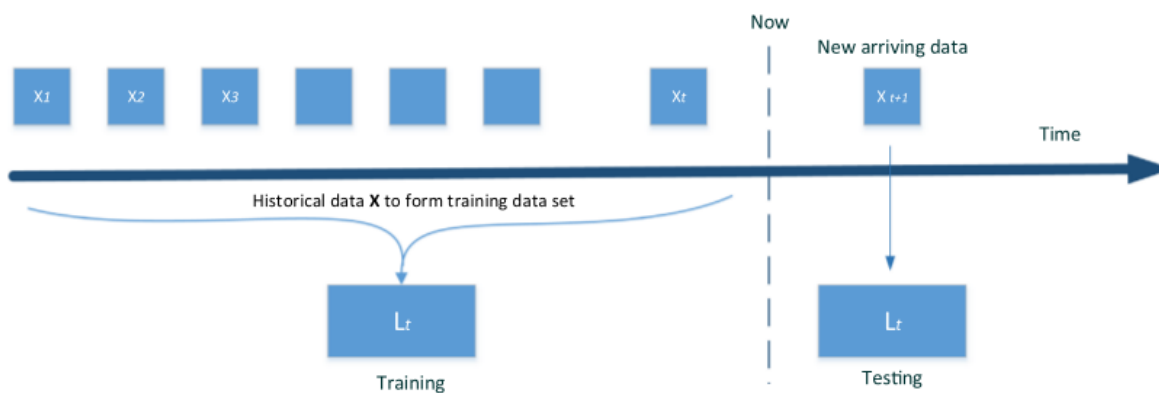
^۱ Drift

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^۱ 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 [۱۹]، [۲۰].

[۲۱] مفهوم یادگیری افزایشی با رانش را این‌گونه بیان می‌کند (رابطه (۱-۲)): فرآیند یادگیری افزایشی در هر زمان t که داده‌های قبلی در دسترس هستند، یک نمونه‌ی هدف x_{t+1} میرسد وظیفه‌اش این است که برچسب y_{t+1} را پیش‌بینی می‌کند. بدین منظور یادگیرنده L_t در فاز یادگیری با استفاده از همه یا انتخاب از داده‌های قبلی برچسب دار ساخته می‌شود.

$$X_{\text{historical}} = (x_1, x_2, \dots, x_t) \quad \text{رابطه (۱-۲)}$$

که این را در شکل (۲-۳) می‌بینیم:



شکل (۲-۳): فرآیند یادگیری افزایشی در هر زمان t

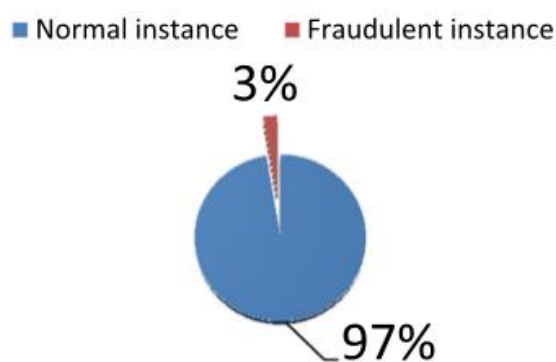
۲-۶-۲ مفهوم توزیع اریب کلاس‌ها^۲

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 [۲۲].

^۱ Adaptive

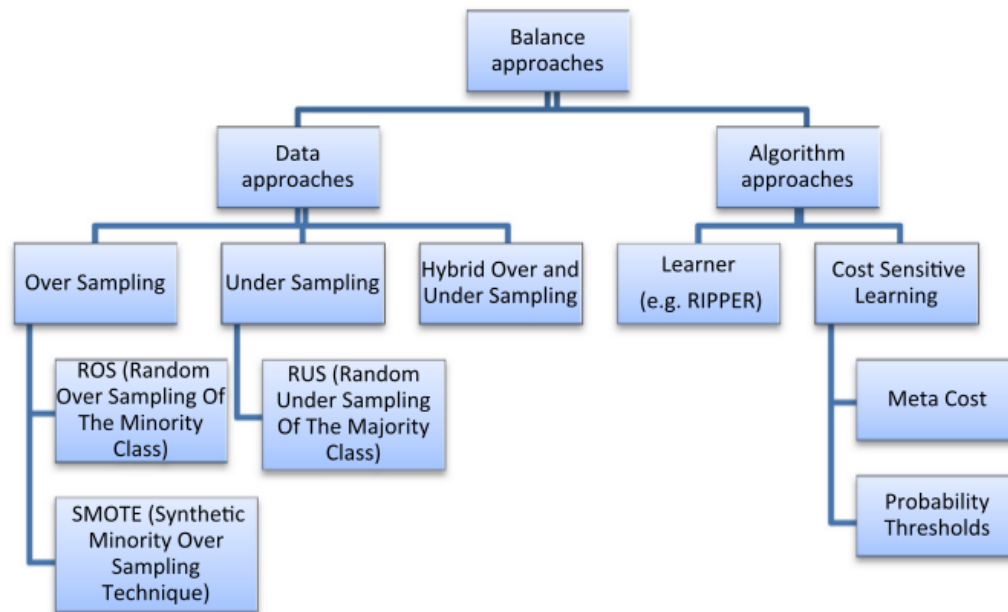
^۲ Skewed class distribution

داده‌ی رقابت‌های داده‌کاوی^۱ UCSD برای تشریح مساله نامتوازن بودن استفاده شده و یک مجموعه داده از دنیای واقعی است که برای تشخیص تراکنش‌های تجارت الکترونیک غیرعادی استفاده شده است. مجموعه داده‌ی آموزش شامل ۱۰۰۰۰۰ تراکنش از ۷۳۷۲۹ مشتری در طی ۹۸ روز است و مجموعه داده تست شامل ۵۰۰۰۰ تراکنش است. داده‌ی آموزشی به شدت نامتوازن است که شامل ۹۷۳۴۶ تراکنش طبیعی و فقط ۲۶۵۴ تراکنش تقلبی است. همانطور که در شکل (۲-۴) می‌بینیم درصد تراکنش‌ها در حدود ۳-۹۷٪ قانونی و جعلی می‌باشد. بنابراین یک مکانیزم متوازن‌سازی نیاز است تا این داده‌ها را با نرخ ۱:۱ میان طبیعی و جعلی متوازن کند. روش‌های متوازن‌سازی داده‌ها در دو سطح می‌تواند طبقه بندی شود: سطح داده و سطح الگوریتم، که تکنیک‌های آن در شکل (۲-۵) نشان داده شده است.



شکل (۲-۴) مجموعه داده آموزشی ناهمگون UCSD

^۱ University of California, San Diego



شکل (۵-۲) روش‌های رسیدگی به داده‌های ناهمگون

۲-۶-۳ مفهوم کاهش داده‌ها

مقیاس بزرگ و ابعاد زیاد مجموعه داده‌ی تقلب و حضور تعداد زیاد ویژگی‌ها/ورودی‌ها/متغیرها فرآیند داده‌کاوی و تشخیص را بسیار دشوار و پیچیده می‌سازد [۲۳]. بعلاوه این شرایط فرآیند تشخیص را نیز کند می‌کند. بنابراین سیستم‌های تشخیص تقلب موجود از روش‌های کاهش داده برای کاهش حجم مجموعه داده‌ها استفاده می‌کند [۲۴]. بعلاوه داده‌ی کم‌حجم مدل و در نتیجه زمان محاسبه را کاهش می‌دهد [۲۵]. روش‌های کاهش داده شامل کاهش ابعاد (Dimension Reduction) و کاهش عددی (Numerosity Reduction) است [۲۶].

کاهش ابعاد شامل استراتژی‌های بسیاری است به نام‌های فشرده‌سازی داده (data compression)، انتخاب داده (feature selection)، ساخت ویژگی (feature construction). مرسوم‌ترین و پرتکرارترین استراتژی‌های استفاده شده در سیستم‌های تشخیص تقلب هستند. استراتژی فشرده‌سازی داده، از طریق استفاده از تکنیک‌های فشرده‌سازی مانند [۲۷]، [۲۸] نمایش داده‌ی اصلی را فشرده می‌کند. در این میان انتخاب ویژگی یک استراتژی دیگر کاهش ابعاد است. مهمترین و مرتبطترین ویژگی‌های انتخاب می‌شوند تا در ساخت مدل استفاده شوند. انتخاب ویژگی توسط [۲۹] نام گذاری شد.

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

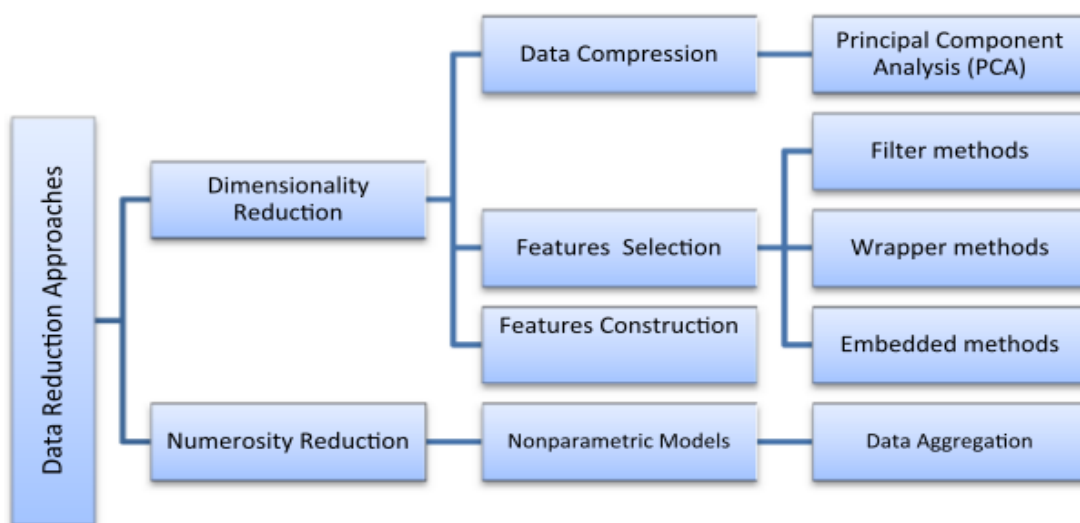
۱. روش‌های تصفیه ۲. روش‌های بسته‌بندی ۳. روش‌های جاسازی

✓ روش تصفیه بعنوان الگوریتم پیش پردازش برای رتبه بندی ویژگی‌ها عمل می‌کند که ویژگی‌های با رتبه بالا انتخاب و به یک پیش‌بینی کننده اعمال می‌شوند.

✓ روش جاسازی شامل انتخاب متغیر بعنوان بخشی از فرآیند یادگیری بدون تقسیم داده به مجموعه‌ی آموزش و تست می‌باشد [۳۰]. ساخت ویژگی در جایی است که یک مجموعه کوچک از ویژگی‌های مفیدتر از مجموعه‌ی اصلی مشتق می‌شود.

در کاهش عددی نیز داده‌ها با نمایش کوچکتری جایگزین می‌شوند. مانند استفاده از تجمیع داده‌ها [۲۲]، [۲۴].

روش‌های کاهش داده شامل کاهش ابعاد و کاهش ارقام در شکل (۶-۲) آورده شده‌اند:



شکل (۶-۲): روش‌های کاهش داده

۲-۶-۴ مفهوم پشتیبانی تشخیص آنلاین

سیستم‌های تشخیص تقلب در دو حالت متفاوت کار می‌کنند که تشخیص آنلاین و آفلاین است که بر اساس انواع تقلب متفاوت است و هریک کاربرد خاص خود را دارند [۱۶].

۲-۶-۵ ادغام در نظام مراقبت‌های بهداشتی

تعداد بسیار کمی از مقالات مورد بررسی تلاش برای ادغام فرآیند داده‌کاوی را در چارچوب تصمیم‌گیری واقعی انجام دادند. تأثیر کشف دانش توسط داده‌کاوی بر میزان کاری و زمان کار حرفه‌ای مراقبت‌های بهداشتی مشخص نیست. مطالعات آینده باید یکپارچگی سیستم توسعه‌یافته را در نظر بگیرند و تأثیر آن بر محیط کار را بررسی کنند [۳۱].

۲-۶-۶ خطای پیش‌بینی و اثر "Black Swan"

در مراقبت‌های بهداشتی، پیش‌بینی بهتر از پیش‌بینی اشتباه بهتر است [۳۲]. 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.

علاوه بر این، مدل پیش‌بینی ممکن است در پیش‌بینی رویدادهای عادی بهتر از موارد نادر باشد. محققان باید مدل‌های پیشرفته‌ای را برای رسیدگی به غیرقابل پیش‌بینی "The Black Swan" توسعه دهند [۳۳].

یک مطالعه [۳۴] یک مسئله مشابه را در توصیه‌های مبتنی بر شواهد برای تجویز پزشکان مطرح کرد. نگرانی آنها این بود که چه مقدار شواهد باید برای تهیه یک توصیه کافی باشد.

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

۲-۶-۷ از دست دادن اطلاعات در پیش‌پردازش

پیش‌پردازش داده‌ها، از جمله دستکاری داده‌های از دست رفته، پرهزینه‌ترین و مهم‌ترین بخش داده‌کاوی است. شایع‌ترین روش مورد استفاده در مقالات مورد بررسی حذف یا حذف داده‌های از دست رفته است. در یک مطالعه، حدود ۴۶.۵٪ از داده‌ها و ۳۶۳ از ۴۱۰ ویژگی به دلیل مقادیر گم شده حذف شدند [۳۵]. در یکی دیگر، محققان [۳۶] تنها قادر به استفاده از ۲۰۶۴ از ۴۹۴۸ مشاهدات (۴۲٪) بودند. با حذف مقادیر از دست رفته و داده‌های پرت، ما مقدار قابل توجهی از اطلاعات را از دست می‌دهیم. پژوهش‌های آینده باید بر روی یافتن یک روش بهتر تخمین مقادیر از دست‌رفته نسبت به حذف تمرکز نمایند. علاوه بر این، تکنیک‌های جمع‌آوری داده‌ها باید به منظور جلوگیری از این موضوع توسعه یابند یا اصلاح شوند.

۲-۶-۸ خودکارسازی فرآیند داده‌کاوی برای کاربران غیرمتخصص

کاربران نهایی داده‌کاوی در مراقبت‌های بهداشتی، پزشکان، پرستاران و متخصصین مراقبت‌های بهداشتی هستند که آموزش‌های محدودی در زمینه تحلیلی دارند. یک راه حل برای این مشکل این است که یک سیستم خودکار (یعنی بدون نظارت انسان) برای کاربران نهایی ایجاد شود [۳۷]. یک ساختار خودکار مبتنی بر ابر برای جلوگیری از خطاهای پزشکی نیز می‌تواند توسعه یابد [۳۸]؛ اما این کار چالش برانگیز خواهد بود زیرا در آن زمینه‌های کاربردی مختلف وجود دارد و یک الگوریتم دقت مشابهی برای تمام برنامه‌های کاربردی ندارد [۳۷].

۲-۶-۹ ماهیت بین رشته‌ای تحقیق و دانش متخصص حوزه

تجزیه و تحلیل بهداشت و درمان یک زمینه تحقیقاتی بین رشته‌ای است [۳۷]. به عنوان یک روش تجزیه و تحلیل، داده‌کاوی باید از ترکیبی از نظر کارشناس از حوزه‌های خاص مراقبت‌های بهداشتی و مشکل مشخص (به عنوان مثال، انکولوژی برای تحقیقات سرطان و متخصص قلب برای CVD) استفاده کند [۳۹]. تقریباً ۳۲٪ از مقالات در تجزیه و تحلیل از نظر متخصص به هیچ شکلی شکل استفاده نمی‌شود. پژوهش‌های آینده باید شامل اعضای از رشته‌های مختلف از جمله مراقبت‌های بهداشتی باشد [۳۱].

۲-۷ انواع ناهنجاری در تشخیص تقلب

ناهنجاری‌ها نمونه داده‌هایی هستند که به میزان قابل توجهی با سایر نمونه داده‌ها متفاوت و ناسازگار هستند [۴۰]. ناهنجاری‌ها همچنین پرت‌ها، اختلالات، مشاهدات غیرواقعی و استثنائات نیز نامیده می‌شوند [۴۱]. در تعریف دیگری، ناهنجاری را بعنوان مشاهده یا زیرمجموعه‌ای از مشاهدات میدانند که تا حدی زیادی از دیگر مشاهدات متفاوت است [۴۲]. منشأ ناهنجاری‌ها میتواند رفتار کلاهبردارانه، خطای انسانی یا شکست سامانه‌ها باشد [۴۳].

ناهنجاری‌ها از چند نظر قابل دسته‌بندی هستند. از نظر ماهیت ناهنجاری‌ها به ۴ دسته نقطه‌ای، جمعی، زمینه‌ای و افقی تقسیم میشوند [۴۴]. 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 [۴۵].

۲-۷-۱ تشخیص ناهنجاری گراف ایستا

ناهنجاری‌ها به یکی از دسته‌های ایستا ساده، ایستای با ویژگی، پویای ساده و پویای با ویژگی تقسیم می‌شوند [۴۵]. 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.

۲-۷-۱-۱ تشخیص ناهنجاری مبتنی بر ساختار

دو نوع روش تشخیص ناهنجاری مبتنی بر ساختار به نام‌های ناهنجاری در گراف‌های ساده^۱ ایستا^۱ و ناهنجاری در گراف ویژگی ایستا^۲ 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 [۴۶] 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 node (ego)* و *one-level neighborhood (an egonet)* و *average betweenness user's* و *betweenness centrality* و *two-level neighborhood (a super egonet)* برای شناسایی کاربران با ساختارهای ارتباطی ناهنجار استفاده کردند.

در طرح تشخیص ناهنجاری در گراف ویژگی ایستا در [۴۷] 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^2) 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

۱ static plain graphs
۲ static attributed graph
۳ Maximum Description Length

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

۲-۷-۱-۲ تشخیص ناهنجاری مبتنی بر اجتماع

در [۴۸] 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 [۴۹] یک روش خوشه بندی متمرکز و تشخیص ناهنجاری در گراف‌ها به نام FocusCo ارائه نمودند. الگوریتم شامل سه گام است ۱. استنتاج وزن‌های ویژگی‌ها ۲. استخراج خوشه‌های گراف‌های ویژگی متمرکز ایستا ۳. تشخیص ناهنجاری.

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

۲-۷-۲ 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 تجزیه، پویای مبتنی بر خوشه یا اجتماع، پویای مبتنی بر مدل‌های احتمالاتی و پویای مبتنی بر پنجره تقسیم می‌شوند.

۲-۷-۲-۱ 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^۱، بیشینه زیرگراف مشترک^۲، فاصله ماتریس همسایگی^۳، فاصله ویرایش گراف^۴، فاصله همینگ برای ماتریس‌های همسایگی گراف‌ها و ... [۵۰].

۲-۷-۲-۲ تشخیص ناهنجاری مبتنی بر فشرده سازی

در این فرآیند یک نمایش گراف فشرده با استفاده از حداقل طول توصیفی و روش فشرده‌سازی با بهره‌گیری از الگوها و تنظیمات داده‌ها با کمترین هزینه‌ی رمزگذاری به‌دست می‌آید. سپس ناهنجاری‌ها به‌عنوان گراف‌هایی که مانع فشرده‌سازی هستند تعریف می‌شوند [۵۰].

۲-۷-۲-۳ تشخیص ناهنجاری مبتنی بر تجزیه

این روش ناهنجاری‌های موقتی را با نمایش مجموعه‌ای از گراف‌های تکامل زمانی^۵ به‌عنوان یک تنسور یا آرایه‌ی چند بعدی تشخیص می‌دهد و factorization یا کاهش بعد انجام می‌دهد. یک روش جدید تجزیه ماتریس فشرده^۶ برای محاسبه تقریب‌های کم مرتبه خلوت در [۵۱] ارائه شده است. خطای بازسازی هرگراف خلوت در طول زمان پیگیری شده و در جایی که تغییر زیاد باشد، گراف منطبق ناهنجاری خواهد بود.

۲-۷-۲-۴ تشخیص ناهنجاری مبتنی بر خوشه یا اجتماع

در مورد روش مبتنی بر اجتماع یا خوشه، به جای زیر نظر گرفتن تغییرات در کل شبکه، یک اجتماع در هر زمان، برای هر حادثه‌ی غیرعادی ای زیر نظر گرفته می‌شود. در [۵۲] 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.

۲-۷-۲-۵ 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 [۵۳] ارائه شد که گام اول شامل مدل‌های بیزین مزدوج برای فرآیندهای شمارش زمان گسسته برای ردیابی جفت پیوندهای میان همه‌ی گره‌ها در گراف برای ارزیابی طبیعی بودن رفتار است. در گام دوم استنتاج شبکه استاندارد روی کاهش زیرمجموعه‌ی گره‌های بالقوه غیرعادی اعمال می‌شود.

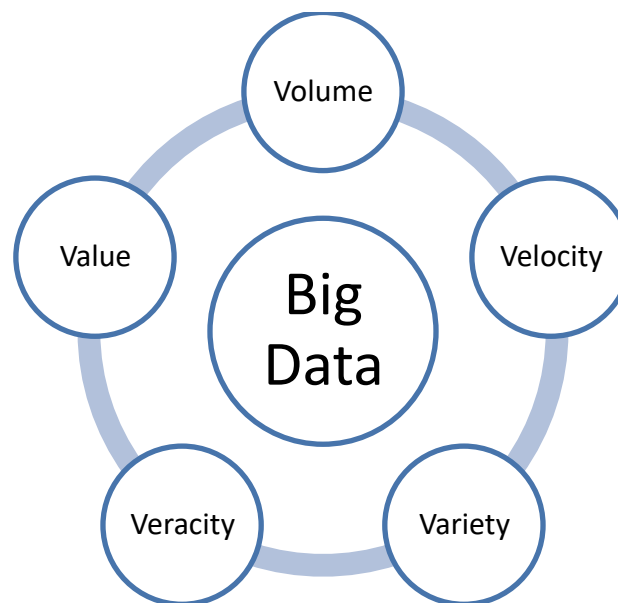
^۱ Error correcting graph matching distance
^۲ Maximum Common Sub graph (MCS)
^۳ distance of adjacency matrices
^۴ Graph Edit Distance (GED)
^۵ time evolving
^۶ Compact Matrix Decomposition (CMD)

۲-۷-۲ تشخیص ناهنجاری مبتنی بر پنجره

الگوریتم‌های تشخیص ناهنجاری برخی روش‌ها را ارائه می‌کنند که محدود به یک چارچوب زمانی است. در [۵۴] روشی ارائه دادند به نام یادگیری الگو و تشخیص ناهنجاری روی جریان‌ها^۱ 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.

۲-۸ تعریف Big Data و کاربرد آن در تقلب

تعریف Big Data در تحقیقات مورد توافق جهانی نیست. بنابراین، از تعریف فراگیر دیمنچکو و همکاران استفاده می‌کنیم. همانطور که در شکل (۲-۷) دیده می‌شود، داده‌های بزرگ را با پنج V تعریف می‌کنند: حجم (Volume) و سرعت (Velocity) و تنوع (Variety)، صحت (Veracity) و ارزش (Value).



شکل (۲-۷): پنج V داده‌های بزرگ

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

^۱ Pattern Learning and Anomaly Detection on Streams (PLADS)

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

هر ساله CMS داده‌های سال گذشته را منتشر می‌کند که حجم وسیعی از داده‌های موجود را افزایش می‌دهد. مجموعه داده‌ها حاوی حدود ۳۰ ویژگی است. هر کدام از ۳۰ ویژگی، از مشخصات جمعیت ارائه‌دهنده و انواع پروسه‌های درمانی، تا مبالغ پرداخت و تعداد خدمات انجام شده می‌باشد؛ بنابراین واجد شرایط Big Variety می‌باشند. علاوه بر این، مجموعه داده ترکیبی مورد استفاده در این تحقیق، ذاتاً شرط داده‌های Big Variety را فراهم می‌کند، زیرا سه منبع اصلی (اما متفاوت) Medicare را با هم ترکیب می‌کند. از آنجایی که CMS یک برنامه دولتی با کنترل کیفیت شفاف و مستندات دقیق است، برای هر مجموعه داده، این مجموعه داده‌ها قابل اعتماد، معتبر و نمایانگر کلیه مطالبات ارائه‌دهندگان شناخته 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.

۲-۹ برخی معیارهای تحلیل شبکه‌ی پزشکان

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

۱. درجه: اهمیت پزشک متناسب با تعداد بیماران مشترک با پزشکان دیگر است.
۲. مقدار خاص^۱: هرچه تعداد بیماران مشترک با سایر پزشکان مهم بیشتر باشد، پزشک مهم‌تر در نظر گرفته می‌شود. اگر پزشک تعداد زیادی از بیماران را به اشتراک بگذارد، اما با پزشکانی که در شبکه مهم نیستند، پزشک مهم در نظر گرفته نمی‌شود [۵۵].
۳. بینابینی^۲: با توجه به اینکه پزشکان می‌توانند برای دیگر پزشکانی که نزدیک‌ترند تأثیرگذار باشند، مثلاً زمانی که پزشک یک پزشک دیگر را به بیمار خود نشان می‌دهد، تعریف معیاری که این مجاورت را نشان دهد امکان‌پذیر است [۵۶].
۴. نزدیکی^۳: پزشکان می‌توانند توسط مقدار سایر پزشکان که در میان یکدیگر شناخته شده‌اند، غیرمستقیم و بدون اطلاع آن‌ها، مرتبط باشند. بنابراین، با توجه به تعداد پزشکان مورد نیاز برای اتصال دو پزشک، می‌توان میزان جدایی پزشکان را محاسبه کرد [۵۷].

۱ eigenvalue
۲ betweenness
۳ closeness

۲-۱۰ کلان داده‌ها^۱ در تشخیص تقلب

در داده‌های کلان، فقط موضوع حجم^۲ مطرح نیست و باید سایر موارد از قبیل تنوع^۳ داده‌ها و سرعت^۴، برقرار باشد تا بتوان به دنیای یادگیری عمیق وارد شد. در موضوع کشف تقلب حوزه درمان نه تنها با حجم انبوهی از داده‌های متنوع مواجه‌ایم، بلکه این داده‌ها و الگوهای مربوط روزبه روز در حال تغییرند.

در سال‌های اخیر علاقه رو به افزایش در کاوش داده‌های مراقبت سلامت برای تشخیص تقلب شکل گرفته است. سیستم‌ها برای پردازش مطالبات الکترونیک پیاده‌سازی شده‌اند تا بصورت اتوماتیک بازرسی و مرور از داده‌های مطالبات را انجام دهند. این سیستم‌ها برای تشخیص اعمال تقلبی، صورتحساب اشتباه، مطالبات تکراری و سرویس‌هایی که تحت پوشش درمانی نیستند، طراحی شده‌اند. قابلیت‌های تشخیص تقلب این سیستم‌ها معمولاً محدود است زیرا تشخیص بطور عمده مبتنی بر قوانین ساده از پیش تعریف‌شده توسط متخصصان امر است. برای رسیدن به تشخیص موثرتر، بسیاری از محققین روش‌های پیچیده‌تر مقابله با تقلب را توسعه داده‌اند که بر اساس داده کاوی، یادگیری ماشین و سایر روش‌های تحلیلی است. روش‌های جدید ارایه شده برخی مزیت‌های اصلی مانند یادگیری خودکار الگوهای تقلب از داده‌ها و تعیین احتمال تقلب برای هر مورد و شناسایی انواع جدید تقلب که قبلاً ثبت نشده‌اند را دارا هستند [۵۸]، [۲۲].

۲-۱۱ رویکردهای کلی روش‌های کشف تقلب

همانطور که در جدول (۱-۲) دیده میشود، رویکردهای کشف تقلب را میتوان به صورت کلی به سه دسته‌ی روش‌های آماری، روش‌های یادگیری ماشین با ناظر، روش‌های یادگیری بدون ناظر و روش‌های یادگیری ماشین ترکیبی تقسیم نمود. اگرچه روش‌های آماری میتواند عملکرد سریعی در شناسایی تقلب داشته باشد و عملکرد مطلوبی در شناسایی انواع جدید تقلب دارد، این روش‌ها ممکن است مطالبات قانونی را به عنوان جعلی شناسایی کند که نقطه ضعف بسیار مهمی در میان روش‌های موجود است. همچنین، این روش‌ها نیاز به تخصص و دانش زیادی در زمینه‌ی روش‌های آماری و تشخیص تقلب دارد.

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

۱ Big Data
۲ volume
۳ variety
۴ velocity

جدول (۱-۲): 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 برچسب دار • overfitting
روش‌های یادگیری ماشین بدون ناظر	<ul style="list-style-type: none"> • هزینه‌ی کم‌تر به دلیل عدم نیاز به داده‌ی برچسب‌دار 	<ul style="list-style-type: none"> • تغییر مداوم وزن ها و پارامترها توسط متخصصان برای شناسایی انواع جدید تقلب • تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب • هزینه‌ی بالای داده‌ی برچسب‌دار
روش‌های یادگیری ماشین ترکیبی	<ul style="list-style-type: none"> • در مقایسه با روش‌های کاوش فرایند دارای نرخ کشف کاذب^۱ کمتر است. • کاهش هزینه‌ی برچسب داده‌ها 	<ul style="list-style-type: none"> • پوشش کم موارد تشخیص تقلب

۲-۱۱-۱ روش‌های با ناظر

در [۵۹] مقایسه‌ای از درخت‌های تصمیم، شبکه عصبی و رگرسیون منطقی به ترتیب نرخ صحیح شناسایی تقلب پزشکی آن‌ها ارائه داده شده‌است. شبکه‌های عصبی می‌توانند به مجموعه داده‌های پیچیده و بزرگ و ارتباطات غیرخطی متغیرها رسیدگی کند. اگرچه کاربرد شبکه عصبی بطور کلی نیازمند متخصصان آماری برای مثلاً تنظیم پارامترها می‌باشد. بعلاوه [۶۰] بر این نکته اشاره دارد که کلاس‌بندی ممکن است کارایی کم و بیش از حد وابستگی^۲ به مجموعه داده‌ی اریب داشته باشد. برای

^۱ false discovery rate

^۲ Overfitting

جلوگیری از مشکل بیش از حد وابستگی به مجموعه داده‌ها [۶۱] یک روش توقف اولیه^۱ در تشخیص تقلب پزشکی مبتنی بر شبکه عصبی‌شان پیاده‌سازی نمودند، که بر اساس استفاده از یک مجموعه آموزشی برای به‌روزرسانی وزن‌ها و بایاس‌ها و یک مجموعه آموزشی دیگر برای توقف آموزش زمانی که شبکه شروع به بیش از حد وابستگی به داده‌ها می‌کند، می‌باشد. آن‌ها همچنین به حل مساله‌ی اختلاف زیاد پیش‌بینی به دلیل سایز کم نمونه‌های با تعداد زیاد ویژگی پرداختند.

در مقایسه، درخت‌های تصمیم قوانین کلی‌ای دارند که تفسیر راحتی دارند. درخت تصمیم می‌تواند داده‌های خلوت را نیز بررسی کند اما ممکن است با افزایش سایز داده منجر به وابستگی بیش از حد به داده و کاهش قابلیت تفسیر نتایج شود. برای مثال [۷] یک مدل امتیازدهی برای احتمال سوءاستفاده ارائه داد و سپس ارائه‌دهندگان را با استفاده از درخت تصمیم، کلاس‌بندی نمود. [۶۲] یک شبکه بیزین پویا از شاخص‌های تقلب ارائه داد که وزن‌ها توسط قدرت پیش‌گویی تقلب هر ویژگی تعیین می‌گردد. طبقه‌بندهای بیزین زمان آموزش کوتاهتری دارند و در بررسی داده‌های خلوت مؤثرتر دیده شده‌اند. [۶۳] از الگوریتم k نزدیکترین همسایه برای طبقه‌بندی نمایه‌ی پزشکان استفاده نموده است. [۶۴] یک روش مبتنی بر بردار ماشین پشتیبان استفاده نموده است. تعدادی از تلاش‌های تشخیص پزشکی شامل ترکیب روش‌های با ناظر می‌باشد.

به‌طور کلی روش‌های باناظر برای تشخیص الگوهای قبلاً کشف شده‌ی تقلب مناسب است. به‌دلیل اینکه این روش‌ها براساس طبقه‌بندی مطالبات قبلی هستند، یک نفر باید از مساله‌ی تخمین بالقوه بیش از حد^۲ آگاه باشد [۵۹].

وجود کلاس‌های با سایز نامتعادل در داده‌های مطالبات نیز می‌تواند منجر به وابستگی بیش از حد به داده‌ها گردد. این مدل‌ها برای مقابله با الگوهای تقلب جدید و تغییرات در قوانین باید به‌طور مرتب به‌روز شوند. عدم توانایی روش‌های نظارت شده برای تشخیص تقلب‌های پویا و انطباق، توجه به روش‌های نامنظم را افزایش داده است [۶۵].

۲-۱۱-۲ روش‌های بدون ناظر

خوشه‌بندی اولین بار روی داده‌های پزشکی برای بخش‌بندی درمان پزشکان عمومی توسط [۶۶] اعمال شد. در [۶۷] و [۲۲] از داده‌های جغرافیایی در یک رویکرد مبتنی بر خوشه‌بندی استفاده نمودند. الگوریتم گروه‌خوشه‌ای بیزین برنولی^۳ [۶۸] با تمرکز بر وقوع ویزیت میان ارائه‌دهندگان و ذینفعان، داده‌های دوتایی را مدل می‌کند. این به‌طور بالقوه می‌تواند یک نوع تقلب در حال ظهور به نام "تقلب توطئه" را نشان دهد که شامل ویژگی‌های بیش از یک عضو از سیستم پزشکی است. این الگوریتم های خوشه‌بندی به بازرسان کمک می‌کند که صورتحساب و متغیر مطلوبشان را گروه‌بندی کنند. [۶۹] یک مرور کلی از روش‌های تشخیص داده‌پرتدر برخی از آزمایش‌ها برای ارزیابی اثربخشی آن ارائه می‌کند. این روش‌های تجزیه تحلیل شامل مدل‌های خطی، طرح جعبه‌ای^۴، تحلیل قله^۵، خوشه‌بندی چند متغیره و ارزیابی متخصص می‌باشد. [۷۰] یک روش تشخیص

^۱ Early stopping

^۲ potential overestimation issues

^۳ Bayesian Bernoulli co-clustering

^۴ boxplots

^۵ peak analysis

داده پرت مبتنی بر چگالی محلی برای شناسایی الگوهای پرداخت نامناسب در سیستم پزشکی استرالیا ارائه می‌دهد. [۷۱]

یک رویکرد یکپارچه که ترکیبی از انتخاب ویژگی، خوشه‌بندی، تشخیص الگو و تشخیص بیرونی است برای شناسایی تقلب در سیستم پزشکی استرالیا ارائه نمودند. [۷۲] یک روش تشخیص ناهنجاری دو مرحله‌ای برای شناسایی بیمارستان‌های جعلی در سیستم مراقبت بهداشت عمومی برزیل ارائه می‌کند. همچنین شامل مطالعات تشخیص داده پرت با داده‌های تجویزی است. [۷۳] یک مدل رفتاری پایه نرمال را برای شناسایی ناهنجاری‌ها برای شناسایی ناهنجاری‌های مربوط به هر نسخه ایجاد می‌کند. [۷۴] یک مدل تشخیص داده پرت مبتنی بر استنتاج بیزی است که با استفاده از توزیع احتمالات و فواصل قابل قبول برای ارزیابی ارجاعات ارائه می‌دهد. [۷۵] استفاده از یک تابع غلظت^۱ را به عنوان یک ابزار تشخیص پیش نمایش داده پرت برای کمک به ارزیابی تقلب پزشکی ارائه می‌دهد. علاوه بر این، ابزارهای صنعتی مبتنی بر تجزیه و تحلیل گراف، تجزیه و تحلیل ارتباطات و انجمن‌ها، ممکن است به بازرسان کمک کند تا روابط، پیوندها و الگوهای پنهان به اشتراک گذاری اطلاعات و تعاملات در گروه‌های بالقوه جعلی ارائه دهندگان و بیماران را آشکار سازند. تعداد و کیفیت ارتباط بین مشاغل را می‌توان با استفاده از شباهت در اطلاعات ارتباطی آن‌ها، مکان، ارائه دهندگان خدمات، دارایی‌ها و وابستگی‌ها تجزیه و تحلیل نمود. ارتباطات بالقوه با بازیکنان درگیر در تقلب ممکن است پرچم‌های قرمز را به ارمغان بیاورند و منجر به تحقیقات آتی گردند. این به طور خاص می‌تواند برای آشکارسازی شبکه‌های سازمان یافته، پیچیده و هماهنگ ارائه دهندگان و بیماران مفید باشد. رویکردهای بدون ناظر به طور کلی مورد استفاده قرار می‌گیرد تا قبل از اینکه متخصصان حوزه را به تحقیق بفرستند فعالیت‌های جعلی را به طور بالقوه برچسب بزنند. بنابراین، یک همکاری نزدیک بین پزشکان، آمارگیران و افرادی که در تصمیم‌گیری شرکت دارند، در مراحل تعیین و تنظیم مدل و تجزیه و تحلیل و تفسیر نتایج سودمند خواهد بود [۶۵].

الگوریتم Apriori یکی دیگر از تکنیک‌هایی است که در تشخیص تقلب استفاده می‌گردد. این الگوریتم (Agrawal و همکاران ۱۹۹۳)، مهم‌ترین الگوریتم کلاسیک برای کاوش اقلام مکرر است. Apriori برای یافتن همه اقلام مکرر در پایگاه داده داده شده DB استفاده می‌شود. بر اساس اصل Apriori هر زیر مجموعه‌ای از اقلام مکرر نیز باید مکرر باشد. به عنوان مثال: اگر XY مجموعه اقلام مکرر است، هر دو A و B باید مجموعه‌های 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^۲ (جستجوی سطح هوشمندانه) شناخته می‌شود که در آن k-آیتم برای کشف (k + ۱) آیتم بکار می‌روند. در ابتدا، مجموعه اقلام ۱-تکراری یافت می‌شود که آستانه پشتیبانی را برآورده می‌کند، توسط L^۱ 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 مجموعه‌های 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، L^۱ برای پیدا کردن L^۲ استفاده می‌شود، مجموعه‌ای از

^۱ concentration function

^۲ breadth-first search

مجموعه‌های مکرر ۲-آیتم که برای پیدا کردن L^3 و غیره استفاده می‌شود، تا زمانی که هیچ مجموعه مکرر k -آیتمی موجود نباشد [۷۶].

روش های مختلفی برای بهبود کارایی الگوریتم Apriori مانند جدول کاهش تراکنش، تقسیم بندی و ... استفاده می‌شود [۷۷]، [۷۸]. در [۷۹]، نویسندگان روشی را برای تفسیر ویژگی‌هایی که مقادیر پیوسته دارند با استفاده از فاصله مساوی عرض باند داخلی^۱ ارائه نمودند که براساس نظر متخصصین پزشکی انتخاب شده است. یک تحقیق دیگر [۸۰]، صورتحساب پزشکی را با استفاده از الگوریتم Apriori تحلیل می‌کند، برخی از اصلاحات را در الگوریتم Apriori موجود پیشنهاد دادند و سپس از اثربخشی آن در اطلاعات مفید ساخته شده در صورتحساب پزشکی استفاده کردند. همچنین از الگوریتم Apriori برای کشف بیماری های مکرر در اطلاعات پزشکی استفاده می‌کند. در [۸۱] روشی برای تشخیص وقوع بیماری با استفاده از الگوریتم Apriori در نقاط خاص جغرافیایی در دوره زمانی خاص ارائه شده است.

بولتن و هاند [۸۲] در سال ۲۰۰۱، PGA^۲ 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 [۸۳]. برای کشف اینکه آیا یک ارائه‌دهنده یک رفتار مطالبه‌ای مشکوک دارد یا نه باید با سایر متخصصان زمینه مشابه مقایسه شود. PGA ابزاری محبوب است برای فهم اینکه چگونه رفتار یک پزشک خاص با رفتار سایر پزشکان در یک گروه خاص مرتبط می‌شود. یک جنبه مهم از این تحلیل این است که چگونه ارائه‌دهندگان با یکدیگر گروه می‌شوند و چرا با هم گروه می‌شوند. به‌طور خاص در صنعت پزشکی متخصصان فوق تخصص بسیاری وجود دارد و گروه‌بندی تقریبی برخی از آن‌ها باهم کار ساده‌ای نیست. بنابراین استفاده از متخصصان برای تشکیل گروه برای اعمال تحلیل PGA نیاز است. فرض بر این است که گروه‌بندی پزشکان ممکن است و سپس مقایسه رفتار می‌تواند آغاز گردد. به‌عنوان مثال ۲۰ دندانپزشک در کدپستی خاصی باهم گروه بندی شده‌اند و توزیع معالجات بررسی شده است، نتیجه میانگین گروهی تعداد دفعات معالجات ارائه شده در هر گروه از بیماران است. خلاصه‌سازی‌های میانگین‌های گروهی در جدول (۲-۲) نشان داده شده‌اند و این مثالی است از اینکه چگونه PGA می‌تواند بکار رود. حال نمایه‌های یک دندانپزشک خاص می‌تواند با میانگین گروه مقایسه شود. اگر دندانپزشکی در طول یک دوره زمانی خاص به‌طور غیرعادی تعداد بالایی کانال ریشه انجام داده باشد، به این معنی نیست که مرتکب تقلب شده است و می‌تواند بر تحقیقات بیشتر روی آن دندانپزشک دلالت داشته باشد. از آنجایی که کانال ریشه یک جراحی دهانی در نظر گرفته می‌شود نرخ بازپرداخت آن به‌طور قابل ملاحظه ای بالاتر از سایر معالجات است و بنابراین محتمل است که دندانپزشک با قصد منحرف از این معالجه برای صورتحساب بیشتر استفاده کرده باشد.

^۱ width binning interval

^۲ Peer Group Analysis

جدول (۲-۲): مثالی از میانگین گروهی برای PGA

Treatment	Average per month	Reimbursement Rate
Cavity treatment	۱۵۰	\$ ۵۰
Pulling teeth	۱۵	\$ ۳۰۰
Root canal	۳	\$ ۱۲۰۰

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 این‌گونه 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 [۸۴].

در سال ۲۰۱۶، در مقاله [۸۵] 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 [۸۶] 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 [۸۷] 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 کلاهدرداری 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 [۸۸]، 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. [۸۸] یک الگوریتم بدون ناظر مبتنی بر فاصله برای ارزیابی خطر تقلب نسخه‌ها ارائه نموده است. در مقاله [۸۸] 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 روزرسانی‌های منظم از مجموعه داده‌های یکپارچه است. همچنین ترکیب چنین سیستمی در مراجع بهداشتی، سازمان‌های تأمین اجتماعی و شرکت‌های بیمه می‌تواند کارآیی را برای اطمینان از رعایت قانون بهبود بخشد، و هزینه‌های حسابرسی متخصص انسانی را بطور چشمگیری کاهش دهد. البته در این روش پزشکان بر اساس تخصصشان تقسیم‌بندی نشده‌اند و از مجموعه داده‌ی واقعی در این تحقیق استفاده نشده است. در مطالعه‌ی [۸۹]، 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 [۹۰]، مشکل شناسایی تقلب در سیستم‌های ارتباطی، به ویژه موارد کلاهدرداری، 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.

۲-۱۱-۳ روش‌های یادگیری ماشین ترکیبی

در [۹۱] 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 [۹۱]. در مطالعه‌ی [۹۲]، 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 (۳-۲) هریک از رویکردها و روش‌هایی که پیش‌تر بحث شد به تفکیک مزایا و معایب بیان شده است:

جدول (۳-۲): 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 [۸۵]	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 Outlier شامل انحراف	<ul style="list-style-type: none"> تست روی مجموعه داده واقعی استخراج ویژگی‌ها و ارزیابی با استفاده از انجام مصاحبه با متخصصان و پزشکان تفسیر نتایج توسط تحلیل گران 	<ul style="list-style-type: none"> اعتبارسنجی دشوار اثربخشی تکنیک تشخیص outlier نیاز به همکاری متخصصان امر برای طراحی معیارها و به‌خصوص تفسیر نتایج دارد 	<ul style="list-style-type: none"> تعداد مطالبات بازپرداخت هر ذینفع مقدار مطالبات بازپرداخت هر ذینفع مقدار مطالبات در ایام تعطیل میانگین تعداد مطالبات بازپرداختی هر ذینفع

		از مدل خطی، انحراف خوشه، انحراف از خوشه تکی، انحراف گرایشی، حداکثر انحراف		<ul style="list-style-type: none"> • تکنولوژی outlier هنوز در مرحله آزمایشگاهی است و خود را در عمل و در اجرای طولانی ثابت نکرده 	<ul style="list-style-type: none"> • میانگین مقدار • مطالبات بازپرداختی • هر ذینفع • کد نسخه • کد دندان • هزینه نسخه • میانگین تعداد نسخه‌های بازپرداختی هر مطالبه
۲	A Novel Page Rank-Based Algorithm to Identify Anomalies [۸۶]	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	<ul style="list-style-type: none"> • Page rank که سابقاً در حوزه تحلیل شبکه به کار رفته بود در حوزه تقلب در مطالبات بیمه نیز مؤثر عمل می‌کند 	<ul style="list-style-type: none"> • وقوع برخی FPها به دلیل اشتراک مشخصه‌های کلی در CPT (نسخه‌ی) پزشکان با تخصص‌های مختلف 	<ul style="list-style-type: none"> • Medical procedure code • national provider identifier • specialty • تعداد procedureها در هر سال
۳	Medical Insurance Fraud Recognition Based on Improved Outlier Detection Algorithm [۸۷]	استفاده از یک الگوریتم تشخیص Outlier بهبود یافته بر اساس خوشه‌بندی k-means	<ul style="list-style-type: none"> • استفاده از مجموعه داده واقعی • کاهش زمان اجرای الگوریتم با یافتن مقدار بهینه k با پیچیدگی زمانی از مرتبه $O(I * k * m * (no))$ و پیچیدگی مکانی از مرتبه $O((n-o)+k) * m$ که نسبت به الگوریتم CBLOF بهبود داشت. 	<ul style="list-style-type: none"> • ویژگی‌های استفاده شده فقط مربوط به بازپرداخت بیمه سلامت هستند و شامل اطلاعات بیمار و نسخه نمی‌باشند 	<ul style="list-style-type: none"> • بازپرداخت مربوط به برونشیت مزمن • بازپرداخت مربوط به بیماری‌های قلبی-ریوی • بازپرداخت مربوط به ذات الریه
۴	Community Detection Algorithm to Find Suspicious Group of Provider Community [۸۸]	ساخت ماتریس ارتباط میان دو پزشک و شناسایی پزشکانی که در یک شبکه بهم متصلند و با سایر پزشکان ارتباطی ندارند، بعنوان شبکه تقلب آمیز، و همچنین استفاده از اپراتورهای DB به جای loop بیش از دو پزشک	<ul style="list-style-type: none"> • تخصیص احتمال (Likelihood) / احتمال تشکیل شبکه انحصاری به هر پزشک • سرعت تشخیص بالا 	<ul style="list-style-type: none"> • استفاده از داده‌های ساختگی • استفاده از برخی پکیج‌های نرم افزاری برای محاسبه ماتریس روابط، که برای مجموعه داده‌های بزرگ بهینه نیست 	<ul style="list-style-type: none"> • procedure code • صورتحساب هر procedure code • health claim data • تاریخ ارائه خدمات • کد پزشک • تعداد ویزیت‌های هر بیمار به ازای هر پزشک

					<ul style="list-style-type: none"> عدم تفکیک پزشکان طبق تخصص آن‌ها post payment
۵	Improving Fraud and Abuse Detection in General Physician Claims: A Data Mining Study [۸۹]	<p>انجام عمل خوشه بندی بر اساس hierarchical clustering method و محاسبه تعداد بهینه خوشه ها بر اساس معیار فاصله Euclidian distance measures با استفاده از شاخص اعتباری بیشینه مقدار ضریب همبستگی سیلویت</p>	<ul style="list-style-type: none"> انجام تحقیق روی پزشکان هر دو بخش عمومی و خصوصی استفاده از مجموعه داده واقعی 	<ul style="list-style-type: none"> حذف داده‌های ناشناس عدم استفاده از روش‌های آماری برای پرکردن داده‌های از دست رفته 	<ul style="list-style-type: none"> درصد بیمارانی که بیش از یکبار در ماه ویزیت شده‌اند میانگین اقلام دارو در یک نسخه میانگین هزینه نسخه دارویی پزشک تعداد نسخ تزریقی/حاوی آنتی بیوتیک تعداد نسخ تزریقی/حاوی آنتی بیوتیکس
۶	A Social Network Analysis Framework for Modeling Health Insurance Claims Data [۹۰]	<p>به‌کارگیری تکنیک‌های برای تحلیل مطالبات بیمه سلامت از طریق نگاشت پزشکان با استفاده از بیماران مشترک بعنوان یک پروکسی برای ارتباط میان آن‌ها</p>	<ul style="list-style-type: none"> در نظر گرفتن رابطه میان پزشکان، پزشک و بیمار، پزشک و ارائه دهنده خدمات استفاده از مجموعه داده واقعی بهبود درک اهمیت ویژگی‌های میان افراد ارزیابی مدل توسط تحلیلگران فرآیند و پزشکان و متخصصان 	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> ID بیمار نوع نسخه اطلاعات زمانی مرتبط به حادثه ID مشاغل درگیر در فرآیند هزینه نسخه تخصص پزشک
۷	An Interactive Machine Learning Based Electronic Fraud and Abuse Detection System in HealthCare Insurance [۹۲]	<p>تولید سناریو توسط متخصصان و پزشکان برای رفتارهای غیرطبیعی و سپس وزن‌دهی actorها با روش وزن‌دهی binary pairwise comparison و محاسبه امتیاز خطای actorها و مطالبات، استفاده از ابزار visualization توسعه یافته تحت</p>	<ul style="list-style-type: none"> سازگار در یک محیط پویا قابل استفاده برای تحلیل , proactive reactive کاهش زمان تحلیل نتایج خروجی توسط کاربران به دلیل استفاده از ابزار visualization استفاده از بیش از یک actor 	<ul style="list-style-type: none"> تغییر مداوم وزن‌ها توسط متخصصان برای شناسایی انواع جدید تقلب تغییر ویژگی‌های فیلد ورودی توسط متخصصان برای شناسایی انواع جدید تقلب 	<ul style="list-style-type: none"> نسبت تعداد نسخه‌ها به تعداد مشخصی از بیمه شدگان نسبت تعداد نسخه‌ها به تعداد مشخصی از پزشکان تعداد کل نسخه‌ها

			QlikView که برای تحلیل بکار می‌رود		
<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> عدم آزمایش روش پیشنهادی با تعداد متخصصان بیشتر 	<ul style="list-style-type: none"> دقت بالا در مقایسه با روش دقت بالا روی همه ۴ تخصص 	<p>محاسبه یک معیار ریسک بر اساس فاصله مهالنویس و چگالی‌ها و محاسبه ریسک و ساخت درخت تصمیم آن</p>	Multi Stage Method to Detect Provider and Patient Fraud [۹۳]	۸

۲-۱۱-۴ جمع‌بندی

موارد زیر نکاتی هستند که در پژوهش‌های انجام شده دیده می‌شوند [۳۹]:

- 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^۱ صورت نگرفته، اگرچه که چنین تحقیقاتی می‌تواند برای عواقب سخت تقلب دسیسه‌گران 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 [۳۹] 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.
 - بیشتر مقالات موجود هنوز روش‌های سنتی داده‌کاوی را پوشش می‌دهند و حمایت زیادی برای تحلیل کلان داده وجود ندارد.
 - کمبود راه‌حل‌های مفهومی کلان داده و کاربرد آن با ابزارها، کتابخانه‌ها و بسترهای کلان داده‌ی جایگزین در تشخیص تقلب مراقبت سلامت دیده می‌شود.
 - ترکیباتی از انواع داده در زمینه‌های مختلف باید بررسی شوند تا روش‌های مؤثر بدست بیایند.

^۱ conspiratorial

- [١] M. E. Johnson and N. Nagarur, "Multi-stage methodology to detect health insurance claim fraud," *Health care management science*, vol. ١٩, no. ٣, pp. ٢٤٩-٢٦٠, ٢٠١٦.
- [٢] 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, ٢٠١٦.
- [٣] E. A. Duman and Ş. Sağiroğlu, "Heath care fraud detection methods and new approaches," in *2017 International Conference on Computer Science and Engineering (UBMK)*, ٢٠١٧, pp. ٨٣٩-٨٤٤: IEEE.
- [٤] 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. ٢, ٨٣no. ١٤, pp. ١٨٥٨-١٨٦٥, ٢٠٠٠.
- [٥] 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)*, ٢٠١٦, pp. ٨٤٥-٨٥١: IEEE.
- [٦] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Systems with Applications*, vol. ٣١, no. ١, pp. ٥٦-٦٨, ٢٠٠٦.
- [٧] 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. ٣٩, no. ٨, pp. ٧٤٤١-٧٤٥٠, ٢٠١٢.
- [٨] A. Garzelli, L. Capobianco, and F. Nencini, "Fusion of multispectral and panchromatic images as an optimisation problem," *Image Fusion*, p. ٢٢٣, ٢٠٠٨.
- [٩] C. Phua, V. Lee, K. Smith, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *arXiv preprint arXiv:1009.6119*, ٢٠١٠.
- [١٠] 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*, ٢٠٠٦, vol. ٢, pp. ١١٢٠-١١٢٤: IEEE.
- [١١] I. Bose and R. K. Mahapatra, "Business data mining—a machine learning perspective ",*Information & management*, vol. ٣٩, no. ٣, pp. ٢١١-٢٢٥, ٢٠٠١.
- [١٢] E. Turban, R. Sharda, and D. Delen, "Decision support and business intelligence systems (required)," *Google Scholar*, ٢٠١٠.
- [١٣] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," *AI magazine*, vol. ١٣, no. ٣, pp. ٥٧-٥٧, ١٩٩٢.
- [١٤] 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*, ٢٠٠٤, vol. ٢, pp. ٧٤٩-٧٥٤: IEEE.
- [١٥] B. Manjula, S. Sarma, A. Govardhan, and L. Naik, "DFFS: Detecting Fraud in Finance Sector," *Int. J. Adv. Eng. Sci. Technol*, vol. ٩, no. ٢, pp. ١٧٨-١٨٢, ٢٠١١.
- [١٦] A. Abdallah, M. A. Maarof, and A. Zainal" ,Fraud detection system: A survey," *Journal of Network and Computer Applications*, vol. ٦٨, pp. ٩٠-١١٣, ٢٠١٦.
- [١٧] 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. ٢٠٠٤٠٣١, ٢٠٠٤.
- [١٨] 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)*, ٢٠١٣, pp. ١-٦: IEEE.
- [١٩] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. ٤٦, no. ٤, pp. ١-٣٧, ٢٠١٤.

- [20] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," *Credit scoring and credit control VII*, pp. 230-200, 2001.
- [21] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1-130, 2009.
- [22] 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.
- [23] C. S. Hilar 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, pp. 1112-1121: Springer.
- [24] 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. 2, pp. 112-120, 2004.
- [25] 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. 290-331, 1999.
- [26] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.
- [27] 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. 9, no. 3, pp. 397-419, 2006.
- [28] 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. 12, no. 4, pp. 438-458, 2009.
- [29] 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, pp. 168-174: IEEE.
- [30] N. Sánchez-Marono, 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, pp. 178-187: Springer.
- [31] 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, p. 24: Multidisciplinary Digital Publishing Institute.
- [32] 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. 11-20, 2006.
- [33] N. N. Taleb, *The black swan: The impact of the highly improbable*. Random house, 2007.
- [34] 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. 109, 2011.
- [35] 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. 21, no. 3, pp. 201-212, 2007.
- [36] 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. 111-119, 2013.
- [37] 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. 269-282, 2013.
- [38] C.-P. Shen et al., "A data-mining framework for transnational healthcare system," *Journal of medical systems*, vol. 36, no. 4, pp. 2060-2070, 2012.

- [39] 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.
- [40] 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.
- [41] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1-58, 2009.
- [42] S. Bendre, "Outliers in Statistical Data," ed :JSTOR, 1994.
- [43] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 80-126, 2004.
- [44] 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.
- [45] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," *Social Networks*, vol. 39, pp. 72-70, 2014.
- [46] 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, pp. 724-730: Springer.
- [47] L. Mookiah, W. Eberle, and L. Holder, "Discovering Suspicious Behavior Using Graph-Based Approach," in *The Twenty-Eighth International Flairs Conference*, 2015.
- [48] 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, pp. 398-409: Springer.
- [49] 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.
- [50] 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, pp. 346-350: IEEE.
- [51] 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.
- [52] C. C. Aggarwal, Y. Zhao, and S. Y. Philip, "Outlier detection in graph streams," in *2011 IEEE 27th International Conference on Data Engineering*, 2011, pp. 399-409: IEEE.
- [53] 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. 740-762, 2010.
- [54] 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, pp. 17-24: IEEE.
- [55] B. Ruhnau, "Eigenvector-centrality—a node-centrality?," *Social networks*, vol. 22, no. 4, pp. 307-370, 2000.
- [56] M. Barthélemy, "Betweenness centrality in large complex networks," *The European physical journal B*, vol. 38, no. 2, pp. 163-168, 2004.
- [57] 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.
- [58] 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. 270-287, 2008.
- [59] 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. 303-308, 2008.

- [٦٠] 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. ٦, pp. ٢٦-٢٩, ٢٠٠٦.
- [٦١] 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)*, ٢٠٠٧, pp. ٥١١-٥١٦: IEEE.
- [٦٢] 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. ٧, no. ٧, pp. ٦٨٧-٧٠٣, ١٩٩٢.
- [٦٣] 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*, ١٩٩٨, pp. ٧٤-٨١: Springer.
- [٦٤] 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*, ٢٠١٠, pp. ٦٥-٧٤.
- [٦٥] T. Ekin, F. Ieva, F. Ruggeri, and R. Soyer, "Statistical medical fraud assessment: exposition to an emerging field," *International Statistical Review*, vol. ٨٦, no. ٣, pp. ٣٧٩-٤٠٢, ٢٠١٨.
- [٦٦] 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. ٣٤, no. ٣, pp. ١٩٦٣-١٩٧٣, ٢٠٠٨.
- [٦٧] R. M. Musal, "Two models to investigate Medicare fraud within unsupervised databases," *Expert Systems with Applications*, vol. ٣٧, no. ١٢, pp. ٨٦٢٨-٨٦٣٣, ٢٠١٠.
- [٦٨] T. Ekina, F. Leva, F. Ruggeri, and R. Soyer, "Application of bayesian methods in detection of healthcare fraud," *chemical engineering Transaction*, vol. ٣٣, ٢٠١٣.
- [٦٩] G. C. Capelleveen, "Outlier based predictors for health insurance fraud detection within US Medicaid," *MS thesis. University of Twente*, ٢٠١٣.
- [٧٠] 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*, ٢٠٠٩, pp. ٩٣-٩٨.
- [٧١] 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 ١٢١*, ٢٠١١, pp. ١٠٣-١١٠.
- [٧٢] 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)*, ٢٠١٧, pp. ٢٢٩-٢٣٨: IEEE.
- [٧٣] V. S. Iyengar, K. B. Hermiz, and R. Natarajan, "Computer-aided auditing of prescription drug claims," *Health care management science*, vol. ١٧, no. ٣, pp. ٢٠٣-٢١٤, ٢٠١٤.
- [٧٤] 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)*, ٢٠١٦, pp. ٣٤٧-٣٥٤: IEEE.
- [٧٥] R. M. Musal and T. Ekin, "Medical overpayment estimation: A Bayesian approach," *Statistical Modelling*, vol. ١٧, no. ٣, pp. ١٩٦-٢٢٢, ٢٠١٧.
- [٧٦] S. Rao and P. Gupta, "Implementing improved algorithm over apriori data mining association rule algorithm ١," ٢٠١٢.
- [٧٧] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. 20th int. conf. very large data bases, VLDB*, ١٩٩٤, vol. ١٢١٥, pp. ٤٨٧-٤٩٩.
- [٧٨] 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*, ٢٠١١, pp. ٤٢١-٤٢٨: IEEE.

- [79] 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, pp. 330-334: IEEE.
- [80] 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, pp. 327-331: IEEE.
- [81] 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, pp. 199-199: IEEE.
- [82] R. J. Bolton and D. J. Hand, "Peer group analysis-local anomaly detection in longitudinal data," *Technical Report*, 2001.
- [83] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.
- [84] 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.
- [85] 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.
- [86] 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, pp. 3664-3667: IEEE.
- [87] 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.
- [88] A. Gangopadhyay and S. Chen, "Health care fraud detection with community detection algorithms," in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2016, pp. 1-5: IEEE.
- [89] 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.
- [90] P. Ferreira, R. Alves, O. Belo, and L. Cortesão, "Establishing fraud detection patterns based on signatures," in *Industrial Conference on Data Mining*, 2006, pp. 526-538: Springer.
- [91] 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, pp. 563-574: Springer.
- [92] 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.
- [93] 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.