

الحمد لله  
الذي هدانا لهذا  
الذي كنا لنهتدي لولا  
أن هدانا الله



دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

پایان نامه‌ی کارشناسی ارشد

گرایش نرم افزار

استحکام الگوریتم‌هایی که از قطعه‌بندی در تشخیص فعالیت استفاده می‌نمایند

نگارش

علی امانی

استاد راهنما

دکتر محمدرضا رزازی

شهریور ۱۴۰۱

## برگ ارزیابی پایان نامه شماره ۱۵۵۲۴۵#

کارشناسی ارشد (تئوری)

علی امانی (۹۹۱۳۱۰۲۵)

دانشکده	مهندسی کامپیوتر
رشته-گرایش	مهندسی کامپیوتر - نرم افزار
عنوان اولیه پایان نامه	بررسی و بهبود استحکام الگوریتم هایی که از قطعه بندی در تشخیص فعالیت استفاده می نمایند
عنوان تغییر یافته	تغییر ندارد
تاریخ تصویب	۱۶-۲۱-JUN
تعداد واحد پایان نامه	۶
تاریخ دفاع	۱۴۰۱/۰۶/۱۹
نوع برگزاری	نشست حضوری :: دانشکده مهندسی کامپیوتر - طبقه دوم آمفی تئاتر دانشکده

### اعضای هیات داوران

#	هیات داوران	نام و نام خانوادگی	کد پورتال	وضعیت حضور
۱	استاد راهنما یک	محمدرضا رزازی	۱۰۰۹۶	حاضر در نشست فیزیکی
۲	استاد راهنما دو	---	---	---
۳	استاد راهنما سه	---	---	---
۴	استاد مشاور یک	---	---	---
۵	استاد مشاور دو	---	---	---
۶	داور داخلی یک	علیرضا باقری	۱۰۹۴۱	حاضر در نشست فیزیکی
۷	داور داخلی دو	---	---	---
۸	داور خارجی یک	حسن نادری	۲۰۰۷۴	حاضر در نشست فیزیکی
۹	داور خارجی دو	---	---	---
۱۰	داور آخر	---	---	---
۱۱	نماینده تحصیلات تکمیلی	علیرضا باقری	۱۰۹۴۱	حاضر در نشست فیزیکی

گواهی می گردد ارزیابی دفاع ۱۸/۰۰ (هجده تمام از بیست)، نمره مقاله یا کار ویژه ۰/۰۰ (صفر) و نمره نهایی ارزیابی دفاع پایان نامه ۱۸/۰۰ (هجده) با درجه بسیار خوب ثبت گردیده است.

مدیریت تحصیلات تکمیلی  
معاونت تحصیلات تکمیلی دانشکده  
مهندسی کامپیوتر



به نام خدا

تاریخ: ۱۴۰۱/۶/۱۴

## تعهدنامه اصالت اثر

این جانب علی امانی متعهد می‌شوم که مطالب مندرج در این پایان‌نامه، حاصل کار پژوهشی تحت نظارت و راهنمایی اساتید دانشگاه صنعتی امیرکبیر بوده و به علت حفظ امانت دستاوردهای دیگران، هرگونه مطالبی که در این پژوهش از آنها استفاده شده، نام صاحب اثر مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مآخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر است. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، نسخه‌برداری، ترجمه و اقتباس از این پایان‌نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است.

نقل مطالب با ذکر مآخذ بلامانع است.

علی امانی

تقدیم به:

مادر فداکارم؛

که سایه‌ی مهربانش، سایه‌سار زندگی‌ام است، او که اسوه صبر و تحمل است و مشکلات مسیر را برایم تسهیل کرده است

و

پدر مهربانم.

سپاس...

سپاس و ستایش معبود یگانه‌ای را که پرتوی الطاف بی‌شمارش بر لحظه‌لحظه‌ی زندگی‌ام ساطع و آشکار است. شاکر و ثناگوی کسی هستم که فکرت و اندیشه را در بستر روحم روان ساخت و بهره‌گیری از خوان گسترده‌ی دانش اساتیدم را نصیب و روزی‌ام گردانید.

از تلاش‌ها، زحمات و راهنمایی‌های ظریف و ارزشمند استاد فرزانه و گران‌مایه‌ام، جناب آقای دکتر محمدرضا رزازی سپاسگزارم که با حمیت و جدیت، مرا به دقت، اندیشه، درک و تعمق واداشتند.

از جناب آقای دکتر علیرضا باقری و دکتر حسن نادری که قبول زحمت کردند و داوری این پایان‌نامه را بر عهده گرفتند کمال تشکر و قدردانی را دارم.

از پدر و مادرم به‌خاطر پشتیبانی‌هایشان صمیمانه تشکر می‌کنم.

با آرزوی موفقیت برای تمام عزیزان

علی امانی

شهریور ۱۴۰۱

## چکیده

در سال‌های اخیر، قطعه‌بندی جریان داده به یک موضوع بسیار مهم در علم یادگیری ماشین تبدیل شده است، زیرا بدین ترتیب داده‌های بیشتری را می‌توان به‌عنوان جریان داده در نظر گرفت. بیشتر این داده‌ها همراه با سرعت بالا، توزیع داده‌های غیر ایستا و طول نامحدود مشخص می‌شوند که نمونه‌هایی از این کاربردها را می‌توان در تشخیص فعالیت، تراکنش‌های کارت اعتباری و مخابرات مشاهده کرد. وقتی توزیع جریان داده در طول زمان تغییر می‌کند، پدیده‌ی رانش مفهوم ایجاد می‌شود. رانش مفهوم، مستلزم تنظیم طبقه‌بندی‌ها برای انطباق با شرایط جدید است. نوع تغییرات را می‌توان به رانش مفهوم ناگهانی، تدریجی، افزایشی و مکرر طبقه‌بندی کرد. داده‌های دنیای واقعی اغلب شامل ترکیبی از انواع مختلفی از رانش‌های مفهوم هستند. در این پایان‌نامه با بهره‌گیری از ۴ روش پرکاربرد در طبقه‌بندی مبتنی بر قطعه‌ی جریان داده‌ها (ACDWM، DWMIL، DUE و KUE)، یک رابطه به‌منظور مقایسه‌ی استحکام الگوریتم‌ها با همدیگر در برابر نویز ارائه می‌شود و همچنین در نظر گرفتن نرخ عدم توازن و رانش مفهوم مکمل آن است. الگوریتم‌های مذکور با ۴ مجموعه داده، مورد بررسی و آزمایش قرار گرفتند. با بررسی نمودارهای به‌دست‌آمده از وارد کردن مقدارهای مختلف نویز، نرخ عدم توازن و رانش مفهوم به داده‌ها، رابطه‌ی کارآمدی به‌دست‌آمده است.

کلمات کلیدی: قطعه‌بندی جریان داده، تشخیص فعالیت، نویز، نرخ عدم توازن، رانش مفهوم.

چکیده .....	و
۱- کلیات تحقیق .....	۱
۱-۱ مقدمه .....	۲
۲-۱ بیان مسئله .....	۲
۳-۱ اهداف تحقیق .....	۵
۴-۱ فرضیه‌ها .....	۵
۵-۱ ساختار پایان‌نامه .....	۶
۲- ادبیات و پیشینه‌ی تحقیق .....	۷
۱-۲ مقدمه .....	۸
۲-۲ قطعه‌بندی .....	۸
۳-۲ طبقه‌بندی .....	۹
۱-۳-۲ ماشین بردار پشتیبان .....	۹
۲-۳-۲ درخت تصمیم .....	۱۰
۴-۲ بررسی دو مقاله‌ی مفید و کاربردی .....	۱۱
۱-۴-۲ مقاله‌ی اول .....	۱۱
۲-۴-۲ مقاله‌ی دوم .....	۲۲
۵-۲ معرفی مجموعه‌داده‌ها .....	۲۷
۱-۵-۲ SKODA .....	۲۷
۲-۵-۲ WISDM .....	۲۸
۳-۵-۲ HCI .....	۲۹
۴-۵-۲ PS .....	۳۲
۶-۲ پیشینه‌ی پژوهش .....	۳۴
۷-۲ مقایسه‌ی کارهای پیشین .....	۳۹
۸-۲ نتیجه‌گیری .....	۴۰
۳- راهکار پیشنهادی .....	۴۱
۱-۳ مقدمه .....	۴۲
۲-۳ تعریف مسئله .....	۴۲
۱-۲-۳ نويز چیست؟ .....	۴۲
۲-۲-۳ استحکام در برابر نويز .....	۴۳
۳-۲-۳ توليد نويز .....	۴۳
۴-۲-۳ الگوریتم‌های استفاده‌شده .....	۴۴
۵-۲-۳ مجموعه‌داده‌های استفاده‌شده در آزمایش‌ها .....	۴۷



۵۰	۶-۲-۳ محاسبه‌ی استحکام
۵۳	۷-۲-۳ محاسبه‌ی نویز، نرخ عدم توازن و رانش مفهوم
۶۰	۳-۳ مراحل راهکار پیشنهادی
۶۰	۱-۳-۳ اعمال نویز گاوسی به الگوریتم‌ها
۶۰	۲-۳-۳ به دست آوردن نمودارها و محاسبه‌ی معیار ناحیه‌ی زیر منحنی
۶۰	۳-۳-۳ محاسبه‌ی انحراف معیار
۶۰	۴-۳-۳ محاسبه‌ی انحراف معیار الگوریتم
۶۰	۵-۳-۳ مقایسه و نتیجه‌گیری
۶۱	۴-۳ نتیجه‌گیری
۶۲	۴-ارزیابی راهکار پیشنهادی
۶۳	۱-۴ مقدمه
۶۳	۲-۴ نتایج ارزیابی
۷۴	۳-۴ نتیجه‌گیری
۷۴	۱-۳-۴ سناریوی اول
۷۵	۲-۳-۴ سناریوی دوم
۷۵	۳-۳-۴ سناریوی سوم
۷۶	۵-نتیجه‌گیری و پیشنهادها
۷۷	۱-۵ مقدمه
۷۷	۲-۵ مزایا و معایب
۷۷	۱-۲-۵ مزایا
۷۷	۲-۲-۵ معایب
۷۸	۳-۵ نتیجه‌گیری کلی
۷۹	۴-۵ پیشنهادها
۸۱	منابع
۸۵	واژه‌نامه‌ی فارسی به انگلیسی
۹۲	واژه‌نامه‌ی انگلیسی به فارسی
۹۹	Abstract

- شکل ۱-۲- فعالیتهای بدنی مورد استفاده برای توان بخشی بیماران آسیب نخاعی..... ۱۲
- شکل ۲-۲- جهت محورها..... ۱۳
- شکل ۳-۲- سیگنال شتاب برای ۱۰ تکرار حرکتهای (الف) انقباض، (ب) چرخش داخلی و (ج) گسترش..... ۱۴
- شکل ۴-۲- قلههای شناسایی شده در سیگنال شتاب: (الف) انتخاب اولیه قلهها (ب) پس از اعمال آستانه‌ی اول و حداقل فاصله..... ۱۵
- شکل ۵-۲- تعیین نوع مرز بر اساس مشخصات سیگنال (P: قله، V: دره)..... ۱۷
- شکل ۶-۲- سیگنال شتاب چهار تکرار خم شدن آرنج که با استفاده از پنجره‌ی کشویی (قسمت پایین) و روش پیشنهادی (خطوط نقطه‌چین) قطعه‌بندی شده‌اند ..... ۲۱
- شکل ۷-۲- تشخیص فعالیت چند ساکن..... ۲۴
- شکل ۸-۲- خوشه‌بندی فازی بر اساس مکان حسگرها..... ۲۷
- شکل ۹-۲- شتاب در محورهای مجموعه داده‌ی WISDM ..... ۲۹
- شکل ۱۰-۲- شتاب پای چپ برای ۶ حرکت هوازی معمولی در مجموعه داده‌ی HCI ..... ۳۰
- شکل ۱۱-۲- حسگرهای شتاب بلوتوث در مجموعه داده‌ی HCI ..... ۳۱
- شکل ۱۲-۲- کلاسهای مختلف فعالیتهای در مجموعه داده‌ی HCI ..... ۳۲
- شکل ۱-۳- مقدار نرخ عدم توازن در مجموعه داده ..... ۴۸
- شکل ۲-۳- نمودارهای سناریوی اول (نویز متغیر) برای مجموعه داده‌های تشخیص فعالیت..... ۵۵
- شکل ۳-۳- نمودارهای سناریوی دوم (رانش مفهوم صفر) برای مجموعه داده‌های تشخیص فعالیت..... ۵۷
- شکل ۴-۳- نمودارهای سناریوی سوم (نرخ عدم توازن صفر) برای مجموعه داده‌های تشخیص فعالیت..... ۵۹
- شکل ۱-۴- نمودارهای حاصل از نرخ عدم توازن متغیر با نویز صفر..... ۶۶
- شکل ۲-۴- نتایج حاصل از نویز ۴۰ درصد..... ۶۸

صفحه	فهرست جداول
۱۷	جدول ۱-۲- شبه کد الگوریتم قطعه‌بندی پویا.....
۳۹	جدول ۲-۲- پیشینه .....
۴۸	جدول ۱-۳- مجموعه داده‌های استفاده شده.....
۶۹	جدول ۱-۴- نتایج آزمایش‌ها در سناریوی اول.....
۷۱	جدول ۲-۴- انحراف معیار نتایج در سناریوی اول بر روی مجموعه داده‌های تشخیص فعالیت.....
۷۱	جدول ۳-۴- نتایج آزمایش‌ها در سناریوی دوم.....
۷۲	جدول ۴-۴- انحراف معیار نتایج در سناریوی دوم بر روی مجموعه داده‌های تشخیص فعالیت.....
۷۳	جدول ۵-۴- نتایج آزمایش‌ها در سناریوی سوم.....
۷۴	جدول ۶-۴- انحراف معیار نتایج در سناریوی سوم بر روی مجموعه داده‌های تشخیص فعالیت.....

## فهرست علائم

## علائم لاتین

ناحیه‌ی زیر منحنی	AUC
دقت	Accuracy
نرخ مثبت واقعی	Tprate
نرخ مثبت کاذب	Fprate

## علائم یونانی

انحراف استاندارد	$\sigma$
میانگین ناحیه‌ی زیر نمودار برای همه‌ی قطعه‌ها در حالت بدون نویز	$\mu$

## بالانویس‌ها

تعداد آزمایشات	N
----------------	---

## زیرنویس‌ها

مقادیر میانگین ناحیه‌ی زیر نمودار برای همه‌ی قطعات در حالت نویزدار	$x_i$
--	-------

## فصل اول

### کلیات تحقیق

## ۱-۱ مقدمه

قطعه‌بندی<sup>۱</sup> داده‌ها یک فرایند ضروری برای تشخیص فعالیت<sup>۲</sup>، در هنگام استفاده از تکنیک‌های یادگیری ماشین<sup>۳</sup> است. مطالعات قبلی در مورد تشخیص فعالیت بدنی بیشتر بر رویکرد پنجره‌ی کشویی<sup>۴</sup> برای قطعه‌بندی تکیه کرده‌اند. با این حال، انتخاب یک اندازه‌ی پنجره‌ی ثابت برای چندین فعالیت با مدت زمان‌های مختلف ممکن است بر دقت تشخیص تأثیر بگذارد، به‌ویژه زمانی که فعالیت‌ها فقط به یک دسته‌ی متحرک یا غیرمتحرک تعلق دارند. در این فصل به بیان کلیات پژوهش حاضر خواهیم پرداخت.

## ۲-۱ بیان مسئله

در این پایان‌نامه به مقایسه‌ی ۴ الگوریتم شناخته‌شده در زمینه‌ی طبقه‌بندی مبتنی بر قطعه‌ی جریان داده‌ها می‌پردازیم. جریان داده‌ها نوعی از داده‌های حجیم هستند و امکان ذخیره‌ی همه‌ی داده‌ها برای پردازش وجود ندارد. الگوریتم‌های مربوط به جریان داده‌ها فقط روی داده‌های اخیر و به صورت افزایشی<sup>۵</sup> کار می‌کنند و در هر لحظه می‌توانند نتیجه را اعلام کنند.

الگوریتم مستحکم<sup>۶</sup> به این معنی است که الگوریتم جواب پیدا کند و با ایجاد نویز در داده‌ها متوقف نشود؛ خروجی برای همه‌ی ورودی‌ها به جواب واقعی نزدیک باشد و تغییرات زیادی نداشته باشد.

الگوریتم مستحکم، خطا در ورودی و محاسبات را تحمل می‌کند و به سمت جواب درست همگرا می‌شود. هر قطعه از جریان داده‌ها طبقه‌بندی می‌شود، سپس الگوریتم‌ها از نظر استحکام مقایسه می‌شوند.

بر اساس رویکرد مورد استفاده برای جمع‌آوری داده‌ها، تشخیص فعالیت را می‌توان به‌طور کلی بر مبنای دو رویکرد مبتنی بر دید و مبتنی بر حسگر طبقه‌بندی کرد.

---

<sup>1</sup> Segmentation

<sup>2</sup> Activity recognition (AR)

<sup>3</sup> Machine learning (ML)

<sup>4</sup> Sliding window

<sup>5</sup> Incremental

رویکرد مبتنی بر دید، غنی از اطلاعات است، اما اغلب نگرانی‌های اخلاقی و حفظ حریم خصوصی، به‌ویژه در برنامه‌ی مراقبت‌های بهداشتی هنگام برخورد با بیماران با مشکل مواجه می‌شود. در مقابل، دستگاه‌های مورد استفاده در رویکرد مبتنی بر حسگر، از جمله حسگرهای پوشیدنی، می‌توانند با هزینه و توان محدود کار کنند و از نظر محیط اطراف یا مکانی که فعالیت‌ها باید انجام شوند، هیچ محدودیتی ندارند. در نتیجه، سامانه‌های تشخیص فعالیت معمولاً رویکرد مبتنی بر حسگر را اتخاذ می‌کنند [۱].

مطالعات متعددی برای بررسی تأثیر موقعیت‌های مختلف حسگر بر دقت تشخیص انجام شده است. این مطالعات نشان می‌دهد که موقعیت حسگر باید عمده‌تاً بر اساس نوع فعالیت مورد مطالعه تعیین شود. اشکال حرکت، از جمله راه رفتن، دویدن و همچنین فعالیت‌های غیرحرکتی، مانند ایستادن و نشستن را می‌توان با دقت بین ۸۳٪ تا ۹۵٪ با قرار دادن حسگر بر روی بخش‌های اندام تحتانی (ران و مچ پا) تشخیص داد. برای افزایش دقت تشخیص فعالیت‌های اندام فوقانی، حسگرها روی مچ دست و بازو قرار می‌گیرند [۲].

تشخیص فعالیت انسانی<sup>۱</sup> با هدف یادگیری و شناسایی انواع مختلف فعالیت‌های انسانی از فرمت‌های مختلف داده مانند تصاویر و سیگنال‌های حسگر بهره می‌برد [۳]. تشخیص فعالیت انسانی به طور گسترده در برخی از کاربردهای دنیای واقعی مانند سامانه‌های زندگی به کمک محیط<sup>۲</sup> و نظارت بر خانه استفاده شده است، زیرا به تشخیص فعالیت‌های غیرمنتظره، مانند سقوط ناگهانی یا موقعیت‌های خطرناک کمک می‌کند.

با افزایش جمعیت سالمندان و پیشرفت در محیط‌های هوشمند، تشخیص فعالیت انسانی به یک زمینه‌ی تحقیقاتی محبوب تبدیل شده است. توسعه‌ی حسگرها و فناوری ارتباطات، امکان تحقق خانه‌های هوشمند را فراهم می‌کند. با نصب انواع مختلف دستگاه‌ها در یک خانه‌ی هوشمند، می‌توان فعالیت‌های ساکنان را توسط بسیاری از روش‌های یادگیری ماشین، نظارت و شناسایی کرد [۴].

تشخیص فعالیت انسانی، توجه زیادی را به خود جلب کرده است، زیرا می‌تواند در نظارت مستمر بر رفتارهای انسان در خانه‌های هوشمند، سامانه‌های نظارت هوشمند، تولید پیشرفته، مراقبت‌های بهداشتی، توان بخشی، تشخیص رفتار غیرعادی، بازی، تناسب‌اندام شخصی و... استفاده‌ی قابل توجهی از آنها کرد. چارچوب‌های

<sup>1</sup> Human activity recognition (HAR)

<sup>2</sup> Ambient assisted living (AAL)

تشخیص فعالیت انسانی راه‌هایی را برای استفاده از داده‌های به‌دست‌آمده ارائه کرده است که توسط حسگرهای مختلف، حرکات بدن، فعالیت‌های سرپایی و وضعیتی و اقدامات کاربر را تحت نظر می‌گیرد [۵]. این روش‌ها عمدتاً شامل استفاده از حسگرهای مبتنی بر ویدئو، حسگرهای پوشیدنی، حسگرهای گوشی هوشمند و حسگرهای بی‌سیم هستند. تشخیص فعالیت انسانی را می‌توان به‌طور کلی به دو دسته‌ی تشخیص فعالیت مبتنی بر حسگر و مبتنی بر ویدئو طبقه‌بندی کرد.

سامانه‌ی مبتنی بر ویدئو از حسگرهای مبتنی بر ویدئو استفاده می‌کند، مانند دوربین‌هایی که برای ضبط ویدئو و تصاویر با هدف تشخیص فعالیت‌ها در زندگی روزمره و رفتارهای انسانی به کار گرفته می‌شوند. سامانه‌های مبتنی بر حسگر از حسگرهای محیطی یا پوشیدنی برای تشخیص فعالیت‌های انسانی بهره می‌برند.

بیشتر تحقیقات تشخیص فعالیت انسانی به دلیل فراگیر بودن، نصب آسان و سهولت استفاده، شامل جمع-آوری داده‌های تولید شده است که از طریق تلفن‌های هوشمند و حسگرهای پوشیدنی داده‌ها را ذخیره می‌کند. حسگرها به طور گسترده در دستگاه‌های هوشمند مانند تلفن‌های همراه، ساعت‌ها و ... تعبیه شده‌اند. آن‌ها می‌توانند به طور مداوم اطلاعات مربوط به حرکت انسان را ثبت کنند.

در سال‌های اخیر، چندین تکنیک یادگیری ماشین مانند «ماشین بردار پشتیبان»<sup>۱</sup>، «جنگل تصادفی»<sup>۲</sup> و «نزدیک‌ترین همسایگان» برای تشخیص فعالیت انسانی به کار گرفته شده است. به‌عنوان مثال، روش تشخیص فعالیت انسانی پیشنهادی، از ویژگی‌های مبتنی بر تجزیه‌ی حالت تجربی مجموعه استفاده می‌کند. ویژگی‌ها و روش انتخاب گزینه‌ها با استفاده از طبقه‌بندهای<sup>۳</sup> «ماشین بردار پشتیبان» و روش «نزدیک‌ترین همسایگان» ارزیابی شده است [۶].

<sup>1</sup> Support vector machine (SVM)

<sup>2</sup> Random forest

<sup>3</sup> Classifiers



یک مدل خوشه‌بندی سلسله‌مراتبی مبتنی بر بی‌نظمی<sup>۱</sup> برای تشخیص فعالیت انسانی پیشنهاد شده است که از یک لایه حسگر و یک لایه طبقه‌بند تشکیل می‌شود. برای برآورد وزن در مدل از روش بی‌نظمی استفاده شده است. استفاده از تکنیک‌های یادگیری ماشین برای شناسایی فعالیت‌ها، نیازمند تلاش بسیار زیادی در آماده‌سازی داده‌ها، پیش‌پردازش داده‌ها و استخراج ویژگی است که به نوبه‌ی خود به تخصص دامنه نیاز دارد. ویژگی‌هایی که با استفاده از این تکنیک‌ها استخراج می‌شوند، برای کارهای مشابه استفاده نمی‌شوند [۷].

### ۳-۱ اهداف تحقیق

- ۱- بررسی استحکام الگوریتم‌هایی که از قطعه‌بندی در تشخیص فعالیت استفاده می‌کنند.
- ۲- بررسی سه عامل نویز، رانش مفهوم<sup>۲</sup> و نرخ عدم توازن در الگوریتم‌های جریان داده که برای قطعه‌بندی به کار می‌روند.
- ۳- ارائه‌ی فرمول پیشنهادی برای استحکام و تأثیر نویز، رانش مفهوم و نرخ عدم توازن در قطعه‌بندی که به مقایسه‌ی الگوریتم‌ها می‌پردازد.
- ۴- آزمایش‌های متعددی که برای وارد کردن نویز گاوسی به الگوریتم‌ها و بررسی معیار ناحیه‌ی زیر منحنی<sup>۳</sup> انجام می‌شود.

### ۴-۱ فرضیه‌ها

- ۱- داده‌ها به بخش‌های کوچک‌تر قطعه‌بندی می‌شوند تا پیش‌بینی بهتر و سریع‌تر گردد.
- ۲- نویز گاوسی<sup>۴</sup> روش مناسبی برای اعمال نویز به الگوریتم‌های طبقه‌بندی شده است.
- ۳- با اضافه کردن نویز گاوسی به الگوریتم‌ها و بررسی معیار ناحیه‌ی زیر منحنی، می‌توان استحکام الگوریتم‌ها را بررسی کرد.

<sup>۱</sup> Entropy

<sup>۲</sup> Concept drift

<sup>۳</sup> Area under the ROC curve (AUC)

<sup>۴</sup> Gaussian noise

۴- عوامل نويز، رانش مفهوم و نرخ عدم توازن در نتیجه‌ی الگوریتم‌های اجرا شده روی مجموعه داده‌ها<sup>۱</sup> تأثیرگذار هستند.

## ۱-۵ ساختار پایان‌نامه

در فصل اول پایان‌نامه کلیات تحقیق بیان شد که شامل مباحثی چون بیان مسئله، سؤالات و فرضیات تحقیق و ... است. فصل دوم شامل ادبیات و پیشینه‌ی تحقیق است. مباحث نظری موضوع بررسی می‌شود و کارهای انجام شده مورد بحث قرار می‌گیرد. در فصل سوم، راهکار پیشنهادی و جزئیات آن شرح داده می‌شود. فصل چهارم به نتایج حاصل از پیاده‌سازی و مقایسه‌ی الگوریتم‌ها می‌پردازد و در نهایت در فصل پنجم، نتیجه‌گیری و پیشنهادهای آتی مطرح می‌گردد.

---

<sup>۱</sup> Datasets

## فصل دوم

### ادبیات و پیشینه‌ی تحقیق

## ۲-۱ مقدمه

تشخیص فعالیت، به عنوان یک کاربرد مهم هوش مصنوعی، در طیف گسترده‌ای از برنامه‌های کاربردی دنیای واقعی مورد استفاده قرار گرفته که از آن جمله می‌توان به مراقبت‌های بهداشتی، خانه‌های هوشمند، امنیت و زندگی به کمک محیط اشاره کرد. هدف از تشخیص فعالیت، تشخیص خودکار فعالیت‌ها یا رفتارهای انسان از یک سری سیگنال‌های دریافتی است.

به طور سنتی، در بینایی کامپیوتر<sup>۱</sup>، مطالعات زیادی در مورد تشخیص حرکات و فعالیت‌ها از تصاویر ثابت و فیلم‌ها کمک گرفته شده است. با این حال، به دلیل مسئله‌ی حریم خصوصی، سامانه‌های تشخیص فعالیت مبتنی بر دید را می‌توان تنها در برخی از محیط‌های محدود استفاده کرد. با توسعه‌ی فناوری حسگر، حسگرهای بی‌سیم مختلف مانند شناسایی فرکانس رادیویی<sup>۲</sup>، شبکه‌ی بی‌سیم، حسگرهای پوشیدنی و ... به طور گسترده در زندگی روزمره‌ی ما در دسترس هستند. تلاش‌های اخیر به سمت توسعه‌ی سامانه‌های تشخیص فعالیت مبتنی بر حسگر تغییر یافته است.

## ۲-۲ قطعه‌بندی

قطعه‌بندی در علم یادگیری ماشین، نوعی یادگیری بدون نظارت است که داده‌ها را بدون نیاز به برچسب و بر اساس برخی معیارهای اندازه‌گیری شده، به چندین بخش قطعه‌بندی می‌کند [۸]. در این موضوع خاص، وظیفه‌ی قطعه‌بندی داده‌ها، شناسایی بخش‌های طولی متغیر از جریان داده‌های چندمتغیره است که احتمالاً حاوی اطلاعات مربوط به فعالیت‌ها هستند. پنجره‌ی کشویی، یک رویکرد رایج برای قطعه‌بندی جریان داده‌های مربوط به تشخیص فعالیت است. اندازه‌ی یک پنجره‌ی کشویی به صورت تجربی و بر اساس محدودیت‌های سخت‌افزاری انتخاب می‌شود. در قطعه‌بندی داده‌های جریانی پیوسته برای تشخیص فعالیت، معمولاً یک پنجره‌ی کشویی روی داده‌های ایستا و پویا<sup>۳</sup> حرکت داده می‌شود. سپس تفاوت بین دو پنجره‌ی مجاور با مقدار آستانه محاسبه می‌شود تا نقطه‌ی شکست پیدا شود [۹].

<sup>1</sup> Computer vision

<sup>2</sup> Radio frequency identification (RFID)

<sup>3</sup> Dynamic

## ۳-۲ طبقه‌بندی

یکی از مراحل بازشناسی الگو، انتخاب نوع مدل طبقه‌بندی است. طبقه‌بند، هسته‌ی اصلی سامانه‌ی بازشناسی الگو است. طبقه‌بند، هر الگوی ناشناخته را بر اساس ویژگی‌های آن، به یکی از کلاس‌های شناخته‌شده نسبت می‌دهد. طبقه‌بندهای متعددی مانند «طبقه‌بند بیز<sup>۱</sup>»، «شبکه‌های عصبی»، «ماشین بردار پشتیبان» و «کمترین فاصله» برای پیش‌بینی فعالیت‌ها پیشنهاد شده‌اند [۱۰]. پس از انتخاب مدل، باید پارامترهای آن را مشخص کرد. پارامترها در طول فرایند یادگیری تعیین می‌شوند. پس از آن که مدل کامل شد، با استفاده از نمونه‌های آزمایشی، سامانه‌ی بازشناسی الگو اعتبارسنجی می‌شود.

## ۱-۳-۲ ماشین بردار پشتیبان

ماشین بردار پشتیبان بیشترین حاشیه‌ی بین دو کلاس را در یک فضای ویژگی خاص دارد [۱۱]. پس ابرصفحه‌ای را انتخاب می‌کند که فاصله‌ی آن از نزدیک‌ترین داده‌ها در هر دو طرف جداکننده‌ی خطی، بیشینه باشد. تابع تصمیم‌گیری برای جداکردن داده‌ها، با استفاده از بردارهای پشتیبان (نزدیک‌ترین داده‌های آموزشی به ابرصفحه جداکننده) تعیین می‌شود. در واقع ابرصفحه‌ی بهینه در ماشین بردار پشتیبان، جداکننده‌ای بین بردارهای پشتیبان است.

مرز تصمیم‌گیری معمولاً با استفاده از یک تابع خطی به دست می‌آید که دو کلاس را از هم جدا می‌کند، اما در بعضی از حالات این مرز به صورت خطی قابل تفکیک نیست؛ بنابراین داده‌ها به فضای ویژگی با ابعاد بالاتر انتقال داده می‌شوند تا به صورت خطی قابلیت جداسازی داشته باشند. ابتدا تابع  $\varphi(x)$  را برای نگاشت به فضای دیگر و انتقال داده‌ها به ابعاد بالاتر پیدا می‌کنیم. تابع تصمیم‌گیری ماشین بردار پشتیبان به این صورت نمایش داده می‌شود:

$$f(x) = \sum_{i=1}^n a_i y_i K(x, x_i) + b \quad \text{رابطه‌ی ۱-۲}$$

<sup>۱</sup> Bayes

در رابطه‌ی ۱-۲،  $f(x)$  خروجی تصمیم،  $y_i$  برچسب داده‌های  $x_i$  و  $x$  داده‌ای است که قرار است طبقه-بندی شود. پارامترهای  $a_i$  و  $b$  در طول آموزش، مقداردهی می‌شوند.  $K(x, x_i) = \varphi(x_i)^T \varphi(x)$  نیز تابع کرنل است.

در ماشین بردار پشتیبان می‌خواهیم مسئله زیر را بهینه‌سازی کنیم:

$$\min \frac{1}{2} ||w||^2 + C \sum_{i=1}^n \epsilon_i \quad \text{subject to } y_i(w^T \varphi(x_i) + b) \geq 1 - \epsilon_i \quad \text{رابطه‌ی ۲-۲}$$

کرنل‌های زیادی وجود دارد؛ اما برای مسئله تشخیص حروف نوری<sup>۱</sup>، بهترین کرنل تابع پایه شعاعی<sup>۲</sup> است:

$$K(x, x_i) = \exp(-\gamma ||x - x_i||^2), \gamma > 1 \quad \text{رابطه‌ی ۳-۲}$$

پارامتر  $C$  در رابطه‌ی ۲-۲ نشان‌دهنده‌ی جریمه و  $\gamma$  در رابطه‌ی ۳-۲ پارامتر کرنل است.

ماشین بردار پشتیبان یک طبقه‌بند دودویی است. ماشین بردار پشتیبان چند کلاسی از ترکیب طبقه‌بندهای دودویی حاصل می‌شود.

### ۲-۳-۲ درخت تصمیم

در حال حاضر استفاده از درخت‌های تصمیم به‌عنوان طبقه‌بند به طور کامل شناخته شده است. درخت تصمیم بر اساس آنالیز داده‌های ورودی و به‌منظور پیدا کردن یک ویژگی به‌عنوان مبنای تصمیم‌گیری برای هر گره<sup>۳</sup> به کار می‌رود. در هر گره، ویژگی‌های مختلف داده‌ها بررسی می‌شود و یک ویژگی گزینش می‌شود که در صورت انتخاب باعث کاهش بی‌نظمی می‌شود. درخت تصمیم به‌عنوان یک طبقه‌بند چندمنظوره و ناپایدار در یادگیری ماشین در نظر گرفته می‌شود [۱۲]. می‌توانیم با آموزش‌های متوالی مجموعه داده‌ها با همان مقادیر اولیه، راه‌حل‌های مختلف را پوشش دهیم.

<sup>1</sup> Optical character recognition (OCR)

<sup>2</sup> Radial basis function (RBF)

<sup>3</sup> Node

## ۴-۲ بررسی دو مقاله‌ی مفید و کاربردی

## ۴-۲-۱ مقاله‌ی اول

مقاله‌ی [۱۳] روش جدیدی است برای قطعه‌بندی پویای<sup>۱</sup> فعالیت‌های بدنی که در راستای توان‌بخشی افراد مبتلا به آسیب نخاعی ارائه و تأیید شده است. در قطعه‌بندی پویای داده‌های خام، ویژگی‌های سیگنال برای تعیین نوع مناسب مرزها تجزیه و تحلیل شده‌اند. سپس مرزهای زمانی برای نشان دادن نقطه‌ی شروع و پایان هر فعالیت شناسایی شده است. برای تأیید روش و ساخت یک مدل، آزمایشی انجام شده است که در آن داده‌ها با استفاده از یک حسگر شتاب‌سنج<sup>۲</sup>، جمع‌آوری شده‌اند. نتایج تجربی با رویکرد پنجره‌ی کشویی<sup>۳</sup> مقایسه شده است که نشان می‌دهد دقت کلی روش پیشنهادی، بیش از ۵٪ بهبود یافته است و همچنین استحکام مدل، بهتر از رویکرد پنجره‌ی کشویی شده است.

هدف از این تحقیق، پیشنهاد یک روش قطعه‌بندی جدید برای تشخیص فعالیت بدنی است که می‌تواند عملکرد طبقه‌بندی را افزایش دهد. برخلاف مطالعات قبلی، این روش مربوط به قطعه‌بندی فعالیت‌های فیزیکی است که به دسته‌ی فعالیت‌های متحرک تعلق دارند. برای دستیابی به این هدف، آزمایشی برای تأیید و مقایسه‌ی روش پیشنهادی با رویکرد پنجره‌ی کشویی انجام شده است. این مقایسه، اثربخشی روش مقاله را به‌ویژه از نظر افزایش دقت تشخیص نشان می‌دهد.

آسیب‌های نخاعی<sup>۴</sup> اندام تحتانی بیماران را تحت تأثیر قرار می‌دهد. در موارد نادر، بیماران آسیب‌های نخاعی بر اساس درجه و محل آسیب، ممکن است دچار فلج کامل شوند. افراد برای بهبود و درمان علائمی مانند قدرت عضلانی کم نیاز به توان‌بخشی دارند. توان‌بخشی از طریق افزایش فعالیت بدنی و توسعه‌ی مهارت‌های حرکتی، اندام فوقانی بیماران را قادر می‌سازد تا فعالیت‌های روزانه را انجام دهند و کیفیت زندگی آن‌ها را ارتقا می‌بخشد.

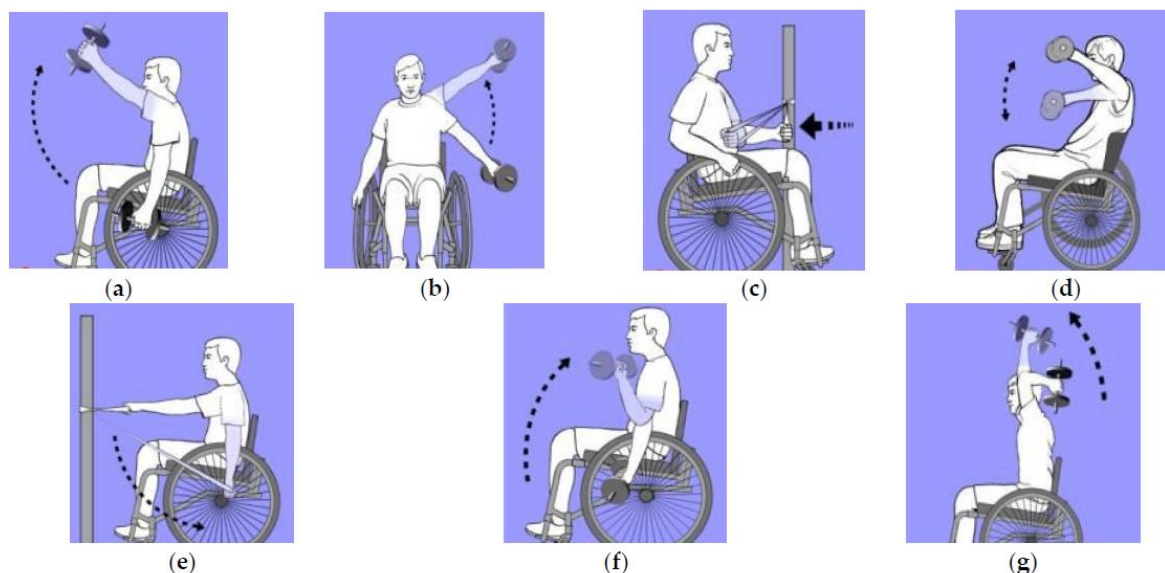
<sup>۱</sup> Dynamic segmentation

<sup>۲</sup> Accelerometer

<sup>۳</sup> Sliding window

<sup>۴</sup> Individuals with spinal cord injuries

قسمت‌های مورد توجه بدن برای تقویت اندام‌های فوقانی آرنج و شانه‌ها هستند. فعالیت‌های اصلی مورد نیاز برای تقویت عضلات شانه، خم شدن شانه<sup>۱</sup>، انقباض شانه<sup>۲</sup>، گسترش<sup>۳</sup>، چرخش داخلی<sup>۴</sup> و چرخش خارجی<sup>۵</sup> هستند. علاوه بر این، فعالیت‌های اصلی که برای تقویت عضلات اصلی آرنج اعمال می‌شود، خم شدن آرنج<sup>۶</sup> و گسترش آرنج<sup>۷</sup> است. تصویری از این فعالیت‌ها در شکل (۱-۲) آورده شده است.



شکل ۱-۲: فعالیت‌های بدنی مورد استفاده برای توان بخشی بیماران آسیب نخاعی: الف) خم شدن شانه. ب) انقباض شانه. ج) چرخش داخلی. د) چرخش خارجی. ه) گسترش. و) خم شدن آرنج. ز) گسترش آرنج [۱۳].

در این مقاله از یک حسگر بی‌سیم استفاده شده که شامل یک شتاب‌سنج سه محوری، یک ژيروسکوپ<sup>۸</sup> سه محوری و یک مغناطیس‌سنج<sup>۹</sup> سه محوری است. با توجه به کارایی شتاب‌سنج‌ها در تشخیص فعالیت، مجموعه داده‌ها با یک شتاب‌سنج سه محوری (دستگاه حسگر برای اندازه‌گیری شتاب در سه جهت متعامد به

<sup>1</sup> Shoulder flexion

<sup>2</sup> Shoulder abduction

<sup>3</sup> Extension

<sup>4</sup> Internal rotation (IR)

<sup>5</sup> External rotation (ER)

<sup>6</sup> Elbow flexion (EF)

<sup>7</sup> Elbow extension (EE)

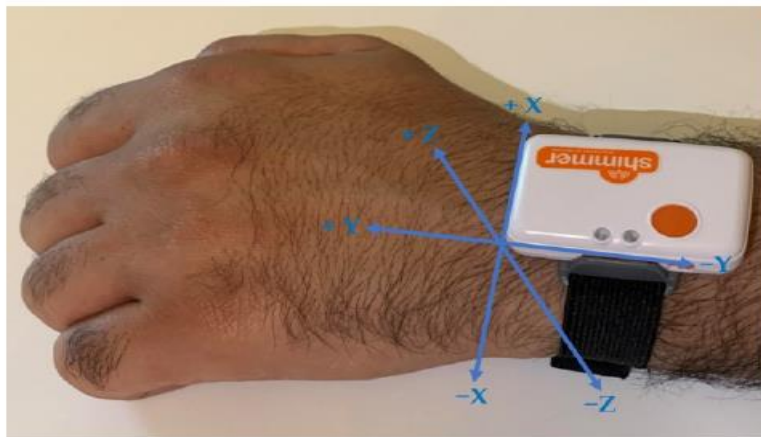
<sup>8</sup> Gyroscope

<sup>9</sup> Magnetometer



طور هم‌زمان) جمع‌آوری شد. ژيروسکوپ و مغناطیس‌سنج حذف شدند؛ زیرا مطالعات قبلی نشان داد که شتاب‌سنج‌ها دقت کلی بالاتری را ارائه می‌دهند. علاوه بر این، مواد فرومغناطیسی که معمولاً در محیط‌های خانگی در دسترس هستند می‌توانند بر مغناطیس‌سنج‌ها تأثیر بگذارند. حسگر برای جمع‌آوری داده‌های شتاب با فرکانس نمونه‌برداری ۳۰ هرتز (محدوده‌ی ۲ گرم) پیکربندی شده است که برای تشخیص فعالیت‌های مشابه کافی است.

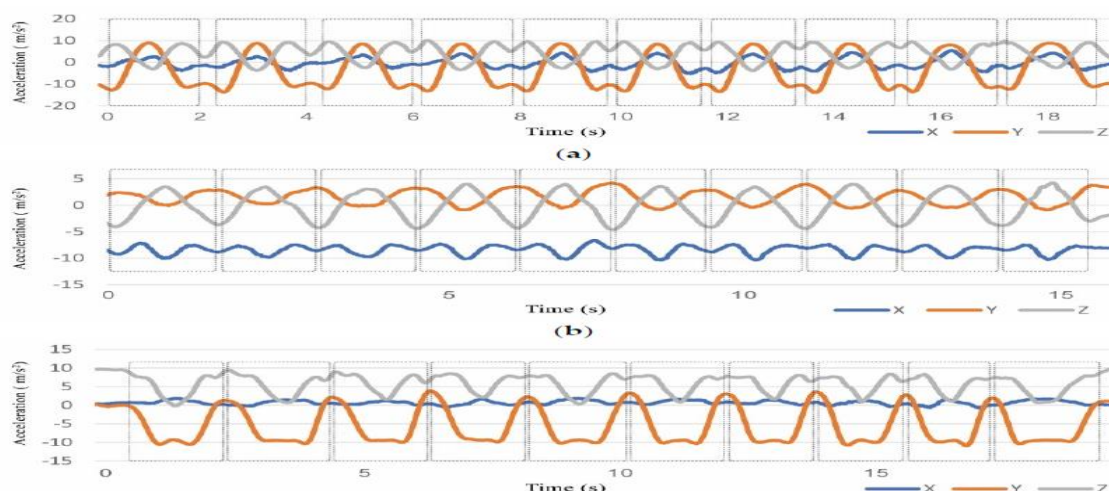
علاوه بر این، مطالعه قبلی نشان داد که نوع و شدت فعالیت‌های انسانی را می‌توان با استفاده از سیگنال‌هایی با نرخ نمونه‌برداری برابر با ۱۰ هرتز تشخیص داد. هنگام تشخیص فعالیت‌های اندام فوقانی، حسگرهایی روی میچ دست و بازو قرار می‌گیرند. فعالیت‌های خاصی مانند گسترش آرنج، خم شدن آرنج و چرخش داخلی فاقد حرکات بازو هستند. پس حسگری که روی بازو قرار می‌گیرد، نمی‌تواند هیچ حرکتی را تشخیص دهد. براین اساس، میچ دست به‌عنوان موقعیت حسگر انتخاب شد. طبق شکل ۲-۲، محور  $X$  در سراسر انگشتان دست، محور  $Y$  در امتداد دست و محور  $Z$  عمود بر محور  $Y$  و در پشت میچ دست قرار دارد.



شکل ۲-۲: جهت محورها [۱۳].

از آنجایی که فعالیت‌های بدنی به‌صورت هم‌زمان انجام می‌شوند، یک الگوی فعالیت را می‌توان با مشاهده‌ی سیگنال شتاب شناسایی کرد. شکل ۲-۳ (الف) داده‌های خام جمع‌آوری شده از یک شتاب‌سنج سه محوری

را در طول ۱۰ تکرار انقباض نشان می‌دهد که در آن هر قسمت محصور در مستطیل نقطه‌چین، نشان‌دهنده‌ی یک تکرار است.



شکل ۲-۳: سیگنال شتاب برای ۱۰ تکرار حرکت‌های (الف) انقباض، (ب) چرخش داخلی و (ج) گسترش [۱۳].

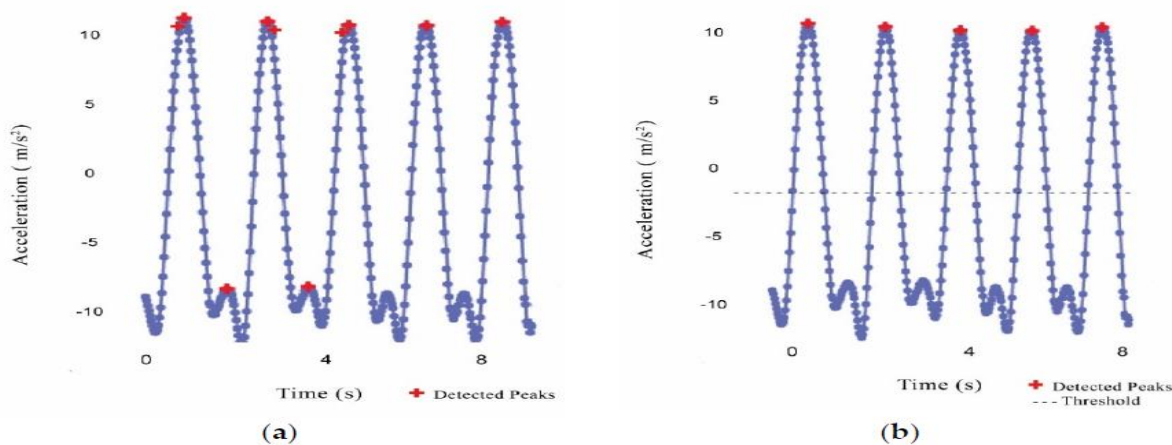
برای تمامی فعالیت‌ها، هر قله در امتداد محور Y مربوط به یک فعالیت است، به جز چرخش داخلی و گسترش که در آن دو قله‌ی متوالی نقطه شروع و پایان فعالیت را نشان می‌دهد. شکل ۲-۳ (ب) و (ج) به ترتیب ۱۰ بار تکرار حرکت‌های چرخش داخلی و گسترش را نشان می‌دهد که در آن تمام نقاط داده‌ی محصور در مستطیل‌های نقطه‌چین به یک تکرار تعلق دارد.

دلیل اصلی این تفاوت، جهت حرکت فعالیت و موقعیت دست در حین حرکت است. روش قطعه‌بندی پیشنهادی، شامل سه مرحله‌ی اصلی است. اولین مرحله، شامل انتخاب قله‌ها در سیگنال شتاب محور Y است؛ زیرا به بهترین شکل شروع و پایان تمامی فعالیت‌های مورد مطالعه را نشان می‌دهد. قله‌ها بر اساس آستانه‌ی اول و فاصله انتخاب شدند که به ترتیب نشان‌دهنده‌ی حداقل مقدار یک قله و حداقل فاصله‌ی بین قله‌ها هستند. مرحله‌ی دوم، انتخاب دره‌هاست. در نهایت، ویژگی‌های سیگنال هر قله برای شناسایی مرزهای قطعه‌بندی مناسب مورد تجزیه و تحلیل قرار می‌گیرد.

در مرحله‌ی اول قله‌هایی تشخیص داده شدند که دارای بیشینه‌ی محلی در سیگنال شتاب محور Y بودند. برای جلوگیری از درج قله‌های مثبت کاذب، طبق شکل ۴-۲ (الف)، از یک میزان آستانه استفاده شد. ارزش قله‌ی شناسایی شده باید برابر یا بزرگ‌تر از آستانه اول باشد. این را می‌توان با گرفتن میانگین جداگانه قله‌ها در مجموعه داده یادگیری همه‌ی فعالیت‌ها و انتخاب حداقل مقدار از بین آنها به شرح زیر محاسبه کرد:

$$\text{رابطه‌ی ۴-۲} \quad \text{Threshold 1} = \min (avg_{max}(PA_1), avg_{max}(PA_2), \dots, avg_{max}(PA_n))$$

در رابطه‌ی بالا avg، نشان‌دهنده‌ی میانگین بیشینه‌ی محلی در محور (Y) است. PA به فعالیت و n به تعداد فعالیت‌ها اشاره دارد. قله‌های شناسایی شده با در نظر گرفتن آستانه‌ی اول در شکل ۴-۲ (ب) نشان‌داده شده است.



شکل ۴-۲: قله‌های شناسایی شده در سیگنال شتاب: (الف) انتخاب اولیه‌ی قله‌ها. (ب) پس از اعمال آستانه‌ی اول و حداقل فاصله [۱۳].

برای جلوگیری از تشخیص بیش از یک قله در نقاط داده‌ای که یک فعالیت را نشان می‌دهند، حداقل فاصله-ی بین قله‌ها مطابق شکل ۴-۲ (الف) اعمال شد. این مقدار را می‌توان با محاسبه‌ی میانگین مدت زمان مورد نیاز برای انجام کوتاه‌ترین فعالیت به صورت زیر به دست آورد:

$$\text{رابطه‌ی ۵-۲} \quad \text{Distance} = \min (avg_{duration}(PA_1), avg_{duration}(PA_2), \dots, avg_{duration}(PA_n))$$

علاوه بر قله‌های به‌دست‌آمده از مرحله‌ی اول، این روش نیاز به شناسایی دره‌ها (کمینه‌های محلی) در سیگنال شتاب محور  $Y$  دارد. انتخاب دره‌ها طبق رابطه‌ی ۲-۶ انجام می‌شود. آستانه‌ی دوم نشان دهنده‌ی بالاترین ارزش یک دره است. برای جلوگیری از تشخیص دره‌های مثبت کاذب از آستانه‌ی دوم استفاده می‌شود. یک دره، زمانی انتخاب می‌شود که کمتر یا مساوی آستانه‌ی دوم باشد. آستانه‌ی دوم را می‌توان با گرفتن میانگین مقادیر دره‌های مثبت واقعی در مجموعه داده‌های یادگیری و تکرار فرایند برای فعالیتی که از یک قله تشکیل شده است، به دست آورد. این فعالیت‌ها شامل انقباض شانه، خم شدن شانه، گسترش آرنج، خم شدن آرنج و چرخش خارجی هستند.

آستانه‌ی دوم را می‌توان با گرفتن میانگین از مقادیر دره‌های مثبت واقعی در مجموعه داده‌های یادگیری به دست آورد. برای شناسایی فعالیت‌هایی که از یک قله تشکیل شده‌اند، از آستانه‌ی دوم استفاده می‌شود. ابتدا میانگین ارزش داده‌ها در مجموعه داده‌های یادگیری برای هر یک از فعالیت‌ها محاسبه می‌شود؛ سپس حداکثر مقدار از بین آنها به عنوان آستانه‌ی دوم طبق رابطه‌ی زیر به دست می‌آید. در این رابطه  $\min$  کمینه‌ی محلی در محور پردازش ( $Y$ ) بوده و  $m$  تعداد فعالیت‌های فیزیکی است که از یک قله در هر تکرار تشکیل شده‌اند.

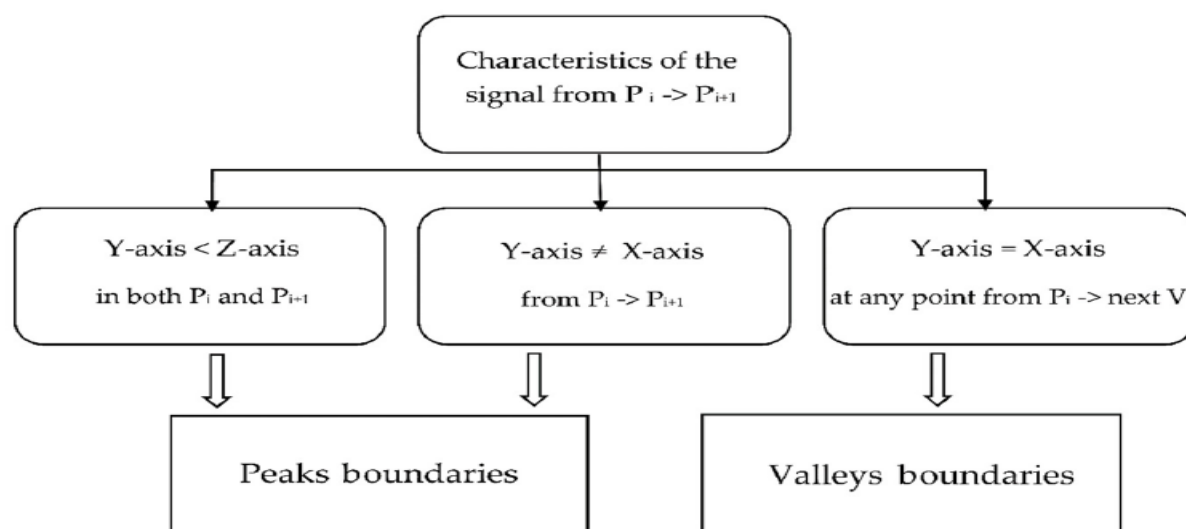
$$\text{Threshold 2} = \text{maximum} (avg_{\min}(PA_1), avg_{\min}(PA_2), \dots, avg_{\min}(PA_m)) \quad \text{رابطه‌ی ۲-۶}$$

#### • تعیین مرزهای قطعه

در قطعه‌بندی فعالیت پویا تعیین مرزهای زمانی قطعه، در بین فعالیت‌های مختلف ضروری است. دو نوع مرز برای فعالیت‌ها بر اساس تعداد قله‌ها در هر فعالیت وجود دارد.

نوع اول مرزهای قله است و زمانی استفاده می‌شود که یک فعالیت مانند چرخش داخلی و گسترش دو قله وجود داشته باشد. این قله‌ها به عنوان مرزهای قطعه در نظر گرفته می‌شوند. نوع دوم، مرزهای دره است و زمانی اعمال می‌شود که یک فعالیت فقط از یک قله تشکیل شود. دو دره که مستقیماً قبل و بعد از هر قله قرار دارند، به ترتیب به عنوان نقطه‌ی شروع و پایان قطعه شناسایی می‌شوند؛ بنابراین طول یک قطعه

به صورت پویا با توجه به مدت زمان فعالیت تغییر می‌کند. برای تعیین نوع مرزها برای قطعه‌بندی، ویژگی‌های سیگنال هر قله مطابق شکل ۵-۲ بررسی می‌شود.



شکل ۵-۲: تعیین نوع مرز بر اساس مشخصات سیگنال (P: قله، V: دره) [۱۳].

مطابق شکل ۵-۲ مرزهای قله در صورتی انتخاب می‌شوند که مقدار محور Y کوچک‌تر از مقدار محور Z در قله‌ی فعلی و قله‌ی بعدی باشد. ویژگی‌های سیگنال بین هر دو قله‌ی متوالی بررسی می‌شود. اگر هیچ تقاطعی بین محور Y و محور X در امتداد این قله‌ها وجود نداشته باشد، مرزهای قله اعمال می‌شود. اگر در هر نقطه بین قله و دره‌ی بعد از آن تقاطع وجود داشته باشد، از مرزهای دره استفاده می‌شود. شبه کدی که نحوه‌ی قطعه‌بندی سیگنال شتاب فعالیت‌ها را توضیح می‌دهد، در جدول ۱-۲ نشان داده شده است. مقادیر ورودی برای قطعه‌بندی پویا در خط ۲ و مقدار خروجی در خط ۵ نشان داده شده است.

حلقه‌ی «for» در خطوط ۹-۲۸ نشان‌دهنده‌ی روند تعیین نوع مرزها در همه قله‌ها به جز آخرین قله است. در خطوط ۱۰-۱۴، ویژگی‌های سیگنال در هر قله بررسی می‌شود. خطوط ۱۵-۲۳، قطعه‌بندی فعالیت‌ها با مرزهای قله یا دره را نشان می‌دهد. در خطوط ۲۸-۳۶ آخرین قله با استفاده از مرزهای دره شناسایی می‌شود.

جدول ۱-۲: شبه کد الگوریتم قطعه‌بندی پویا [۱۳].

الگوریتم ۱: قطعه‌بندی پویا

```
01 Input:
02 S: a set of tri-axial accelerometer data
03 Output:
04 A set of segments:  $Seg = \{seg_1, seg_2, \dots, seg_n\}$ 
05 Peaks = indices of all peaks in Y-axis using Threshold1 and Distance
06 valleys = indices of all valleys in Y-axis using Threshold2
07 p = total number of peaks
08 v = total number of valleys
09 for i = 0 to p-2 do
10   If Y-axis value is smaller than Z-axis value at peak(i) and peak(i+1)
11   Or no intersection between X-axis and Y-axis from peak(i) to peak(i+1) then
12     for h = peaks(i) to peaks(i+1) do
13       Add S (h) to Segi
14     end for
15   else if intersection exists between X-axis and Y-axis at any point from peak(i) to
       next valley then
16     for k = 1 to v-1 do
17       If valleys(k) is the valley that directly follows peaks(i)
18       AND valleys(k-1) is the valley that directly precedes peaks(i) then
19         for h = valleys(k-1) to valleys(k) do
20           Add S (h) to Segi
21         end for
22       end if
23     end for
```

```

24  else
25  i = i+1
26  end if
27  end for
28  for k = 1 to v-1 do
29  If valleys(k) is the valley that directly follows peaks(p-1)
30  AND valleys(k-1) is the valley that directly precedes peaks(p-1)
31  AND intersection exists between X-axis and Y-axis from peaks(p-1) to valleys(k)
    then
32  for h = valleys(k-1) to valleys(k) do
33  Add S (h) to Segp-1
34  end for
35  end if
36  end for

```

در این آزمایش، ۱۰ فرد سالم (۳ مرد و ۷ زن) بین ۲۵-۵۰ سال برای انجام فعالیت‌ها انتخاب شدند. قبل از آزمایش، همه‌ی شرکت‌کنندگان یک فرم رضایت را آگاهانه امضا کردند.

قبل از شروع جلسه‌ی فعالیت به شرکت‌کنندگان توصیه‌ها و دستورالعمل‌هایی برای اجرای صحیح تمرین‌ها داده شد. قبل از هر فعالیت، یک ویدئوی نمایشی کوتاه برای یادآوری چگونگی عملکرد بهینه نمایش داده شد. از هر شرکت‌کننده خواسته شد تا ۱۰ تکرار از تمام فعالیت‌ها (مجموعاً ۷۰۰ تکرار) را انجام دهد. علاوه بر این، از آنها خواسته شد که تکرارهای یک فعالیت مشابه را با فاصله زمانی تقریباً ۱۰ ثانیه جدا کنند تا شروع هر گروه جدید از تکرارها را مشخص کنند.

داده‌های خام به دست آمده از حسگرهای پوشیدنی، مانند شتاب‌سنج‌ها، دارای نویز و خطا هستند. پیش-پردازش یک مرحله‌ی ضروری برای مدل‌سازی و پیش‌بینی است. در این تحقیق، پیش‌پردازش در دو مرحله اجرا می‌شود:

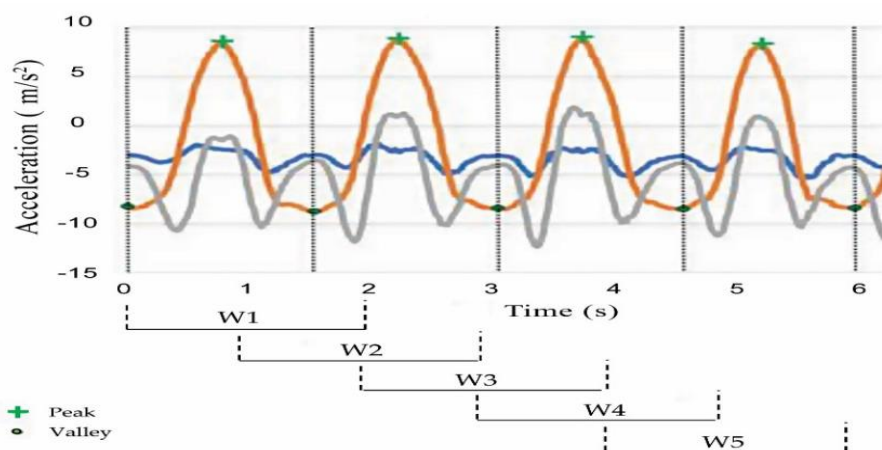
در مرحله‌ی اول، یک فیلتر میانگین متحرک<sup>۱</sup> برای حذف نویز با فرکانس بالا اعمال شد. این فرایند معادل فیلتر پایین گذر است. نحوه‌ی انتخاب طول بهینه، یکی از جنبه‌های مهم فیلتر میانگین متحرک است. مقادیر مختلف برای طول فیلتر می‌تواند بر عملکرد شناسایی تأثیر بگذارد. در تحقیق اخیر با آزمایش مقادیر مختلف این نتیجه گرفته شد که طول ۱۰ برای فیلتر منجر به تولید داده‌هایی با نویز کمتر بدون از دست دادن اطلاعات کلیدی می‌شود.

در مرحله‌ی دوم، داده‌های نامطلوب حذف شد. از شرکت‌کنندگان خواسته شد تا هر گروه از فعالیت‌ها را با زمانی به اندازه‌ی تقریباً ۱۰ ثانیه جدا کنند؛ بنابراین بخشی از داده‌های جمع‌آوری شده با زمانی مطابقت دارد که هیچ فعالیت‌ای انجام نشده است. این بخش‌ها به صورت دستی حذف شدند. شایان ذکر است که این مرحله فقط برای مجموعه داده‌های یادگیری انجام شد. به بیماران واقعی چنین دستورالعمل‌هایی داده نشد. از آنجایی که روش پیشنهادی، مبتنی بر تشخیص قله‌ها و دره‌هاست، وجود این بخش‌ها بر عملکرد الگوریتم تأثیری نخواهد داشت.

در تشخیص فعالیت، از دو روش پنجره‌ی کشویی و قطعه‌بندی مبتنی بر قله‌ها و دره‌ها برای مقایسه استفاده شد. روش پنجره کشویی با اندازه ثابت به طول ۲ ثانیه و هم‌پوشانی ۵۰ درصد، دارای بالاترین دقت تشخیص بود. روش دوم، روش قطعه‌بندی پیشنهادی بود. شکل ۲-۶ نتیجه‌ی استفاده از هر دو روش را برای قطعه-بندی سیگنال شتاب خم کردن آرنج نشان می‌دهد.

<sup>۱</sup> Moving average filter (MAF)





شکل ۲-۶: سیگنال شتاب چهار تکرار خم شدن آرنج که با استفاده از پنجره‌ی کشویی (قسمت پایین) و روش پیشنهادی (خطوط نقطه‌چین) قطعه‌بندی شده‌اند [۱۳].

برای تولید داده‌هایی که می‌توانند برای الگوریتم یادگیری ماشین مناسب باشند، ویژگی‌های چندگانه مانند ویژگی‌های دامنه‌ی زمان، فرکانس و ویژگی‌های اکتشافی از تمام قطعه‌های به‌دست‌آمده با استفاده از هر دو روش محاسبه شد.

برای تشخیص فعالیت‌های انسانی مبتنی بر حسگر، استفاده از ویژگی‌های حوزه‌ی زمانی رایج است و دلیل آن سادگی و اثربخشی بالای آن است. در این روش، فقط از ویژگی‌های حوزه‌ی زمان استفاده شده است، زیرا ویژگی‌های حوزه‌ی فرکانس، نیازمند محاسبات و حافظه بالایی هستند. استفاده از ویژگی‌های حوزه‌ی فرکانس در برنامه‌های بلادرنگ و کم مصرف غیرممکن است.

به دلیل استفاده‌ی مکرر در مطالعات قبلی، تشخیص فعالیت از ماشین بردار پشتیبان برای آموزش و آزمایش مدل طبقه‌بندی استفاده شد. مدل اعتبارسنجی متقابل ده برابری<sup>۱</sup> برای آموزش به کار برده شد. داده‌های ۹ نفر به طور تصادفی به مجموعه‌های آموزشی و آزمایشی به ترتیب با استفاده از ۹۰ و ۱۰ درصد داده‌ها تقسیم شد.

<sup>1</sup> Ten-fold cross validation

## ۲-۴-۲ مقاله‌ی دوم

در مقاله‌ی [۱۴]، روش‌های جدید قطعه‌بندی با استفاده از مجموعه داده‌های چند ساکن<sup>۱</sup> CASAS ارزیابی شد و بهبود عملکرد تشخیص فعالیت چند ساکن<sup>۲</sup> با استفاده از تکنیک‌های قطعه‌بندی تأیید شد. سپس یک روش قطعه‌بندی مبتنی بر تشخیص نقطه‌ی تغییر<sup>۳</sup> و خوشه‌بندی فازی<sup>۴</sup> پیشنهاد شد که می‌تواند عملکرد تشخیص فعالیت چند ساکن را افزایش دهد.

برای قطعه‌بندی رویداد حسگر، یک روش خوشه‌بندی فازی با یک روش مبتنی بر تشخیص نقطه‌ی تغییر ترکیب می‌شود. روش خوشه‌بندی فازی برای طبقه‌بندی رویدادهای حسگر استفاده می‌شود و سپس تکنیک تشخیص نقطه‌ی تغییر برای بررسی انتقال فعالیت<sup>۵</sup> و تعیین توالی قطعه‌بندی به کار می‌رود. روش پیشنهادی به طور قابل توجهی عملکرد تشخیص فعالیت چند ساکن را در مقایسه با روش‌های قطعه‌بندی پایه و پیشرفته بهبود می‌بخشد.

قطعه‌بندی داده‌های حسگر، می‌تواند عملکرد تشخیص فعالیت را بهبود ببخشد. علاوه بر این، قطعه‌بندی رویدادهای حسگر، مشاهدات مناسب بیشتری را با شناسایی شروع و پایان فعالیت‌ها ارائه می‌دهد که الگوهای رفتاری و ایجاد برخی ویژگی‌های جدید مانند مدت زمان فعالیت را مدیریت می‌کند.

برای تشخیص فعالیت چند ساکن، نه تنها باید نقطه‌ی شروع و پایان فعالیت تشخیص داده شود، بلکه باید مشخص شود کدام ساکن باعث ایجاد رویداد حسگر شده است. شکل ۲-۷ نمای کلی از ساختار قطعه‌بندی فعالیت چند ساکن پیشنهادی را نشان می‌دهد که شامل چهار مرحله است، اولین قدم، استفاده از انواع حسگرهای محیطی برای جمع‌آوری داده‌های حسگر در خانه‌ی هوشمند است که چند نفر در آن ساکن هستند.

<sup>1</sup> Multi-resident

<sup>2</sup> Multi-resident activity recognition

<sup>3</sup> Change point detection (CPD)

<sup>4</sup> Fuzzy c-means (FCM)

<sup>5</sup> Activity transition

در مرحله‌ی دوم، همه‌ی حسگرها بر اساس مکان‌هایشان خوشه‌بندی شده‌اند و داده‌ها با استفاده از روش خوشه‌بندی فازی بر اساس مکان حسگرها به زیرمجموعه‌های مختلف تقسیم شده است. مرحله‌ی سوم، شناسایی انتقال فعالیت در زیرمجموعه‌های مختلف و قطعه‌بندی است. مرحله‌ی نهایی تشخیص فعالیت چند ساکن با استفاده از مدل‌های طبقه‌بندی در یادگیری ماشین و یادگیری عمیق پایه و پیشرفته مطابق شکل ۲-۷ است. چهار تکنیک قطعه‌بندی برای تشخیص فعالیت چند ساکن ارزیابی شده است.

اولین رویکرد، پنجره‌ی کشویی مبتنی بر زمان ایستا است در این روش رویدادهای حسگر بر اساس یک مدت زمان ثابت تقسیم می‌شوند.

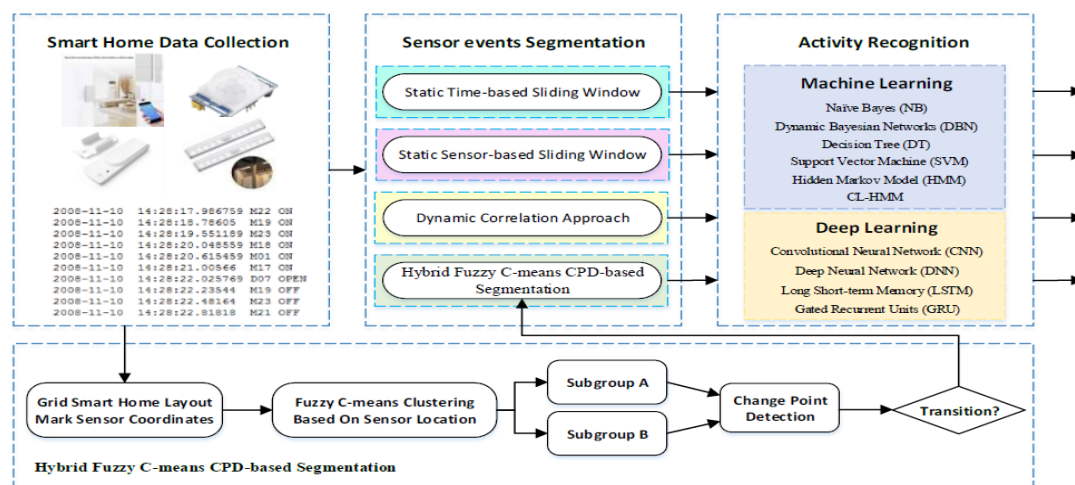
رویکرد دوم، پنجره‌ی کشویی مبتنی بر حسگر ایستا است، که در آن رویدادهای حسگر به قطعه‌هایی با تعداد مشخص و ثابتی از رویدادها تقسیم می‌شوند.

سومین رویکرد از زمان پویا و همبستگی حسگر استفاده می‌کند. در این روش همبستگی محصول-لحظه‌ی پیرسون<sup>۱</sup> بین جفت رویدادهای حسگر محاسبه می‌شود. مدت زمان مربوط به هر قطعه از رویدادهای حسگر متغیر است.

در رویکرد چهارم، رویدادهای حسگر با خوشه‌بندی مکان حسگر به زیرمجموعه‌های مختلف تقسیم می‌شود. سپس یک روش تشخیص نقطه‌ی تغییر در هر زیرگروه اعمال می‌شود تا رویدادهای حسگر بین دو انتقال فعالیت را به‌عنوان یک فعالیت تشخیص دهد.

---

<sup>۱</sup> Pearson product-moment correlation (PMC)



شکل ۲-۷: تشخیص فعالیت چند ساکن [۱۴].

خوشه‌بندی یکی از شاخه‌های یادگیری بدون نظارت است. در این فرایند، نمونه‌ها به خوشه‌هایی که اعضای آن مشابه یکدیگر هستند، تقسیم می‌شوند؛ بنابراین خوشه، مجموعه‌ای از اشیاء است که اشیاء موجود در آن با یکدیگر مشابه بوده و با اشیاء موجود در خوشه‌های دیگر مشابه نیستند. برای مشابه بودن می‌توان معیارهای مختلفی را در معیار قرار داد. می‌توان معیار فاصله را برای خوشه‌بندی در نظر گرفت و اشیایی که به یکدیگر نزدیک‌تر هستند را به‌عنوان یک خوشه مدنظر قرار داد. به این نوع خوشه‌بندی، خوشه‌بندی مبتنی بر فاصله نیز گفته می‌شود.

الگوریتم خوشه‌بندی فازی بر اساس بهینه‌سازی تابع هدف با استفاده از تئوری فازی عمل می‌کند. درجه‌ی عضویت هر نقطه داده در مرکز خوشه‌بندی با یک عدد نشان داده می‌شود. با بهینه‌سازی تابع هدف، درجه عضویت هر نقطه‌ی نمونه در همه مراکز خوشه به دست می‌آید. یک شیء می‌تواند به چندین خوشه تعلق داشته باشد.

برای درک بهتر خوشه‌بندی فازی<sup>۱</sup> مفهوم مجموعه‌های فازی و تفاوت آنها با مجموعه‌های کلاسیک بررسی می‌شود. می‌توان برای هر مجموعه‌ی کلاسیک، یک تابع تعلق تعریف کرد که مقدار این تابع تعلق برای اعضای مجموعه ۱ و برای بقیه ۰ است. در مجموعه‌های کلاسیک، مقدار آن ۰ یا ۱ است.

در اینجا با مفهوم عدم قطعیت مواجه هستیم. عدم قطعیت در زندگی روزمره نیز وجود دارد. هوای سرد و آب داغ، مثال‌هایی از مجموعه‌های فازی هستند. تفاوت اصلی مجموعه‌های فازی و مجموعه‌های کلاسیک این است که تابع تعلق مجموعه‌های فازی، دو مقدار ۰ یا ۱ نیست، بلکه می‌تواند هر مقداری بین ۰-۱ را اختیار کند. اگر اعضای یک مجموعه فازی تنها دارای تابع تعلق ۰ و ۱ باشند، این مجموعه‌ی فازی یک مجموعه کلاسیک خواهد بود. یک عضو مجموعه‌ی مرجع می‌تواند با درجه‌ی عضویت‌های مختلف، عضو مجموعه‌های فازی تعریف شده روی مجموعه‌ی مرجع باشد.

روش خوشه‌بندی فازی نتایج انعطاف‌پذیرتری را در مقایسه با خوشه‌بندی سخت سنتی<sup>۲</sup> ارائه می‌کند، زیرا اشیا می‌توانند به خوشه‌های متفاوت تعلق داشته باشند. در بیشتر موارد، اشیای یک مجموعه‌داده را نمی‌توان به خوشه‌های کاملاً مجزا تقسیم کرد؛ بنابراین در خوشه‌بندی فازی به هر شیء و هر خوشه، وزنی داده می‌شود که نشان‌دهنده‌ی درجه‌ی عضویت شیء به خوشه است. تابع عضویت  $\mu_A(X)$  تابعی است که درجه‌ی عضویت یک شیء  $X$  به مجموعه‌ی  $C$  را نشان می‌دهد و مقدار ۰-۱ دارد.

الگوریتم خوشه‌بندی باید مقدار تابع هدف را به حداقل برساند. اگر مقدار درجه‌ی عضویت تغییر زیادی نکند، الگوریتم پایان می‌پذیرد، زیرا به بهینه‌ی محلی (نسبی) رسیده است.

در محیط‌هایی که چند ساکن وجود دارد یک آستانه‌ی ساده اعمال نمی‌شود. خوشه‌بندی فازی رویدادهای حسگر را براساس مکان آن‌ها در خوشه‌های مختلف قرار می‌دهد.

هر رویداد حسگر به شکل  $e = \langle d, t, s, v \rangle$  است،  $d$  تاریخ جمع‌آوری داده‌های حسگر،  $t$  برچسب زمانی در این تاریخ،  $s$  حسگر و  $v$  مقدار حسگر است.

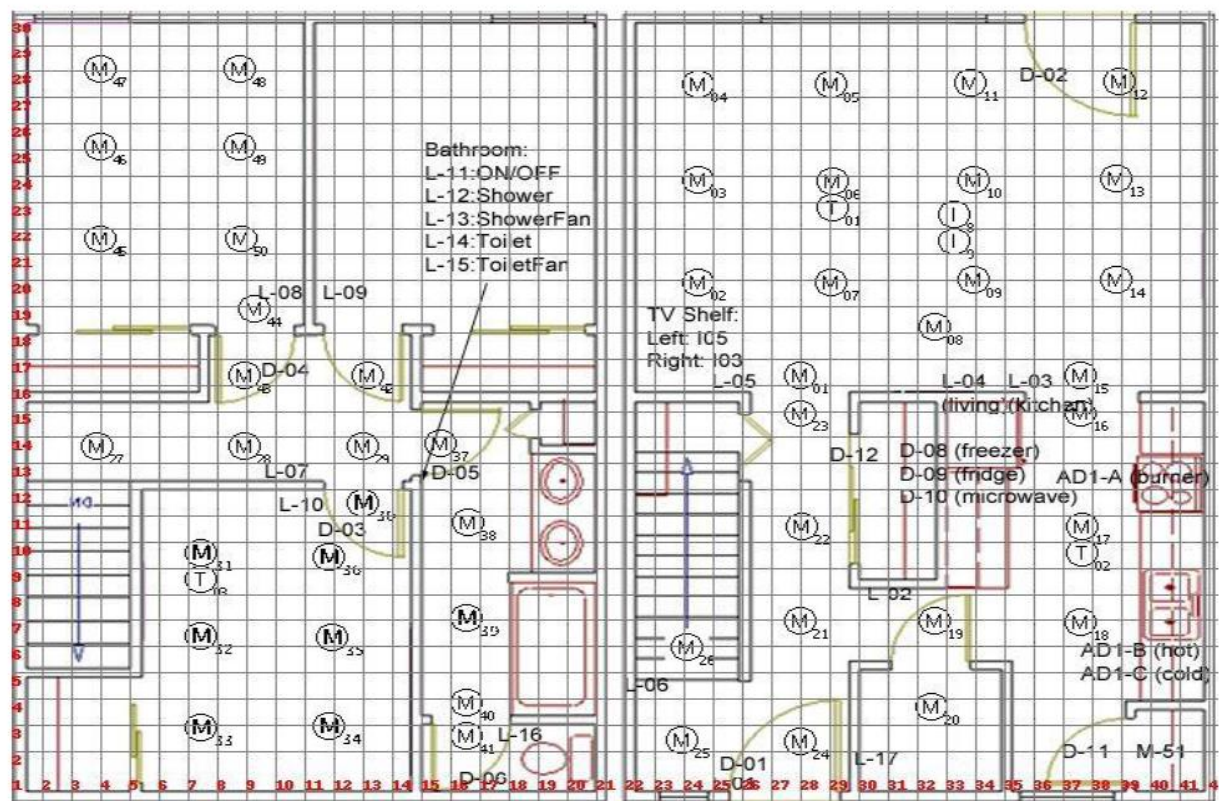
<sup>۱</sup> Clustering

<sup>۲</sup> K-means

فرض می‌شود که فعالیت‌های موازی در مکان‌های مختلف رخ می‌دهند. یک مکان می‌تواند تنها توسط یک ساکن اشغال شود یا فعالیت‌های مشترک در آن انجام شود. برای تشخیص فعالیت‌ها از رویدادهای حسگر پیوسته، چند دنباله از رویدادهای حسگر تولید می‌شود و هر فعالیت به‌عنوان دنباله‌ای از  $n$  رویداد حسگر  $\langle e_1, e_2, \dots, e_n \rangle$  در نظر گرفته می‌شود. قرار گرفتن رویدادهای حسگر در این دنباله‌ها بر اساس خوشه‌بندی فازی آن‌هاست.

مختصات حسگر مطابق شکل ۲-۸ به‌عنوان یک ویژگی معرفی شده است. روش خوشه‌بندی فازی برای اختصاص رویدادهای حسگر به دنباله‌های مختلف به کار می‌رود. این رویدادها با  $e = \langle t, s, v, c \rangle$  مشخص شده‌اند که در آن دو رویداد حسگر در زمان‌های  $t_i$  و  $t_{i+1}$  مقایسه شده است.  $c$  نشان‌دهنده‌ی خوشه‌ی حسگرها است.

اگر مکان دو حسگر به دو خوشه‌ی مختلف تعلق داشته باشد، دو رویداد حسگر به دنباله‌های مختلف اختصاص داده می‌شود، در غیر این صورت رویدادهای حسگر به‌عنوان یک دنباله در نظر گرفته می‌شود. به‌عنوان مثال، در زمان  $t_i$  تا  $t_{i+5}$ ، به ترتیب حسگرهای  $M_{18}, M_{19}, M_{21}, M_{19}, M_{20}$  را فعال می‌کند. نتایج خوشه‌بندی فازی نشان می‌دهد که هر پنج مکان حسگر به یک خوشه تعلق دارند. پس یک ساکن واحد این حسگرها را فعال کرده است یا چند ساکن فعالیت‌ی مشترک را در یک مکان انجام داده‌اند. این حسگرها در یک دنباله قرار می‌گیرند. در زمان  $t_{i+6}$  حسگر  $M_{05}$  فعال می‌شود. حسگرهای  $M_{05}$  و  $M_{05}$  در خوشه‌های متفاوتی هستند. پس امکان ندارد یک ساکن فقط این حسگر را فعال کند؛ بنابراین این رویداد حسگر به دنباله‌ی دیگری تعلق دارد. اگر حسگرهای  $S_i$  و  $S_{i+1}$  در یک خوشه نباشند، رویدادهای آن‌ها به زیرگروه‌های مختلف تقسیم می‌شوند.



شکل ۲-۸: خوشه‌بندی فازی بر اساس مکان حسگرها [۱۴].

## ۲-۵ معرفی مجموعه داده‌ها

### ۲-۵-۱ SKODA

این مجموعه داده شامل ۱۰ حرکت انجام شده در سناریوهای تعمیر و نگهداری خودرو است. ۲۰ حسگر در بازوی چپ و راست فرد قرار داده شده است. همچنین شامل ویژگی‌های شتاب ۳ بعدی برای هر حسگر است [۱۵]. هر حرکت حدود ۷۰ بار تکرار می‌شود.

در این مجموعه داده ۳۱ کلاس فعالیت، انتخاب شده است. تمام کلاس‌های فعالیت شامل حرکت دست(ها) یا بازو(های) کاربر است که با استفاده از یک مجموعه‌ی توزیع شده از ۵ حسگر اینرسی گرفته شده است. این مژول‌ها در پوشش فرد ادغام شده و با بالاتنه تراز شده‌اند. در مقاله‌ی اصلی تنها زیرمجموعه‌ای از شش

کلاس فعالیت شامل باز کردن کاپوت موتور، بستن درب موتور، باز کردن صندوق عقب، بررسی لولاهای صندوق عقب، بستن صندوق عقب و بررسی درپوش باک سوخت برای تجزیه و تحلیل انتخاب شده است.

## ۲-۵-۲ WISDM

WISDM با استفاده از شتاب‌سنج‌های تعبیه‌شده در تلفن‌ها جمع‌آوری می‌شود. یک تلفن در جیب‌های جلوی شلوار هر فرد قرار داده شده است. شش فعالیت منظم پیاده‌روی، دویدن، بالارفتن از پله، پایین آمدن از پله، نشستن و ایستادن در نظر گرفته شده است [۱۶].

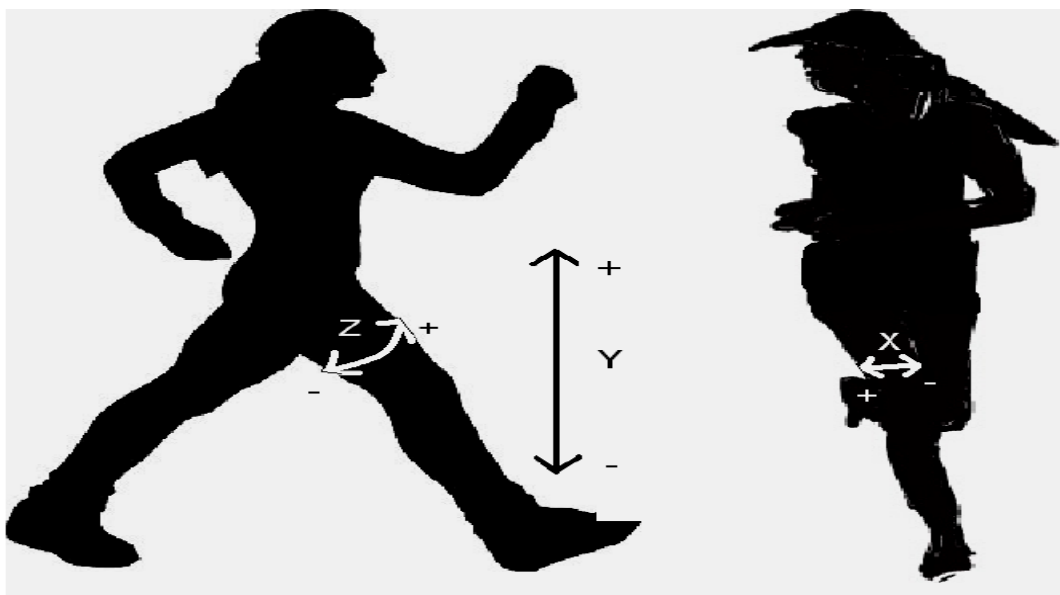
برای جمع‌آوری داده‌ها برای یادگیری تحت نظارت، لازم بود تعداد زیادی از کاربران، تلفن هوشمند مبتنی بر اندروید را در حین انجام برخی فعالیت‌های روزمره همراه داشته باشند. سپس از ۲۹ داوطلب برای حمل تلفن هوشمند در حین انجام مجموعه‌ای از فعالیت‌ها کمک گرفته شد. این افراد تلفن اندرویدی را در جیب جلویی شلوار خود حمل کردند و از آنها خواسته شد برای دوره‌های زمانی مشخصی راه بروند، آهسته بدونند، از پله‌ها بالا بروند، از پله‌ها پایین بیایند، بنشینند و بایستند.

جمع‌آوری داده‌ها توسط یک برنامه روی تلفن همراه کنترل شده است. این برنامه، از طریق یک رابط کاربری گرافیکی ساده، اجازه می‌دهد تا نام کاربر ثبت شود، جمع‌آوری داده‌ها را شروع و متوقف می‌کند و فعالیت‌های انجام شده را با برچسب‌گذاری مشخص می‌کند. این برنامه اجازه می‌دهد نوع داده‌های جمع‌آوری شده از حسگرهایی مانند سامانه‌ی موقعیت‌یاب جهانی و شتاب‌سنج کنترل شده و تعداد دفعات آن را جمع‌آوری می‌کند. در همه‌ی موارد، داده‌های شتاب‌سنج هر ۵۰ میلی‌ثانیه جمع‌آوری می‌شود؛ بنابراین ۲۰ نمونه در ثانیه به وجود می‌آید. جمع‌آوری داده‌ها توسط یکی از اعضای تیم WISDM برای اطمینان از کیفیت داده‌ها نظارت می‌شود.

در این مطالعه شش فعالیت راه‌رفتن، دویدن، بالارفتن از پله، پایین آمدن از پله، نشستن و ایستادن در نظر گرفته شده است. این فعالیت‌ها به دلیل اجرای منظم توسط تعداد زیادی از افراد در روال روزانه انتخاب شده‌اند. این فعالیت‌ها شامل حرکاتی است که اغلب برای دوره‌های زمانی قابل توجهی رخ می‌دهد، بنابراین تشخیص آنها آسان‌تر می‌شود. علاوه بر این، بیشتر این فعالیت‌ها شامل حرکات تکراری است که تشخیص



آن‌ها را آسان‌تر می‌کند. داده‌های شتاب هر یک از این فعالیت‌ها در سه محور ثبت می‌شود. محور  $Tz$  حرکت جلوی پا و محور  $Y$ ، حرکت رو به بالا و پایین را ثبت می‌کند. محور  $X$ ، حرکت افقی پای کاربر را ثبت می‌کند. شکل ۲-۹ این محورها را نسبت به یک کاربر نشان می‌دهد.

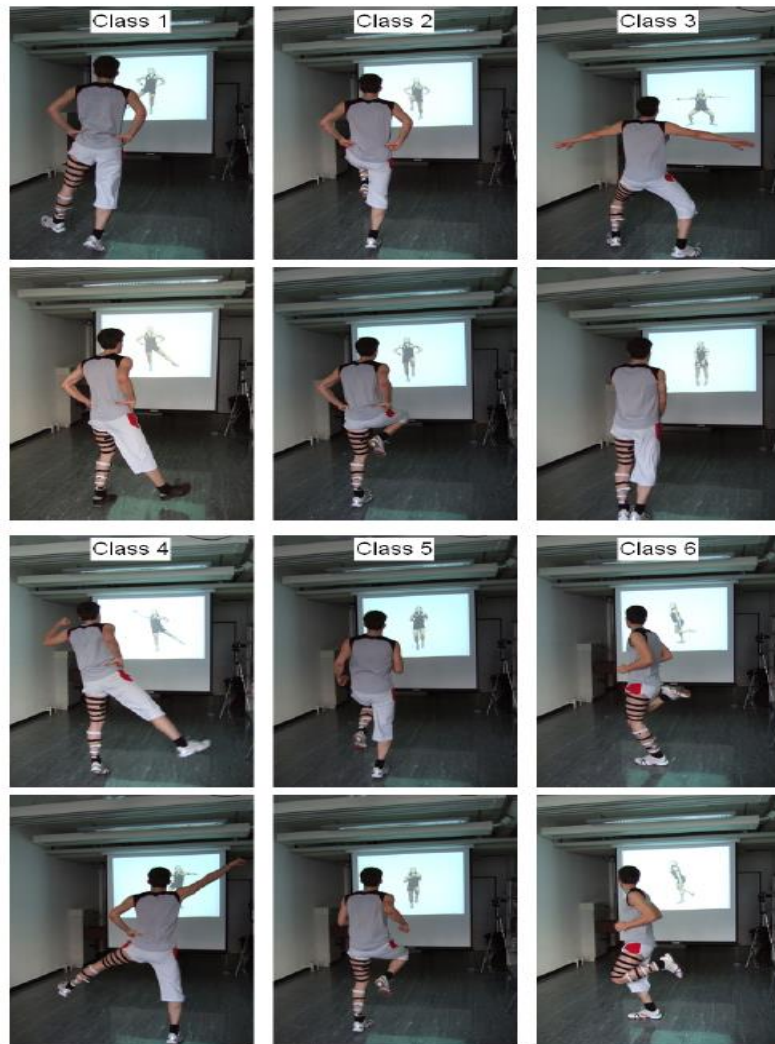


شکل ۲-۹: شتاب در محورهای مجموعه داده‌ی WISDM [۱۶].

## HCI ۳-۵-۲

HCI بر تغییرات ناشی از جابه‌جایی حسگرها تمرکز دارد. حرکات شامل حرکات بازو یا دست است که اشکال مختلف مثلث رو به بالا، مثلث وارونه و دایره را توصیف می‌کند. هشت حسگر به بازوی پایین سمت راست هر فرد متصل شده است [۱۷]. هر حرکت برای بیش از ۵۰ تکرار و هر تکرار برای ۵ تا ۸ ثانیه ضبط می‌شود.

در این مجموعه داده، شتاب پای چپ برای شش حرکت هوازی معمولی مختلف مطابق شکل ۲-۱۰ ثبت شده است.



شکل ۲-۱۰: شتاب پای چپ برای شش حرکت هوازی معمولی در مجموعه داده‌ی HCI [۱۷].

سناریوی تناسب‌اندام شامل ۶ کلاس ضربات تلنگر، بلندکردن زانو، جک‌های پرش، پرش سوپرمن، دویدن از روی زانو و دویدن به عقب است. برای هر کلاس، نحوه‌ی حرکات بدن در دو ردیف نشان داده شده است. برای ایجاد این مجموعه داده، ۱۰ حسگر شتاب بلوتوث مطابق شکل ۲-۱۱ در پای فرد قرار داده می‌شود که ۵ حسگر در ساق پا و ۵ حسگر در ران است.



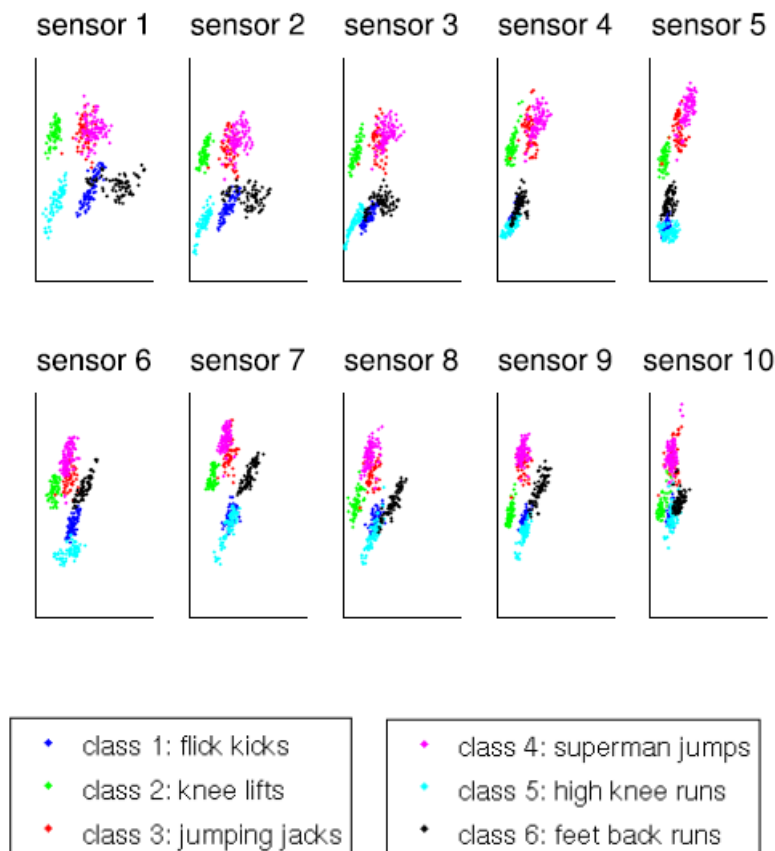
شکل ۲-۱۱: حسگرهای شتاب بلوتوث در مجموعه داده‌ی HCI [۱۷].

در سناریوی HCI (سمت چپ) هشت حسگر به بازوی راست متصل شده است که شامل شش حسگر در ساعد و دو حسگر روی بازو است. در سناریوی تناسب اندام (سمت راست) ۱۰ حسگر به پای چپ فرد وصل شده است که شامل پنج حسگر در ران و پنج حسگر دیگر در پایین ساق پا است.

برای داده‌های هر حسگر، بزرگی شتاب محاسبه شد و ویژگی‌های میانگین و واریانس بر اساس یک پنجره‌ی کشویی هشت ثانیه‌ای با دو سوم هم‌پوشانی استخراج گردید. فضای ویژگی دوبعدی حاصل برای هر موقعیت حسگر در شکل ۲-۱۲ نشان داده شده است. تفاوت کمتری در فضای ویژگی بین حسگرهایی که به طور مستقیم در مجاورت هم قرار دارند، مشاهده می‌شود. حسگرهایی که بیشتر از هم فاصله دارند، تفاوت بیشتری در فضای ویژگی دارند.

همچنین شباهت کمتری بین موقعیت حسگرهای ۵ و ۶ وجود دارد، حتی اگر در مجاورت هم قرار بگیرند، زیرا در بخش‌های مختلف یک اندام قرار دارند. برای مکان‌های ۴، ۵ و ۷-۱۰ می‌توان هم‌پوشانی‌های بسیار قوی را بین کلاس‌های «ضربه‌های تلنگر» و «حرکت بالای دو زانو» مشاهده کرد. همچنین هم‌پوشانی‌های

بسیار قوی بین کلاس‌های «جک‌های پرش» و «پرش‌های سوپرمن» برای همه موقعیت‌های حسگر وجود دارد.



شکل ۲-۱۲: کلاس‌های مختلف فعالیت‌ها در مجموعه داده‌ی HCI [۱۷].

PS ۴-۵-۲

اطلاعات PS با چهار گوشی هوشمند در چهار موقعیت بدن جمع‌آوری شده است [۱۸]. گوشی‌های هوشمند مجهز به شتاب‌سنج، مغناطیس‌سنج و ژيروسکوپ بودند. از چهار شرکت‌کننده خواسته شد تا شش فعالیت راه‌رفتن، دویدن، نشستن، ایستادن، راه‌رفتن به طبقه بالا و پایین را برای چند دقیقه انجام دهند.

ابتدا برنامه روی هر چهار گوشی هوشمند راه‌اندازی شد و سپس آن‌ها را در چهار وضعیت بدن یک شرکت‌کننده قرار دادند. پس از اتمام فعالیت، آن‌ها از بدن شرکت‌کننده حذف شدند و برنامه متوقف شد. این عمل باعث ایجاد نویز (خوشه‌های غیرطبیعی) در شروع و پایان هر فعالیت شد. این قسمت‌های پر از نویز قبل از تجزیه و تحلیل داده‌ها حذف شدند.

سپس داده‌های جمع‌آوری شده برای استخراج ویژگی با استفاده از رویکرد پنجره‌ی کشویی به قطعه‌های کوچک تقسیم شدند. انتخاب یک اندازه‌ی پنجره‌ی مناسب مهم است و می‌توان مقادیر مختلفی را برای آن تعیین کرد. پنجره‌ی کشویی ۲ ثانیه‌ای انتخاب شده است، زیرا اندازه‌ی پنجره‌ی ۲ ثانیه به‌عنوان یک مقدار مؤثر و کافی برای عملکرد تشخیص فعالیت معقول در نظر گرفته می‌شود.

هر حسگر مقادیر را در امتداد سه بعد، شامل محور  $x$ ، محور  $y$  و محور  $z$  گزارش می‌دهد. به‌عنوان مثال، شتاب‌سنج شتاب را بر حسب متر بر مجذور ثانیه ( $m/s^2$ )، مغناطیس‌سنج میدان مغناطیسی را به میکرو-تسلا ( $\mu T$ ) وژیروسکوپ سرعت چرخش را بر حسب رادیان بر ثانیه ( $rad/s$ ) در امتداد محورها گزارش می‌دهد. هر محور جهت‌گیری یک گوشی هوشمند بر عملکرد (دقت) الگوریتم‌های طبقه‌بندی تأثیر می‌گذارد زیرا مقدار حسگرهایی مانند شتاب‌سنج با توجه به جهت گوشی هوشمند تغییر می‌کند. اکثر کارهای موجود در حین ارزیابی الگوریتم‌های طبقه‌بندی مختلف، جهت‌گیری ثابتی را فرض می‌کنند. به همین دلیل بُعد چهارم به سه بعد موجود هر حسگر اضافه می‌شود که به آن بزرگی حسگر<sup>۱</sup> می‌گویند. بزرگی حسگر بر خلاف سه محور دیگر شتاب‌سنج وژیروسکوپ، به جهت‌گیری حساس نیست. بزرگی حسگر برای هر حسگر با استفاده از رابطه‌ی زیر محاسبه می‌شود:

$$Magnitude = \sqrt{x^2 + y^2 + z^2} \quad \text{رابطه‌ی ۲-۷}$$

<sup>۱</sup> Magnitude

اکنون چهار بعد شامل  $x$ ،  $y$ ،  $z$  و  $magnitude$  برای هر حسگر داریم. برای هر پنجره‌ی کشویی با ۵۰ درصد هم‌پوشانی، دو ویژگی حوزه‌ی زمانی میانگین و انحراف استاندارد برای چهار بعد هر حسگر استخراج شد؛ بنابراین در مجموع،  $4 \times 2 = 8$  ویژگی برای هر حسگر و  $8 \times 3 = 24$  ویژگی برای همه‌ی حسگرها محاسبه شد.

## ۲-۶ پیشینه‌ی پژوهش

در مقاله‌ی [۱۹] یک مدل مبتنی بر شبکه‌ی عصبی عمیق شامل شبکه‌ی عصبی کانولوشن<sup>۱</sup> و شبکه‌ی عصبی بازگشتی<sup>۲</sup> پیشنهاد شده است که استخراج ویژگی‌ها و طبقه‌بندی خودکار فعالیت‌ها را نیز انجام داده است. این داده‌ها از حسگرهای پوشیدنی مانند ژيروسکوپ و شتاب‌سنج به دست آمده‌اند و به شکل داده‌های سری زمانی هستند، زیرا هر رویداد، دارای یک برچسب زمانی است. برای تشخیص فعالیت، استخراج ویژگی‌های زمانی از داده‌های خام حسگرها مهم است. آزمایش‌ها بر روی داده‌های خام به دست آمده از حسگرهای پوشیدنی با پیش‌پردازش اسمی انجام شده است و شامل هیچ‌گونه تکنیک استخراج ویژگی دستی نیست. نتایج آزمایش‌ها نشان داده است که مدل پیشنهادی به عملکرد بهتری نسبت به سایر معماری‌های مشابه دست‌یافته است.

در مقاله‌ی [۲۰] یک چارچوب یکپارچه پیشنهاد شده که با نظارت ضعیف مبتنی بر تعبیه‌ی هسته‌ی<sup>۳</sup> توزیع‌ها همراه است و برای قطعه‌بندی جریان داده‌های حسگر، استخراج ویژگی از هر قطعه و آموزش یک طبقه‌بند نهایی برای تشخیص فعالیت طراحی شده است. قطعه‌بندی داده‌ها و استخراج ویژگی، دو مرحله‌ی مهم در توسعه‌ی مدل‌های مبتنی بر یادگیری ماشین هستند که برای تشخیص فعالیت مبتنی بر حسگر ضروری هستند. با استفاده از ویژگی‌های فوریه‌ی تصادفی، یک نسخه‌ی تسریع شده برای داده‌های حجیم ارائه شده است. آزمایش‌هایی روی چهار مجموعه داده‌ی (SKODA، WISDM، HCI و PS) انجام شده است تا کارایی و مقیاس‌پذیری چارچوب پیشنهادی تأیید شود.

<sup>۱</sup> Convolutional neural network (CNN)

<sup>۲</sup> Recurrent neural network (RNN)

<sup>۳</sup> Kernel

یک چارچوب سراسری برای تشخیص فعالیت مبتنی بر حسگر بیان می‌شود که قطعه‌بندی، استخراج ویژگی و همچنین طبقه‌بندی نهایی را امکان‌پذیر می‌سازد. جریان مداوم داده‌های فعالیت، ابتدا به قطعه‌هایی با نظارت ضعیف<sup>۱</sup> تقسیم می‌شود. یک تابع توزیع چگالی برای استخراج ویژگی‌های آماری کافی از هر قطعه و یادگیری یک طبقه‌بند برای تشخیص فعالیت استفاده می‌شود. کل فرایند به‌عنوان ماشین اندازه‌گیری پشتیبانی<sup>۲</sup> با نظارت ضعیف شناخته می‌شود. پارامترهای قابل یادگیری اجزای مختلف به طور مشترک به‌روزرسانی خواهد شد. علاوه بر این، از تکنیک ویژگی‌های فوری<sup>۳</sup> تصادفی<sup>۴</sup> برای توسعه‌ی یک نسخه‌ی تسریع شده در جهت مقابله با مسئله‌ی مقیاس‌پذیری استفاده می‌شود [۲۰].

در مقاله‌ی [۲۱] یک تکنیک جدید به نام قطعه‌بندی دودویی وحشی<sup>۴</sup> برای تخمین تعداد و مکان‌های نقاط تغییر در داده‌ها پیشنهاد شده است. تعداد نقاط تغییر می‌تواند با توجه به حجم نمونه تا بی‌نهایت افزایش یابد. قطعه‌بندی دودویی وحشی بر خلاف قطعه‌بندی دودویی<sup>۵</sup> استاندارد، به دلیل مکانیزم محلی‌سازی تصادفی، حتی برای نقاط تغییر با فاصله‌ی کم یا اندازه‌ی پرش‌های بسیار کوچک کار می‌کند. از سوی دیگر، علی‌رغم استفاده از محلی‌سازی، قطعه‌بندی دودویی وحشی به انتخاب پارامتر پنجره یا دهانه نیازی ندارد.

مقاله‌ی [۲۲] به مشکل تشخیص نقاط تغییر برون‌خط می‌پردازد. مدل‌های تشخیص نقاط تغییر به‌عنوان مدل‌های ساختار پنهان در نظر گرفته می‌شوند و مربوط به فضای قطعه‌بندی پنهان هستند. فضای قطعه-بندی‌های یک دنباله، برای تعداد ثابتی از نقاط تغییر را به دو دسته‌ی شمارش قطعه‌بندی‌ها و خلاصه‌ای از قطعه‌بندی‌های ممکن در نمایه‌های نقطه‌ی تغییر یا قطعه تقسیم می‌کند.

در دسته‌ی اول، یک الگوریتم برنامه‌نویسی پویا برای محاسبه‌ی قطعه‌بندی‌های محتمل به دست آمده است. در دسته‌ی دوم، یک الگوریتم برنامه‌نویسی پویای رو به عقب و یک الگوریتم رو به عقب از نوع صاف کردن به دست آمده که برای محاسبه‌ی دو نوع نمایه‌ی نقطه‌ی تغییر و قطعه مشتق کاربرد دارد. روش‌های پیشنهادی

<sup>۱</sup> Weakly-supervised

<sup>۲</sup> Support measure machine (SMM)

<sup>۳</sup> Random fourier features (RFF)

<sup>۴</sup> Wild binary segmentation (WBS)

<sup>۵</sup> Binary segmentation (BS)

برای جستجو در فضای قطعه‌بندی و تعداد متوالی نقاط تغییر مفید هستند و مجموعه‌ای از ابزارهای ارزیابی برای مدل‌های نقاط تغییر را ارائه می‌دهند که در چارچوب بیزی<sup>۱</sup> و غیر بیزی اعمال می‌شوند.

در مقاله‌ی [۲۳] یک توسعه از الگوریتم برنامه‌نویسی پویای کلاسیک را برای تشخیص جهش در سیگنال‌های قطعه‌ای همراه با نویز پیشنهاد می‌کند. داده‌ها حاصل آزمایش‌های رابط مغز و رایانه<sup>۲</sup> هستند و ساختار زمانی دارند و می‌توانند معیاری برای بررسی الگوریتم پیشنهادی بر روی داده‌های دنیای واقعی باشند. برای نشان دادن پتانسیل این روش، یک سیگنال دوبعدی شبیه‌سازی شده در نظر گرفته شده است. سیگنال شامل ۲۰۰ مشاهده است که ۸ بخش آن دارای مدت‌زمان متغیر است و به طور متناوب از یکی از نیمه حلقه‌های نمایش داده‌شده نمونه‌برداری می‌شود.

در مقاله‌ی [۲۴] مشکل تشخیص نقاط تغییر در مجموعه داده‌های حجیم بررسی شده است که به صورت سری زمانی هستند. تشخیص نقاط تغییر با به حداقل رساندن تابع هزینه بر روی تعداد و مکان‌های نقاط تغییر در نظر گرفته شده است که شامل روش‌های تعیین شده مانند احتمال جریمه و حداقل طول توصیف برای تشخیص نقاط تغییر است. همچنین روش جدیدی برای یافتن حداقل توابع هزینه و تعداد و مکان بهینه‌ی نقاط تغییر معرفی شده است که دارای هزینه‌ی محاسباتی خطی از نظر تعداد مشاهدات است. این روش با روش‌هایی مقایسه شده است که هزینه‌ی محاسباتی درجه‌ی دوم یا درجه‌ی سوم دارند. الگوریتم پیشنهادی با الگوریتم قطعه‌بندی دودویی برای تشخیص نقاط تغییر مقایسه شده است و نشان می‌دهد که دقت رویکرد پیشنهادی منجر به بهبود قابل توجهی در دقت قطعه‌بندی داده‌ها شده است.

در مقاله‌ی [۲۵] یک مطالعه‌ی گسترده برای توصیف اندازه‌ی پنجره، تعیین تأثیر آن در فرایند تشخیص فعالیت و کمک به روشن شدن برخی از مفروضات در طول طراحی سامانه ارائه شده است. برای این منظور، برخی از پرکاربردترین روش‌های تشخیص فعالیت برای طیف وسیعی از اندازه‌ها و فعالیت‌های پنجره ارزیابی شده‌اند. بهترین مبادله بین سرعت تشخیص و دقت در بازه‌ی ۱-۲ ثانیه در نظر گرفته شده است. دستوالعمل‌هایی برای پیکربندی سامانه با توجه به برنامه‌ها و فعالیت‌های هدف به طراحان ارائه شده است.

<sup>۱</sup> Bayesian

<sup>۲</sup> Brain-computer interface (BCI)



مجموعه داده‌ی استفاده شده شامل داده‌های حرکتی ۱۷ داوطلب است که ۳۳ فعالیت تناسب‌اندام را انجام می‌دهند، درحالی که مجموعه‌ای از ۹ حسگر اینرسی متصل به قسمت‌های مختلف بدن داده‌ها را ثبت می‌کنند.

در مقاله‌ی [۲۶] یک روش مبتنی بر هسته برای تجزیه و تحلیل نقطه‌ی تغییر در یک سری زمانی معرفی شده است. تجزیه و تحلیل نقطه‌ی تغییر در یک نمونه‌ی بدون برچسب از مشاهدات به بررسی تغییر در توزیع نمونه و تخمین لحظه‌ی آن می‌پردازد. یک آمار آزمایشی بر اساس توزیع فیشر به عنوان معیاری برای همگنی بین قطعه‌ها پیشنهاد می‌شود. از یک مجموعه داده‌ی پیشنهادی در مسابقه BCI استفاده شده است که در طی ۴ جلسه‌ی بدون بازخورد روی ۳ فرد جمع‌آوری می‌شود. از فرد خواسته شده است که فعالیت‌های متفاوتی را انجام دهد. زمانی که فرد از یک فعالیت به فعالیت دیگر تغییر وضعیت می‌دهد، تصادفی است.

در مقاله‌ی [۲۷] اولین بررسی گسترده و مقایسه‌ی تجربی تکنیک‌های پیشنهادی در زمینه‌ی استخراج پایگاه‌های داده‌ی سری زمانی انجام شده است. همه‌ی این الگوریتم‌ها از دیدگاه داده کاوی دارای کمبودهای جبران ناپذیری هستند. یکی از متداول‌ترین نمایش‌های مورد استفاده، تقریب خطی دسته‌ای است. این نمایش توسط محققان مختلف برای استفاده در خوشه‌بندی، طبقه‌بندی، نمایه‌سازی و استخراج قوانین مرتبط با داده‌های سری زمانی استفاده شده است.

در مقاله‌ی [۲۸] روشی برای تحلیل نقطه‌ی تغییر ارائه شده است. تخمین ناپارامتری تعداد نقاط تغییر و موقعیت‌هایی که در آن رخ می‌دهند، بر روی مجموعه‌ای از مشاهدات چند متغیره با ابعاد دلخواه انجام شده است. تخمین بر اساس خوشه‌بندی سلسله‌مراتبی است و دو الگوریتم قطعه‌بندی و تجمعی پیشنهاد شده است. در روش قطعه‌بندی، تخمینی از تعداد و مکان نقاط تغییر تحت مفروضات استاندارد ارائه شده است. رویکرد پیشنهادی با روش‌های دیگر در یک شبیه‌سازی مقایسه شده است. روش‌های تحلیل خوشه‌ای برای ارزیابی عملکرد و امکان مقایسه‌ی ساده‌ی تخمین‌های مکان و زمان استفاده شده‌اند. از داده‌های ژنوم بلیکلی و ورت استفاده شده است. نمونه‌های ژنوم ۵۷ فرد مبتلا به تومور مثانه برای شناسایی تغییرات در تعداد کپی

اسیددی‌اکسی‌ریبونوکلیک<sup>۱</sup> آنها از هیبریداسیون مقایسه‌ای ژنومی آرایه‌ای اسکن شده‌اند. شدت هیبریداسیون نسبی باتوجه‌به سیگنال مرجع ژنوم طبیعی ثبت شده است. این مشاهدات در صورتی نرمال هستند که نسبت مودال در مقیاس لگاریتمی صفر باشد.

در مقاله‌ی [۲۹] یک الگوریتم برنامه‌ریزی پویای هرس شده برای یک راه حل بهینه پیشنهاد می‌شود. در تشخیص برون‌خط، یک الگوریتم برنامه‌نویسی پویا<sup>۲</sup> نقاط تغییر را تشخیص می‌دهد تا تابع هزینه‌ی درجه دوم را به حداقل برساند و پیچیدگی را از  $\Theta(Kn^2)$  به  $\Theta(Kn)$  کاهش دهد که  $n$  در آن تعداد مشاهدات است. در بدترین حالت پیچیدگی در زمان  $O(Kn^2)$  و فضای  $O(Kn)$  رخ می‌دهد که معادل برنامه‌ریزی پویای کلاسیک است. توالی‌ها با استفاده از یک سیگنال ثابت، سینوسی یا مستطیلی شبیه‌سازی می‌شود. برای امواج سینوسی و مستطیلی، دامنه‌ها و فرکانس‌های مختلفی در نظر گرفته می‌شود. در این مجموعه‌داده نویز گاوسی واریانس ۱، نویز یکنواخت واریانس ۱، نویز مجذور کای واریانس ۱ و نویز کای واریانس ۱ در نظر گرفته می‌شود.

در مقاله‌ی [۳۰] یک الگوریتم برای هرس مجموعه‌ی نقاط تغییر ارائه شده است. این روش بر اساس نمایش عملکردی هزینه‌ی قطعه‌بندی‌ها است. اگر یک پارامتر تک‌بعدی در هر قطعه وجود داشته باشد، پیچیدگی الگوریتم در بدترین حالت قرار دارد. در رویکرد پیشنهادی، بدترین حالت معادل پیچیدگی الگوریتم همسایگی است. برای یک تابع هزینه‌ی خاص، نشان داده می‌شود که حتی اگر هیچ نقطه‌ی تغییری در سیگنال وجود نداشته باشد، هرس کارآمد است.

در این پروژه از مجموعه‌داده‌ی Gene Expression Omnibus استفاده می‌شود. آرایه‌ی چندشکلی تک-نوکلئوتیدی<sup>۳</sup> در این مجموعه‌داده از آزمایش‌ها ساخته شده است. آرایه‌های چندشکلی تک‌نوکلئوتیدی امکان مطالعه‌ی افزایش و کاهش تعداد کپی اسیددی‌اکسی‌ریبونوکلیک در طول ژنوم را فراهم می‌کند. برای این توالی، اغلب از یک مدل نقاط تغییر با پیچیدگی درجه‌ی دو استفاده می‌شود.

<sup>1</sup> Deoxyribonucleic acid (DNA)

<sup>2</sup> Dynamic programming (DP)

<sup>3</sup> Single nucleotide polymorphisms (SNP)

هدف مقاله‌ی [۳۱] مطالعه‌ی ایجاد یک مدل خرگوش مناسب و انتخاب پروتکل بهینه برای کرایوآبلاسیون لاپاراتومی در سرطان پانکراس بود. بافت تومور VX2 به پانکراس خرگوش تلقیح شد تا مدل سرطان پانکراس ساخته شود. سپس خرگوش‌های حامل تومور به طور تصادفی به ۴ گروه تقسیم شدند. مدل کارسینوم پانکراس VX2 با موفقیت ایجاد شد و سطوح انولاز اختصاصی نورون سرم به طور مداوم پس از تلقیح افزایش یافت.

## ۷-۲ مقایسه‌ی کارهای پیشین

در جدول ۲-۲ خلاصه‌ای از کارهای پیشین و مزایا و معایب هر یک آورده شده است.

جدول ۲-۲: پیشینه

مرجع	روش پیشنهادی	مزایا	معایب
[۲۰]	شبکه‌ی عصبی عمیق	دقت مناسب و سرعت بالا	مقداردهی پارامترهای اولیه
[۲۱]	قطعه‌بندی دودویی وحشی	پیچیدگی محاسباتی کم	زمان اجرای زیاد
[۲۲]	تعبیه‌ی هسته‌ی توزیع‌ها برای قطعه-بندی	کارایی و مقیاس پذیری بالا	پیچیدگی اجرای زیاد
[۲۳]	الگوریتم برنامه‌نویسی پویا	کاوش فضای قطعه‌بندی برای تعداد متوالی نقاط تغییر	پیچیدگی محاسباتی بالا
[۲۴]	توسعه‌ای از الگوریتم برنامه‌نویسی پویای کلاسیک	دقت مناسب و سرعت بالا	مقداردهی پارامترهای اولیه
[۲۵]	به حداقل رساندن یک تابع هزینه بر روی تعداد و مکان‌های ممکن نقاط تغییر	افزایش دقت قطعه‌بندی	هزینه‌ی محاسباتی بالا
[۲۶]	استفاده از طیف وسیعی از اندازه‌های	افزایش سرعت و دقت	پیچیدگی محاسباتی بالا

	پنجره	تشخیص	
[۲۷]	روش مبتنی بر هسته برای تجزیه و تحلیل نقطه‌ی تغییر در یک دنباله	دقت مناسب	زمان اجرای زیاد
[۲۸]	تقریب خطی دسته‌ای	افزایش دقت قطعه‌بندی	مقداردهی پارامترهای اولیه
[۲۹]	تخمین ناپارامتری تعداد نقاط تغییر	پیچیدگی محاسباتی کم	مقداردهی پارامترهای اولیه
[۳۰]	استفاده از نمایش عملکردی هزینه‌ی قطعه‌بندی‌ها	پیچیدگی محاسباتی کم	زمان اجرای زیاد
[۳۱]	استفاده از پروتکل بهینه برای کرایو ابلاسیون	افزایش دقت	پیچیدگی اجرای زیاد

## ۸-۲ نتیجه‌گیری

در این فصل به کارهای پیشین انجام شده در زمینه‌ی قطعه‌بندی مبتنی بر دید و مبتنی بر حسگر پرداختیم. در فصل بعدی به ارائه‌ی راهکار پیشنهادی برای مقایسه‌ی استحکام الگوریتم‌ها می‌پردازیم.

## فصل سوم

### راهکار پیشنهادی

## ۳-۱ مقدمه

طبقه‌بندی مبتنی بر قطعه‌ی جریان داده<sup>۱</sup> در سال‌های اخیر به یک موضوع بسیار مهم در علم یادگیری ماشین تبدیل شده است، زیرا داده‌های بیشتری را می‌توان به‌عنوان داده‌های جریانی در نظر گرفت. بیشتر این داده‌ها با سرعت بالا، توزیع داده‌های غیرایستا<sup>۲</sup> و طول نامحدود مشخص می‌شوند که نمونه‌هایی از این برنامه‌ها را می‌توان در تراکنش‌های کارت اعتباری و مخابرات مشاهده کرد.

توزیع جریان داده‌ها در طول زمان تغییر می‌کند، به این پدیده رانش مفهوم گفته می‌شود. رانش مفهوم، مستلزم تنظیم طبقه‌بندها برای انطباق با شرایط جدید است. نوع تغییرات را می‌توان به رانش مفهوم ناگهانی<sup>۳</sup>، تدریجی<sup>۴</sup>، افزایشی<sup>۵</sup> و مکرر<sup>۶</sup> طبقه‌بندی کرد. برنامه‌های کاربردی دنیای واقعی همیشه ترکیبی از انواع مختلفی از رانش‌های مفهوم هستند. در این فصل به بیان راهکار پیشنهادی خواهیم پرداخت و جزئیات مسئله را مورد بررسی قرار خواهیم داد.

## ۳-۲ تعریف مسئله

در این پایان‌نامه از چهار روش جدید در طبقه‌بندی مبتنی بر قطعه‌ی جریان داده‌ها برای تشخیص فعالیت استفاده می‌شود. این چهار روش DUE، KUE، DWMIL و ACDWM هستند. هدف اصلی این پایان‌نامه تمرکز بر استحکام الگوریتم‌ها در برابر نویز و در نظر گرفتن عدم توازن داده و رانش مفهوم است.

## ۳-۲-۱ نویز چیست؟

در یادگیری ماشین، نویز یک نوع پیچیدگی در داده‌ها است. مجموعه داده‌های ناشی از اندازه‌گیری‌های واقعی اغلب حاوی داده‌های نویز هستند. دلایل ایجاد نویز می‌تواند متفاوت باشد، اما معمولاً به دلیل نقص،

<sup>1</sup> Data stream classification

<sup>2</sup> Non-static

<sup>3</sup> Abrupt drift (sudden drift)

<sup>4</sup> Gradual drift

<sup>5</sup> Incremental drift

<sup>6</sup> Recurring drift

کالیبراسیون معیوب تجهیزات اندازه‌گیری یا خطای انسانی است. نویز در داده‌ها ممکن است عملکرد یک مدل یادگیری ماشین را از نظر دقت در طبقه‌بندی کاهش دهد.

نویزها به دو دسته‌ی نویز ویژگی<sup>۱</sup> و نویز کلاس<sup>۲</sup> تقسیم می‌شوند. نویز ویژگی خطایی است که به مقادیر نسبت داده‌شده وارد می‌شود که به آن ویژگی‌های گم شده<sup>۳</sup> یا «نمی‌دانم»<sup>۴</sup> یا مقادیر اشتباه<sup>۵</sup> می‌گویند. نویز کلاس می‌تواند برچسب‌گذاری متناقض نمونه‌ها باشد. نمونه‌ها بیش از یک‌بار، اما با یک طبقه‌بندی متمایز ظاهر می‌شوند. ممکن است طبقه‌بندی اشتباه انجام شود (خروجی نادرست) و نمونه‌ها با کلاس‌های اشتباه برچسب‌گذاری شوند [۳۶].

### ۳-۲-۲ استحکام در برابر نویز

طبقه‌بندهای یادگیری ماشین نظارت‌شده، اغلب با دقت آنها برای وظایف خاص محک زده می‌شوند؛ اما یکی دیگر از ویژگی‌های مهم و گاه نادیده گرفته‌شده‌ی طبقه‌بند، استحکام آن در برابر نویز است. از آنجایی که داده‌های دنیای واقعی اغلب حاوی نویز هستند، در چنین مواردی یک طبقه‌بند باید بتواند تا حدی در برابر نویز مقاومت کند. اندازه‌گیری توانایی طبقه‌بند برای مقاومت در برابر نویز را می‌توان با آموزش آن بر روی یک مجموعه داده‌ی بدون نویز و سپس روی یک کپی نویز از همان مجموعه داده انجام داد. نتایج را می‌توان با معیارهایی مانند معیار دقت یا میانگین ناحیه‌ی زیر منحنی در همه‌ی قطعه‌ها اندازه‌گیری کرد.

### ۳-۲-۳ تولید نویز

هنگام ایجاد مجموعه داده‌های نویز، مواردی مانند مکان و نحوه‌ی معرفی و نوع نویز باید در نظر گرفته شود. در مورد نوع نویز، باید ویژگی‌های طبقه‌بندی و ویژگی‌های پیوسته را مدنظر داشت. برای ویژگی‌های موجود، از نویز گاوسی استفاده می‌شود و مقادیر نویز بر اساس توزیع گاوسی اضافه می‌شود. نویز با توجه به تابع چگالی احتمال زیر به داده‌ها اعمال می‌شود:

<sup>1</sup> Attribute noise

<sup>2</sup> Class noise

<sup>3</sup> Missing

<sup>4</sup> Don't know

<sup>5</sup> Erroneous

$$f(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp \left[ -\frac{(x-\mu)^2}{2\sigma^2} \right] \quad \text{رابطه ی ۱-۳}$$

X متغیر تصادفی گاوسی،  $\mu$  مقدار میانگین<sup>۱</sup> و ویژگی محاسبه شده در مجموعه داده و  $\sigma$  حراف استاندارد<sup>۲</sup> است [۳۶].

### ۳-۲-۴ الگوریتم های استفاده شده

۱- الگوریتم KUE: در این الگوریتم، یک روش گروهی جدید به نام Kappa Updated Ensemble (KUE) پیاده سازی شده است. یادگیری از جریان داده ها در حضور رانش مفهوم، یکی از بزرگ ترین چالش های یادگیری ماشین است. الگوریتم های طراحی شده برای چنین سناریوهایی باید اندازه ی بالقوه نامحدود داده ها به همراه ماهیت دائماً در حال تغییر آن ها و نیاز به پردازش بلادرنگ را در نظر بگیرند. رویکردهای گروهی طبقه بندها به دلیل قابلیت پیش بینی بالا و مکانیسم مؤثر در کاهش رانش مفهوم، در مطالعه ی جریان داده محبوبیت قابل توجهی یافته است. الگوریتم KUE ترکیبی از رویکردهای گروهی<sup>۳</sup> بر خط و مبتنی بر قطعه است که از آمار کاپا برای وزن دهی پویا و انتخاب طبقه بندهای پایه استفاده می کند. طبقه بندهای جدید با استفاده از نمونه های جدید و توزیع پواسون برای دستیابی به تنوع بیشتر به روز شده است.

هر طبقه بند پایه در KUE می تواند در رأی گیری شرکت کند و استحکام KUE را افزایش دهد. یک مطالعه ی تجربی گسترده نشان می دهد که KUE می تواند با وجودی که پیچیدگی محاسباتی پایینی دارد، از دیگر الگوریتم های پیشرفته بر روی جریان داده های رانش مفهوم استاندارد و نامتوازن بهتر عمل کند.

۲- الگوریتم DUE: این الگوریتم برای یادگیری جریان داده های نامتوازن با رانش مفهوم استفاده شده است. نکته مهم در این الگوریتم این است که برای داده های دارای عدم توازن نیز مناسب است. تنها چند الگوریتم خاص به موضوع مشترک رانش مفهوم و عدم توازن به دلیل پیچیدگی آن می پردازند. در همین حال، الگوریتم های قطعه ای موجود برای طبقه بندی جریان داده های غیرایستای نامتوازن، همیشه نیاز به ذخیره ی

<sup>1</sup> Mean

<sup>2</sup> Standard deviation

<sup>3</sup> Ensemble



داده‌های قبلی دارند که حافظه‌ی زیادی را مصرف می‌کند. برای غلبه بر این مسائل، الگوریتم DUE معرفی شده است. این الگوریتم در مقایسه با تکنیک‌های موجود، پنج مزیت دارد:

- (۱) بدون نیاز به دسترسی به داده‌های قبلی، یک قطعه از داده را در یک‌زمان یاد می‌گیرد.
  - (۲) در روش به‌روزرسانی مدل بر نمونه‌هایی که نادرست طبقه‌بندی شده‌اند، تأکید می‌کند.
  - (۳) می‌تواند به‌موقع به انواع مختلفی از رانش‌های مفهوم، واکنش نشان دهد.
  - (۴) می‌تواند با تغییر کلاس اکثریت<sup>۱</sup> به کلاس اقلیت<sup>۲</sup> با شرایط جدید سازگار شود.
  - (۵) تعداد محدودی از طبقه‌بندها را برای اطمینان از کارایی بالا نگه می‌دارد.
- آزمایش‌ها روی مجموعه داده‌های مصنوعی و واقعی، اثربخشی DUE را در یادگیری جریان داده‌های نامتوازن غیرایستا نشان می‌دهند.

۳- الگوریتم DWMIL: در این الگوریتم یک روش یادگیری افزایشی مبتنی بر قطعه به نام اکثریت وزن دار پویا برای یادگیری عدم توازن بر روی جریان داده‌های دارای رانش مفهوم و مشکل عدم توازن کلاس پیشنهاد می‌شود. رانش مفهوم که در جریان داده‌ها رخ می‌دهد، دقت و ثبات فرایند یادگیری برخط را به خطر می‌اندازد. اگر جریان داده نامتوازن باشد، تشخیص و درمان رانش مفهوم، چالش‌برانگیز خواهد بود. از یک چارچوب گروهی با وزن‌دهی پویای طبقه‌بندهای پایه با توجه به عملکرد آنها در قطعه‌ی فعلی استفاده می‌شود. در مقایسه با روش‌های موجود، این الگوریتم چهار قابلیت دارد:

- (۱) می‌تواند برای جریان داده‌های بدون رانش مفهوم، پایدار بماند و به‌سرعت با رانش مفهوم جدید سازگار شود.

(۲) کاملاً افزایشی است، یعنی هیچ داده‌ی قبلی ذخیره نمی‌شود.

(۳) تعداد محدودی از طبقه‌بندها را برای اطمینان از کارایی بالا نگه می‌دارد.

<sup>1</sup> Majority class

<sup>2</sup> Minority class

۴) ساده است و تنها به یک پارامتر آستانه نیاز دارد.

آزمایش‌ها بر روی مجموعه داده‌های مصنوعی و واقعی با رانش مفهوم نشان می‌دهد که این الگوریتم بهتر از الگوریتم‌های پیشرفته و با هزینه‌ی محاسباتی کمتر عمل می‌کند.

۴- الگوریتم ACDWM: در این الگوریتم، یک روش یادگیری افزایشی پیشنهاد شده است که مبتنی بر قطعه‌ای به نام اکثریت وزنی پویا برای مقابله با جریان داده‌های نامتوازن دارای رانش مفهوم است. اگر جریان داده نامتوازن باشد، تشخیص رانش مفهوم سخت است. الگوریتم‌های گروهی برای طبقه‌بندی داده‌های جریانی با رانش مفهوم مؤثر هستند، یک طبقه‌بند برای هر قطعه از داده‌ی ورودی ساخته می‌شود و وزن مربوط به آن برای مدیریت رانش مفهوم تنظیم می‌شود. با این حال، تنظیم وزن‌ها برای دستیابی به تعادل بین ثبات و سازگاری طبقه‌بندها در الگوریتم گروهی دشوار است.

علاوه بر این، هنگامی که جریان داده نامتوازن است، استفاده از یک قطعه با اندازه ثابت برای ساخت یک طبقه‌بند می‌تواند مشکلات بیشتری ایجاد کند. قطعه‌ی داده ممکن است شامل نمونه‌های کلاس اقلیت با تعداد بسیار کم باشد یا حتی هیچ نمونه‌ای از کلاس اقلیت نداشته باشد و فقط نمونه‌های کلاس اکثریت را داشته باشد. طبقه‌بندی که بر روی چنین قطعه‌ای ساخته شده است، ناپایدار است. از یک چارچوب گروهی با وزن‌دهی پویای طبقه‌بندها بر اساس عملکرد طبقه‌بندی آنها بر روی قطعه‌ی داده‌ی فعلی استفاده می‌شود. اندازه‌ی قطعه به طور پویا توسط آزمون‌های فرضیه‌ی آماری برای رسیدن به پایداری طبقه‌بند روی قطعه‌ی فعلی انتخاب می‌شود. این الگوریتم در مقایسه با روش‌های موجود دارای سه مزیت به شرح زیر است:

۱) می‌تواند در هنگام پردازش جریان داده‌های بدون رانش مفهوم، پایداری خود را حفظ کند و به سرعت با رانش مفهوم جدید سازگار شود.

۲) کاملاً افزایشی است، یعنی نیازی به ذخیره‌سازی داده‌های قبلی ندارد و تعداد محدودی از طبقه‌بندها را برای اطمینان از کارایی بالا ذخیره می‌کند.

۳) به صورت پویا، اندازه‌ی قطعه را در محیط رانش مفهوم انتخاب می‌کند.

آزمایش‌ها بر روی مجموعه داده‌های مصنوعی و واقعی که حاوی رانش مفهوم هستند نشان می‌دهند که ACDWM نسبت به هر دو روش مبتنی بر قطعه و برخط برتری دارد.

### ۳-۲-۵ مجموعه داده‌های استفاده شده در آزمایش‌ها

در آزمایش‌های انجام شده برای تعریف رابطه‌ها، از ۴ مجموعه داده‌ی مربوط به تشخیص فعالیت استفاده شده است. مجموعه داده‌های مصنوعی استفاده شده با مولد رانش مفهوم تولید می‌شوند. این مجموعه داده‌ها شامل انواع مختلفی از رانش واقعی<sup>۱</sup> و رانش مجازی<sup>۲</sup> هستند. رانش مفهوم قبلی با تنظیم دستی نرخ عدم توازن<sup>۳</sup> در همه‌ی مجموعه داده‌ها با نمونه‌برداری کم، وارد جریان داده می‌شود تا بتوان به تعریف رانش مفهوم مشترک نزدیک شد. نرخ عدم توازن در آزمایش‌ها به شرح زیر تغییر می‌کند:

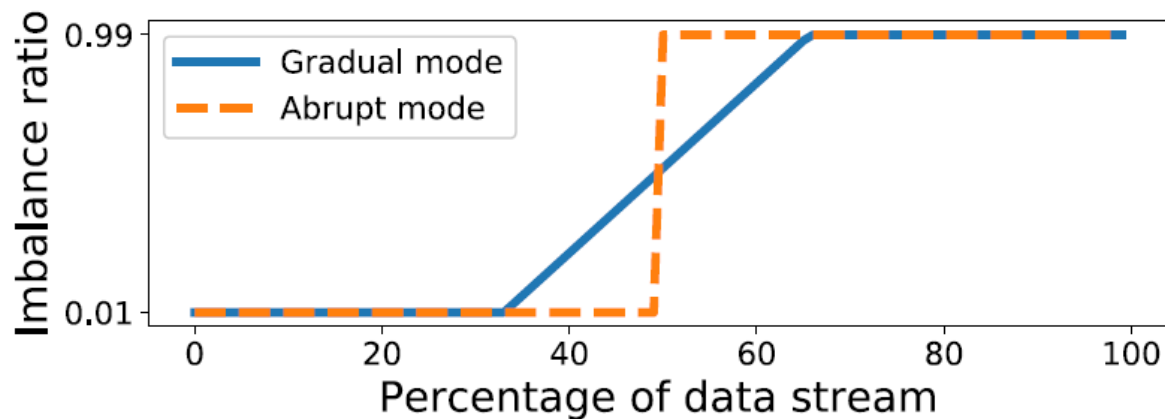
(۱) رانش ناگهانی: نرخ عدم توازن در ابتدا ۰/۰۱ تنظیم شده است. پس از نیمی از جریان داده، نرخ عدم توازن ناگهان به ۰/۹۹ تغییر می‌کند، یعنی کلاس اکثریت با نرخ عدم توازن ۰/۰۱ به کلاس اقلیت تبدیل می‌شود. اقدامات مقدماتی در موقعیت رانش ناگهانی تنظیم مجدد می‌شود.

(۲) رانش تدریجی: نرخ عدم توازن در ابتدا ۰/۰۱ تعیین می‌شود. پس از یک سوم جریان داده، نرخ عدم توازن به تدریج شروع به افزایش می‌کند تا زمانی که در دو سوم جریان داده به ۰/۹۹ برسد. اقدامات مقدماتی در موقعیت‌های شروع و پایان رانش تدریجی تنظیم مجدد می‌شود.

<sup>۱</sup> Real drift

<sup>۲</sup> Virtual drift

<sup>۳</sup> Imbalance ratio



شکل ۳-۱: مقدار نرخ عدم توازن در مجموعه داده [۳۲].

نرخ عدم توازن در اینجا به درصد نمونه‌های کلاس مثبت اشاره دارد. برای کنترل نرخ عدم توازن، نمونه‌برداری کم روی ۱۰۰۰ نمونه در جریان داده‌ی اصلی انجام می‌شود. اگر نرخ عدم توازن اولیه در این قطعه<sup>۱</sup> کوچک‌تر از نرخ عدم توازن اختصاص داده‌شده باشد، کلاس اکثریت کمتر نمونه‌برداری می‌شود و اگر نرخ عدم توازن تخصیص داده‌شده از نرخ عدم توازن اولیه در این قطعه کوچک‌تر باشد، کلاس اقلیت کمتر نمونه‌برداری می‌شود. از آنجایی که نرخ عدم توازن در هر مجموعه داده متفاوت است، موقعیت رانش مفهوم پس از نمونه‌برداری کم نیز متفاوت است.

مجموعه داده‌های استفاده شده به شرح زیر هستند:

جدول ۳-۱: مجموعه داده‌های استفاده شده

نام مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	تعداد کلاس
HCI	۷۳۳۴۸	۵۷	۶
Opportunity	۶۶۲۴۸	۲۴۳	۷
Skoda	۶۹۶۹۷	۶۰	۳۱

<sup>۱</sup> Chunk

۴	۱۸	۷۵۰۹۹	WISDM
---	----	-------	-------

۱- مجموعه داده‌ی Skoda mini checkpoint: این مجموعه داده شامل ۱۰ حرکت انجام شده در سناریوی تعمیر و نگهداری خودرو است. ۲۰ حسگر شتاب سه بعدی (۶۰ ویژگی) دارد و نمونه‌های قطعه‌بندی شده و پیوسته در مجموعه داده وجود دارد.

۲- مجموعه داده‌ی HCI gestures: این مجموعه داده با انجام ۵ حرکت با دست به صورت آزاد یا هدایت شده در مقابل تخته سیاه به دست می‌آید. ۸ حسگر شتاب سه بعدی (۲۴ ویژگی) دارد و نمونه‌های قطعه‌بندی شده و پیوسته در مجموعه داده وجود دارد.

۳- مجموعه داده‌ی HCI Tabletop Gestures: برای جمع‌آوری این مجموعه داده، ۹ حرکت نوشتن با استفاده از القای کف دست<sup>۱</sup> در ۳ اندازه و روی چندین سطح لمسی را بررسی کرده که استفاده از موس در حالت نشسته و ایستاده، استفاده از تبلت ایستاده، استفاده از میز لمسی نشسته و ایستاده را شامل می‌شود.

۴- مجموعه داده‌ی WISDM: در جمع‌آوری این مجموعه داده، داده‌های خام شتاب‌سنج و حسگر ژيروسکوپ از تلفن هوشمند و ساعت هوشمند با نرخ ۲۰ هرتز جمع‌آوری می‌شود. این مجموعه داده از ۵۱ آزمودنی جمع‌آوری شده است. آن‌ها ۱۸ فعالیت را به مدت ۳ دقیقه انجام می‌دهند. داده‌های حسگر برای هر دستگاه (تلفن یا ساعت) و هر نوع حسگر (شتاب‌سنج یا ژيروسکوپ) در یک فهرست متفاوت ذخیره می‌شود؛ بنابراین ۴ فهرست داده وجود دارد. در هر فهرست راهنما ۵۱ فایل مربوط به ۵۱ موضوع آزمون وجود دارد. فرمت هر ورودی یکسان است. علاوه بر داده‌های حسگر سری زمانی خام، نمونه‌هایی نیز تولید می‌شود که داده‌های حسگر را با استفاده از یک پنجره به اندازه‌ی ۱۰ ثانیه توصیف می‌کند [۳۲].

۵- مجموعه داده‌ی Opportunity: از این مجموعه داده برای تشخیص فعالیت انسانی با حسگرهای پوشیدنی<sup>۲</sup>، شیء<sup>۳</sup> و محیطی<sup>۴</sup> استفاده می‌شود. این مجموعه داده به عنوان معیار الگوریتم‌های تشخیص فعالیت

<sup>۱</sup> Palm alphabet

<sup>۲</sup> Wearable sensors

<sup>۳</sup> Object sensors

<sup>۴</sup> Ambient sensors

انسانی، مانند طبقه‌بندی و قطعه‌بندی خودکار داده‌ها ایجاد شده است. ۶ اجرا برای هر کاربر از ۴ کاربر انجام شده است. از این تعداد، ۵ مورد فعالیت‌های روزانه‌ی زندگی هستند و اجرای ششم یک اجرای تمرینی<sup>۱</sup> است که در آن کاربران دنباله‌ای از فعالیت‌ها را اجرا می‌کنند. حسگرهای پوشیدنی شامل ۷ واحد اندازه‌گیری اینرسی<sup>۲</sup>، ۱۲ حسگر شتاب<sup>۳</sup> سه‌بعدی، ۴ حسگر محلی‌سازی<sup>۴</sup> سه‌بعدی است. حسگرهای شیء از ۱۲ جسم با شتاب سه‌بعدی و سرعت چرخش ۲ بعدی داده جمع‌آوری می‌کند. حسگرهای محیطی شامل ۱۳ حسگر سوئیچ<sup>۵</sup> و ۸ حسگر شتاب سه‌بعدی است.

### ۳-۲-۶ محاسبه‌ی استحکام

برای تعریف رابطه‌ی استحکام نیاز به تعریف دو پارامتر اندازه‌ی قطعه<sup>۶</sup> و نرخ عدم توازن داریم:

(۱) اندازه‌ی قطعه: باتوجه به این‌که روش‌های گروهی طبقه‌بندها نیاز به تقسیم به چندین قطعه را دارند، مقدار این متغیر می‌تواند در خروجی نهایی بسیار تأثیرگذار باشد، زیرا در انتها برای نتیجه‌گیری باید عملیات رأی‌گیری انجام شود تا خروجی نهایی انتخاب شود. برای مثال اگر مقدار اندازه‌ی قطعه، عدد ۱۰ باشد، تعداد قطعه‌بندهای انتهایی برابر با ۱۰ بوده و عملیات رأی‌گیری باید بین ۱۰ خروجی انجام گردد. هرکدام از این خروجی‌ها (۰ یا ۱) رأی بیشتری جمع کند، به‌عنوان خروجی نهایی برگزیده می‌شود. مقدار این متغیر باید به‌صورت تقریبی انتخاب شود. برای مثال اگر مقدار این متغیر بسیار پایین در نظر گرفته شود (۲-۵)، خروجی به‌دست‌آمده قابل استناد نیست و کم‌برازش<sup>۷</sup> می‌شود؛ چون تعداد طبقه‌بندهای به‌دست‌آمده کم است. در نتیجه رأی‌گیری بین تعداد کمی خروجی انجام می‌شود و نتیجه‌ی به‌دست‌آمده از اعلان نظر تعداد کمی از طبقه‌بندها خواهد بود. اگر مقدار این متغیر بسیار بالا در نظر گرفته شود، زمان اجرای برنامه بسیار زیاد خواهد شد و عملاً با سرعت اجرای کم، قابلیت استفاده در دنیای واقعی را ندارد. ضمن اینکه امکان بروز

<sup>1</sup> Drill

<sup>2</sup> Inertial measurement units

<sup>3</sup> Acceleration sensors

<sup>4</sup> Localization

<sup>5</sup> Switch

<sup>6</sup> Chunk Size

<sup>7</sup> Underfit

پدیده‌ی بیش برآزش<sup>۱</sup> نیز وجود دارد. ممکن است طبقه‌بندها به داده‌های آموزشی خیلی نزدیک شوند و قابلیت انعطاف خود را برای پیش‌بینی داده‌های جدید از دست بدهند.

(۲) نرخ عدم توازن: نرخ عدم توازن در روش‌های طبقه‌بند، میزان تأثیر داده‌های نامتوازن را مشخص می‌کند. ایجاد قطعه‌بند بر روی داده‌های نامتوازن بسیار سخت و زمان‌بر است. هر چه این مقدار بیشتر باشد تأثیر داده‌های نامتوازن بیشتر خواهد بود و قطعه‌بند نامتوازن‌تر می‌گردد.

برای تعریف رابطه‌ی مورد نظر، ۴ الگوریتم بر روی ۴ مجموعه‌داده با اندازه‌ی قطعه‌های مختلف مورد ارزیابی قرار گرفت. نمودارهای این فصل حاصل میانگین نتایج برای تمام قطعه‌ها است. پارامتر مورد ارزیابی ناحیه‌ی زیر منحنی است.

ابتدا به بررسی ۴ مفهوم بسیار مهم در طبقه‌بندها خواهیم پرداخت:

(۱) مثبت واقعی<sup>۲</sup>: نمونه‌هایی که به کلاس ۱ (کلاس داده‌های نرمال) تعلق داشتند و درست طبقه‌بندی شدند.

(۲) مثبت کاذب<sup>۳</sup>: نمونه‌هایی که به کلاس ۱ (کلاس داده‌های نرمال) تعلق داشتند و نادرست طبقه‌بندی شدند.

(۳) منفی واقعی<sup>۴</sup>: نمونه‌هایی که به کلاس ۰ (کلاس داده‌های غیرنرمال) تعلق داشتند و درست طبقه‌بندی شدند.

(۴) منفی کاذب<sup>۵</sup>: نمونه‌هایی که به کلاس ۰ (کلاس داده‌های غیرنرمال) تعلق داشتند و نادرست طبقه‌بندی شدند.

حال رابطه‌های مربوط به معیارها را شرح می‌دهیم.

---

<sup>۱</sup> Overfit

<sup>۲</sup> True positive (TP)

<sup>۳</sup> False positive (FP)

<sup>۴</sup> True negative (TN)

<sup>۵</sup> False negative (FN)

**Accuracy** = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative) رابطه‌ی ۲-۳

**Precision** (Sensirivity) = True Positive / (True Positive + False Positive) رابطه‌ی ۳-۳

**Recall** (Specificity) = True Positive / (True Positive + False Negative) رابطه‌ی ۴-۳

معیار دقت برای مقایسه‌ی مجموعه‌داده‌های متوازن استفاده می‌شود. هر مدلی که دقت آن بیشتر باشد، مدل بهتری است، اما برای مقایسه‌ی دو معیار دیگر نیاز به یک رابطه‌ی جدید داریم. معیار دیگری به نام **f\_score** یا **f1\_score** وجود دارد که از ترکیب دو معیار بالا حاصل می‌شود.

**f\_score** = 2 \* (Precision \* Recall) / (Precision + Recall) رابطه‌ی ۵-۳

هر مدلی **f\_score** بیشتری داشته باشد، مدل بهتری است. حال می‌توانیم نرخ هریک از روابط بالا را به دست آوریم:

**True positive rate:** TPrate = TP / (TP + FN) رابطه‌ی ۶-۳

**True negative rate:** TNrate = TN / (FP + TN) رابطه‌ی ۷-۳

**False positive rate:** FPrate = FP / (FP + TN) رابطه‌ی ۸-۳

**False negative rate:** FNrate = FN / (TP + FN) رابطه‌ی ۹-۳

حال با استفاده از روابط بالا می‌توان معیار ناحیه زیر منحنی را تعریف کرد که برای مجموعه‌داده‌های نامتوازن مناسب است:

**AUC** = (1 + TPrate - FPrate) / 2 رابطه‌ی ۱۰-۳

به دلیل پخش بودن مقادیر میانگین ناحیه‌ی زیر منحنی برای همه قطعه‌ها و بالا و پایین رفتن آن در نمودارها، می‌توان از فرمول انحراف استاندارد برای مقایسه استحکام الگوریتم‌ها کمک گرفت با این تفاوت که به جای میانگین، از مقدار میانگین ناحیه زیر منحنی برای همه‌ی قطعه‌ها در حالت بدون نویز استفاده می‌شود. با توجه به نتایج به دست آمده در نمودارهای سناریوی اول، می‌توان به رابطه زیر برای محاسبه استحکام رسید:



$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} \quad \text{رابطه ی ۱۱-۳}$$

در رابطه ی ۱۱-۳ علامت  $\sigma$  انحراف استاندارد جمعیت،  $x_i$  مقادیر میانگین ناحیه ی زیر نمودار برای همه ی قطعه ها در حالت نويزدار،  $\mu$  میانگین ناحیه ی زیر نمودار برای همه ی قطعه ها در حالت بدون نويز و  $N$  تعداد آزمایشات است.

الگوریتمی که انحراف استاندارد کمتری دارد، به ازای درصد نويز متفاوت، استحکام بیشتری دارد و معیار مورد محاسبه (ناحیه ی زیر منحنی) تغییرات کمتری دارد.

۳-۲-۷ محاسبه ی نويز، نرخ عدم توازن و رانش مفهوم

برای یافتن رابطه ای به منظور مقایسه ی استحکام الگوریتم های طبقه بندی مبتنی بر قطعه ی جریان داده ها در برابر نويز، آزمایشاتی در شرایط مختلف انجام گرفت.

برای محاسبه ی رابطه ی استحکام، از نتایج حاصل از ۳ سناریوی مختلف استفاده شده است. برای آزمایش در این مرحله، ۳ معیار مختلف نويز، رانش مفهوم و نرخ عدم توازن مورد استفاده قرار گرفتند. این ۳ سناریو به شرح زیر است:

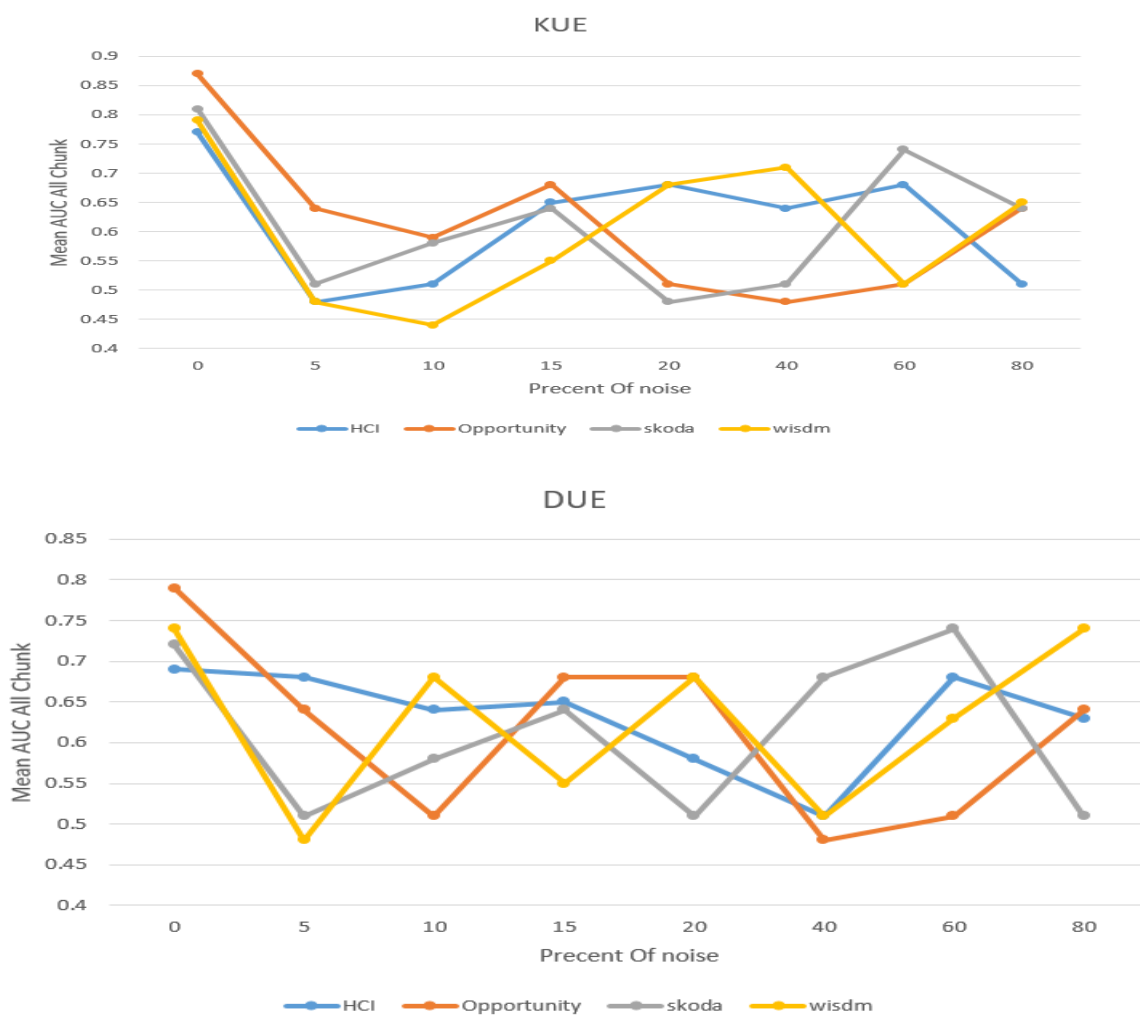
(۱) رانش مفهوم و نرخ عدم توازن برابر با صفر بوده و نويز، متغیر است.

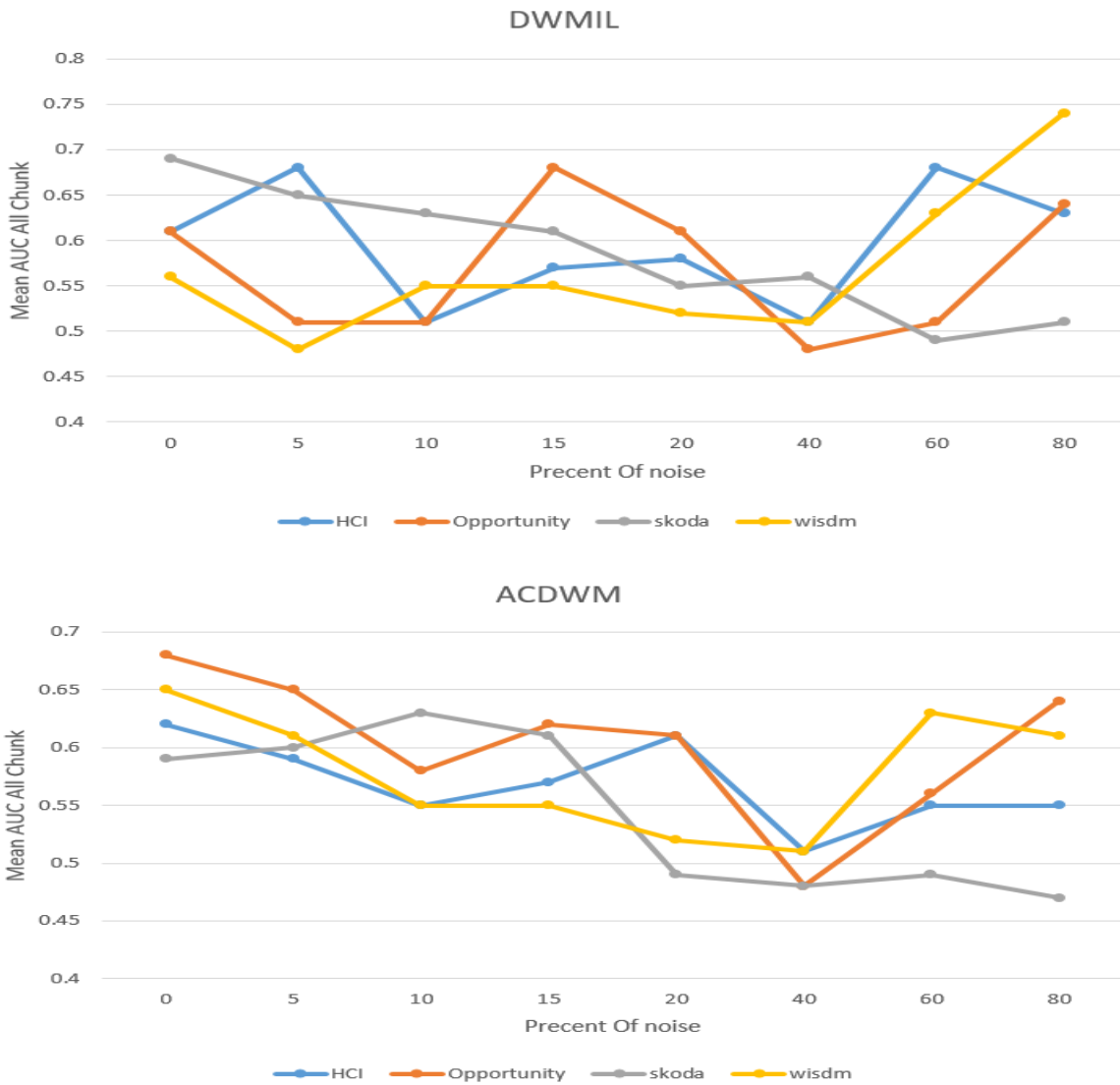
(۲) رانش مفهوم، برابر با صفر بوده و مقادیر نويز و نرخ عدم توازن، متغیر است.

(۳) نرخ عدم توازن برابر با صفر بوده و مقادیر رانش مفهوم و نويز، متغیر است.

برای هر سناریوی مذکور، ۴ الگوریتم بر روی ۴ مجموعه داده آزمایش شده است:

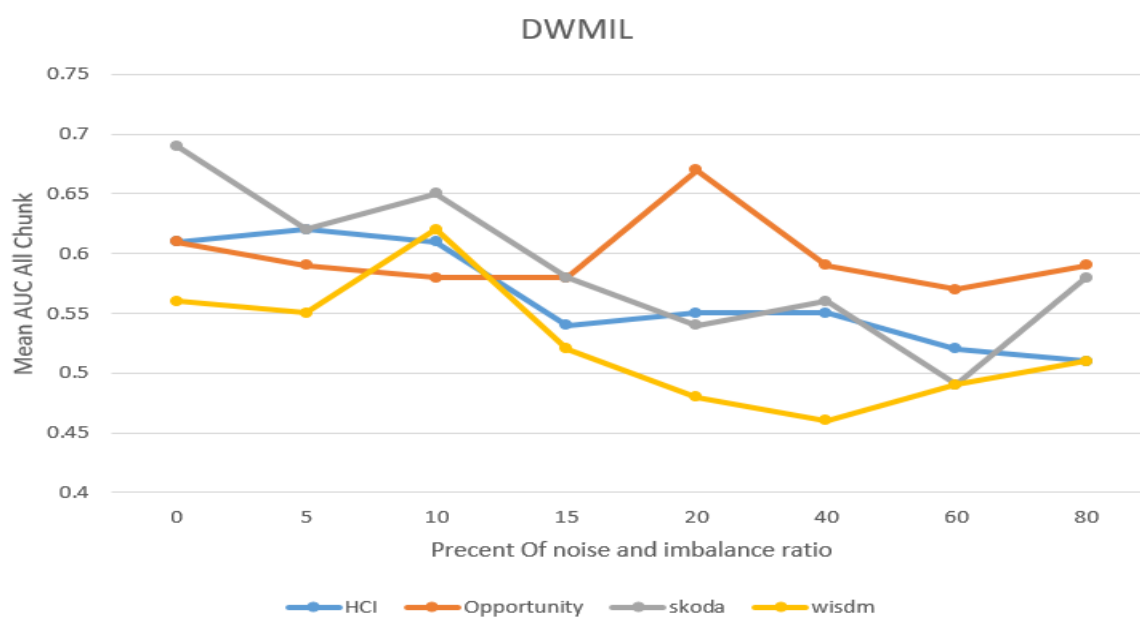
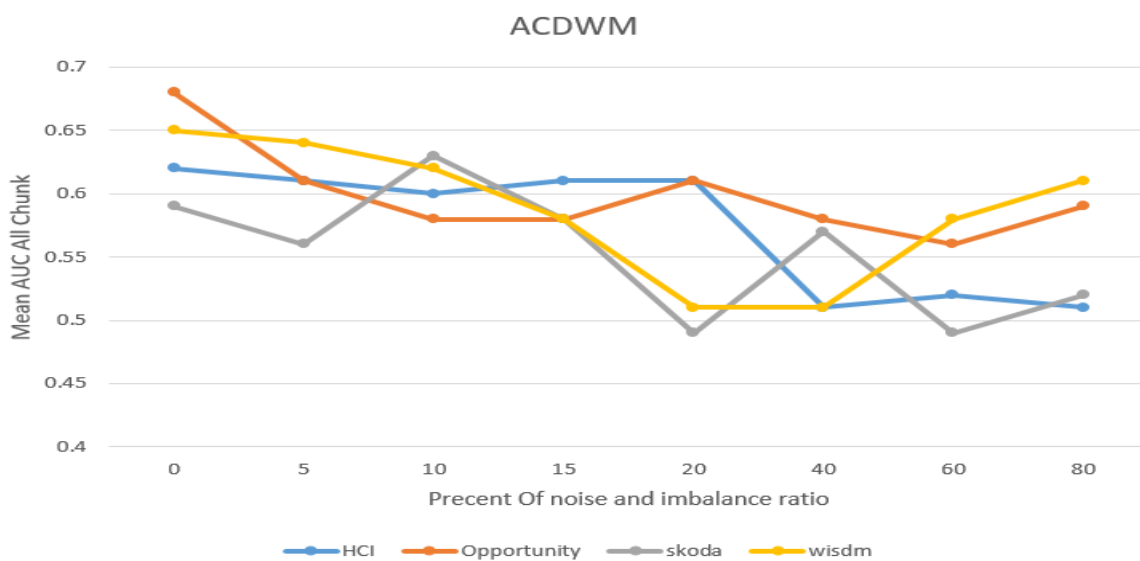
در ادامه نمودارهای مربوط به الگوریتم ها بر روی مجموعه داده های تشخیص فعالیت در سناریوی اول نمایش داده شده است.

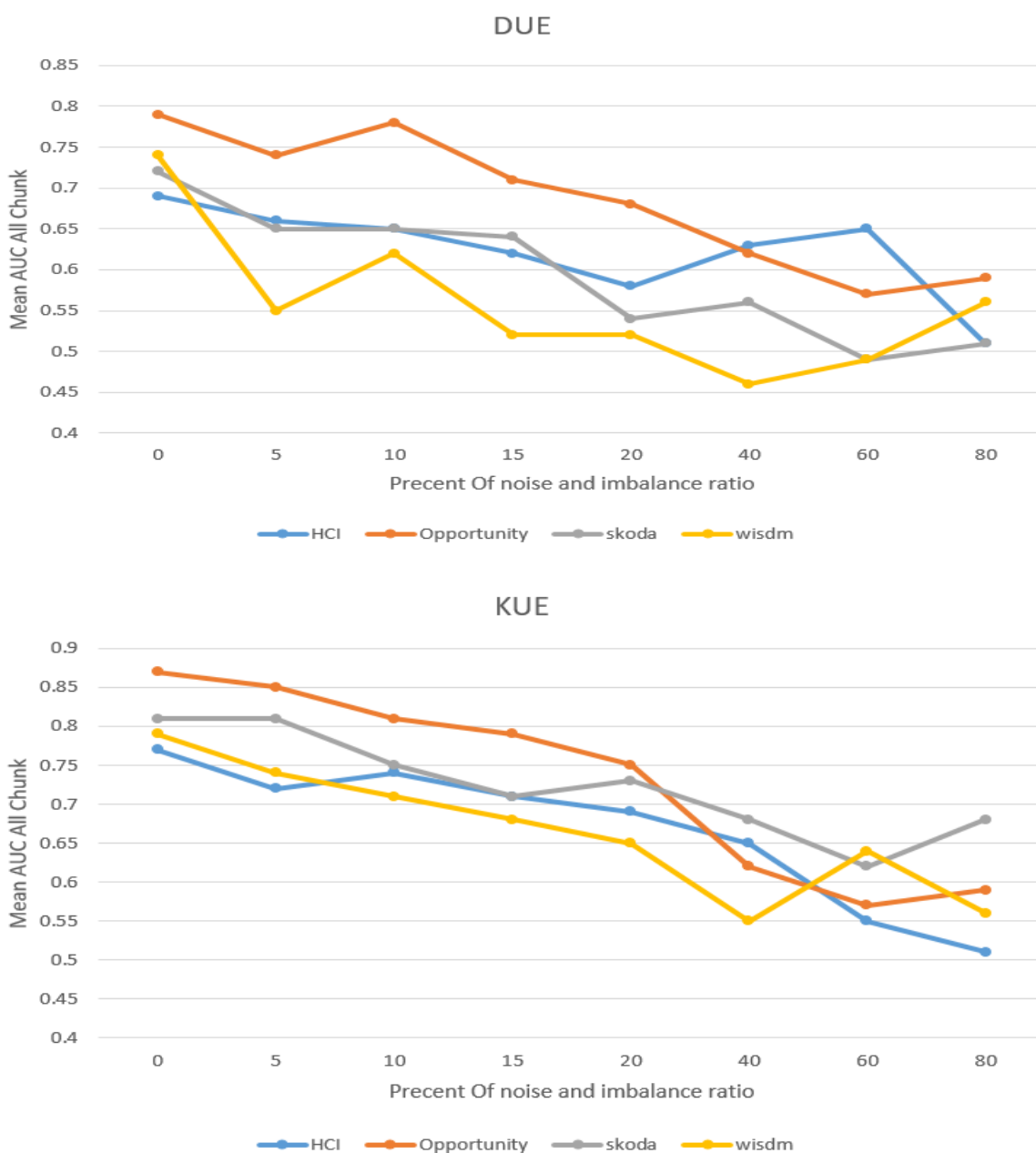




شکل ۳-۲: نمودارهای سناریوی اول (نویز متغیر) برای مجموعه داده‌های تشخیص فعالیت

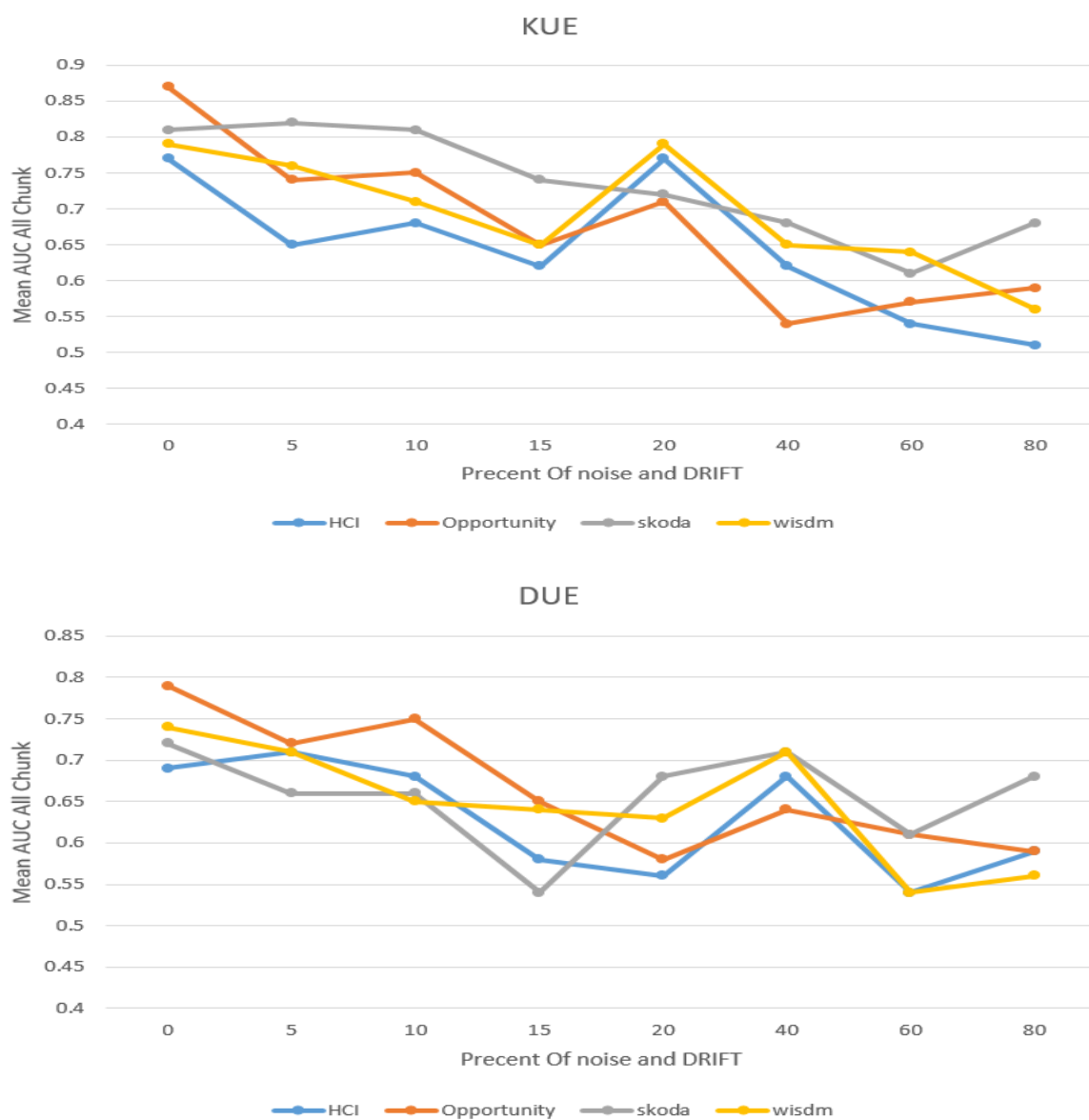
در ادامه، نمودارهای مربوط به الگوریتم‌ها بر روی مجموعه داده‌های تشخیص فعالیت در سناریوی دوم مشاهده می‌شود.

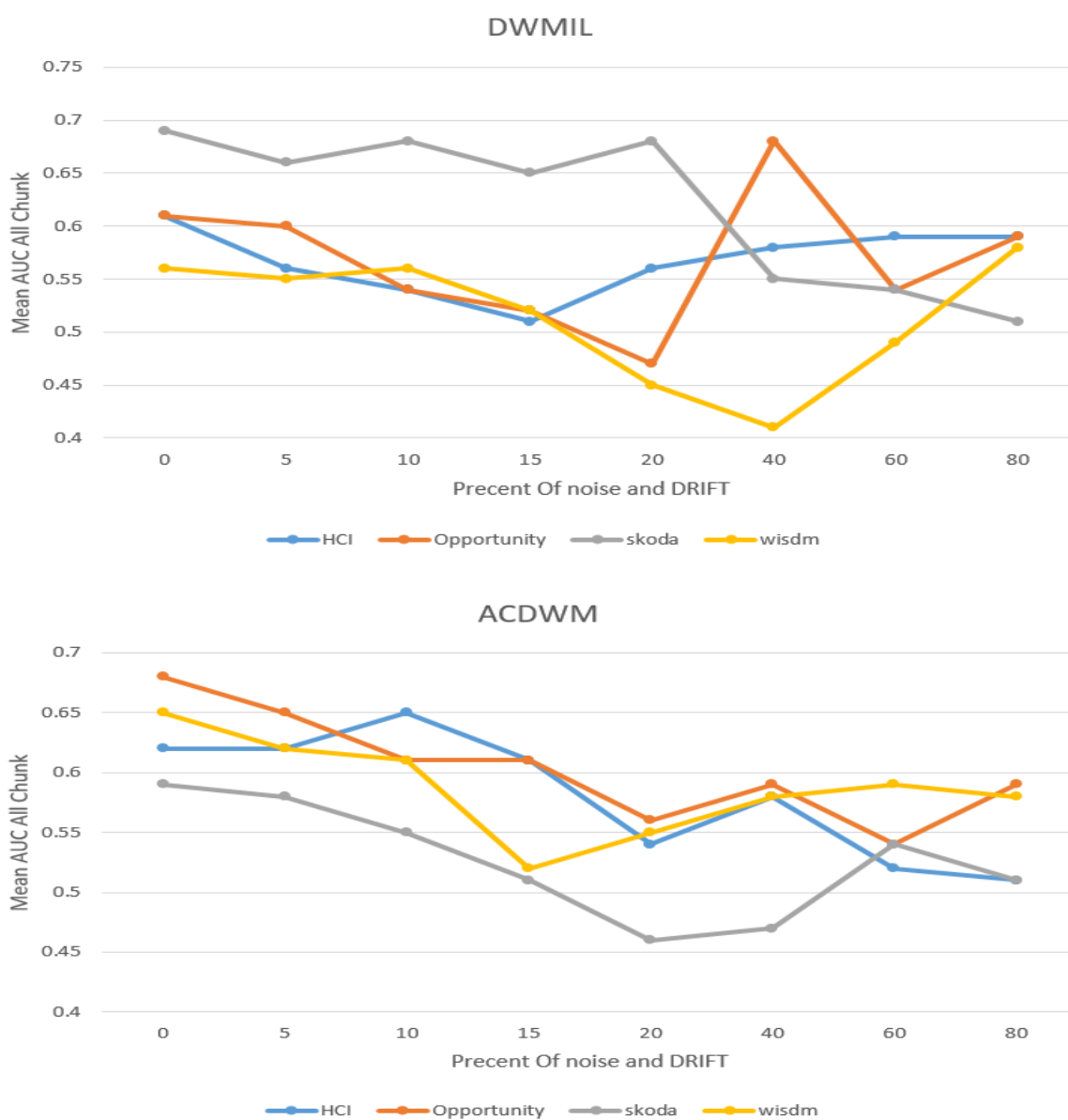




شکل ۳-۳: نمودارهای سناریوی دوم (رانس مفهوم صفر) برای مجموعه داده‌های تشخیص فعالیت

در ادامه نمودارهای مربوط به الگوریتم‌ها بر روی مجموعه داده‌های تشخیص فعالیت در سناریوی سوم مشاهده می‌شود.





شکل ۳-۴: نمودارهای سناریوی سوم (نرخ عدم توازن صفر) برای مجموعه داده‌های تشخیص فعالیت

### ۳-۳ مراحل راهکار پیشنهادی

#### ۳-۳-۱ اعمال نویز گاوسی به الگوریتم‌ها

ابتدا نویز گاوسی طبق رابطه‌ی ۳-۱ به الگوریتم‌ها اعمال می‌شود. درصد نویز به ترتیب ۰، ۵، ۱۰، ۱۵، ۲۰، ۴۰، ۶۰ و ۸۰ است.

#### ۳-۳-۲ به دست آوردن نمودارها و محاسبه‌ی معیار ناحیه‌ی زیر منحنی

نمودارهای ۴ الگوریتم در ۳ سناریوی ذکر شده برای مجموعه‌داده‌های تشخیص فعالیت به طور جداگانه به دست می‌آید. این نمودارها معیار میانگین ناحیه‌ی زیر منحنی برای همه‌ی قطعه‌ها را نشان می‌دهند.

#### ۳-۳-۳ محاسبه‌ی انحراف معیار

انحراف معیار با جای‌گذاری میانگین ناحیه‌ی زیر منحنی برای همه‌ی قطعه‌ها در حالت بدون نویز به جای میانگین در رابطه‌ی ۳-۱۱ محاسبه می‌شود. به دلیل تغییرات زیاد نتایج نشان‌داده‌شده در نمودارها، انحراف معیار برای مقایسه انتخاب می‌شود. این معیار برای هر الگوریتم به ازای همه‌ی مجموعه‌داده‌ها محاسبه می‌شود.

#### ۳-۳-۴ محاسبه‌ی انحراف معیار الگوریتم

انحراف معیار الگوریتم با جای‌گذاری اعداد مرحله‌ی ۳-۳-۳ در رابطه‌ی ۳-۱۱ محاسبه می‌گردد. این معیار نشان‌دهنده‌ی استحکام الگوریتم روی مجموعه‌داده‌های تشخیص فعالیت است. این معیار برای همه‌ی الگوریتم‌ها محاسبه می‌شود. در این مرحله،  $\mu$  در رابطه‌ی ۳-۱۱ همان میانگین نتایج در مرحله‌ی ۳-۳-۳ است.

#### ۳-۳-۵ مقایسه و نتیجه‌گیری

الگوریتمی که انحراف معیار کمتری دارد، به عنوان الگوریتم با استحکام بیشتر شناخته می‌شود.



### ۳-۴ نتیجه‌گیری

در این فصل به بیان روابط به‌دست‌آمده از آزمایش‌ها بر روی چهار الگوریتم DUE، KUE، DWMIL و ACDWM پرداخته شد. هر الگوریتم بر روی ۴ مجموعه داده مورد بررسی و ارزیابی قرار گرفت و رابطه‌ی کارآمدی از بررسی این نتایج به‌دست‌آمد. در فصل بعد به نتایج حاصل از آزمایش‌های مختلف خواهیم پرداخت و نتیجه‌گیری نهایی تشریح خواهد شد.

## فصل چهارم

### ارزیابی راهکار پیشنهادی

## ۴-۱ مقدمه

در سال‌های اخیر مطالعه‌ی جریان‌داده‌ها مورد توجه زیادی قرار گرفته است. در اکثر مطالعات انجام‌شده، توزیع طبقاتی جریان‌داده‌ها نسبتاً متوازن است. اگر یک جریان‌داده نامتوازن باشد، برخورد با مسئله‌ی رانش مفهوم دشوارتر است. روش‌های نمونه‌گیری کنونی عموماً با ارزیابی شباهت نمونه‌های مجموعه‌ی اقلیت گذشته و نمونه‌های فعلی، نمونه‌های دریافتی گذشته را به‌صورت انتخابی در مجموعه‌ی فعلی جذب می‌کنند. با این حال، ارزیابی شباهت به‌راحتی تحت‌تأثیر عوامل دشواری داده‌ها قرار می‌گیرد. این تکنیک‌های نمونه‌برداری بیش از حد، توزیع طبقه‌ی اکثریت را نادیده گرفته‌اند؛ بنابراین خطر هم‌پوشانی، طبقات را تهدید می‌کنند. در این فصل به بیان نتایج حاصل از آزمایش‌های مختلف خواهیم پرداخت.

## ۴-۲ نتایج ارزیابی

در این بخش نتایج حاصل از ارزیابی ۴ الگوریتم بر روی ۴ مجموعه‌داده با استفاده از درصد نویزهای مختلف نمایش داده می‌شود. لازم به ذکر است نتایج گزارش‌شده از گرفتن میانگین ۱۰ بار اجرای مستقل به دست می‌آید تا پایداری نتایج حفظ شود. همچنین نتایج حاصل از گرفتن میانگین تمام قطعه‌های آزمایش شده است. معیار مورد ارزیابی ناحیه‌ی زیر منحنی است که در فصل قبل کامل شرح داده شد.

در الگوریتم KUE از ترکیب قطعه‌بندهایی با نرخ وزن کاپا استفاده شده است. ما هنگام طراحی یک سامانه-ی یادگیری انتخاب‌های فراوانی مانند نحوه‌ی نمایش، پارامترهای یادگیری و داده‌های آموزشی داریم. این تنوع سبب ایجاد نوعی از واریانس در عملکرد سامانه می‌شود. در نتیجه اگر سامانه‌های مختلفی وجود داشته باشد و از نتایج آنها استفاده شود، این امکان وجود دارد که توزیع خطا، حول هدف متمرکز شود و با افزایش نمونه‌گیری از این توزیع به نتیجه‌ی بهتری برسیم. وقتی تعداد نمونه‌ها بیشتر شود احتمال افزایش خطا در خروجی با اضافه‌کردن نویز به داده‌ها بسیار بالا می‌رود. در نتیجه خروجی نهایی در برخی از نمودارها مشاهده می‌شود. برای گرفتن نتیجه‌ی مناسب، طبقه‌بندهای مورد استفاده در روش‌های گروهی طبقه‌بندها باید شرایط زیر را داشته باشند:

۱) هر یک به تنهایی در حد قابل قبولی دقیق باشند. البته نیازی به دقت کامل آنها نیست.

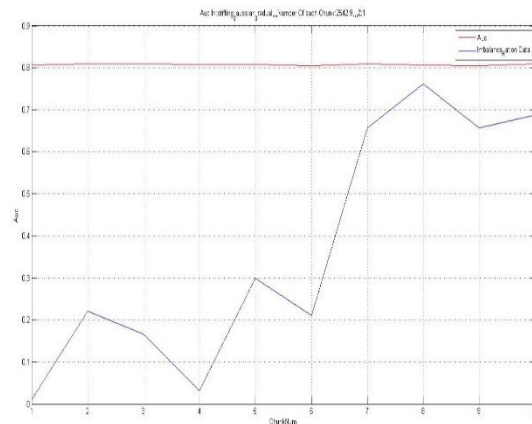
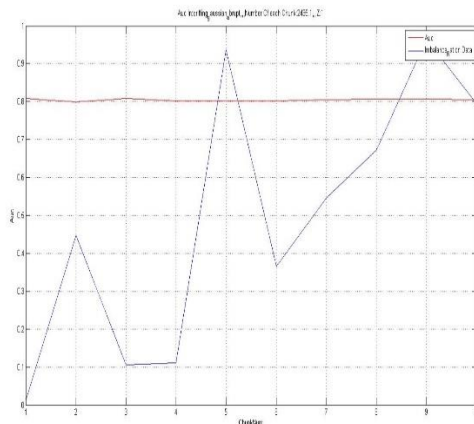
۲) هر کدام به عنوان مکمل دیگری عمل کنند. به این معنا که همگی نباید مشابه هم باشند و نتیجه یکسانی تولید کنند.

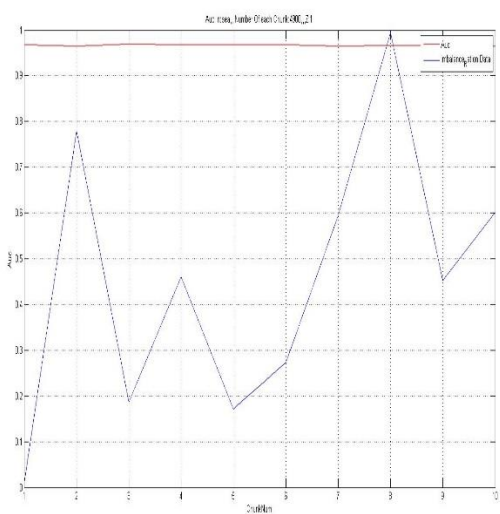
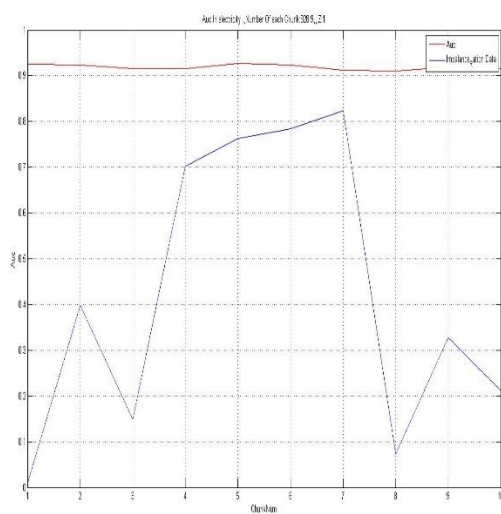
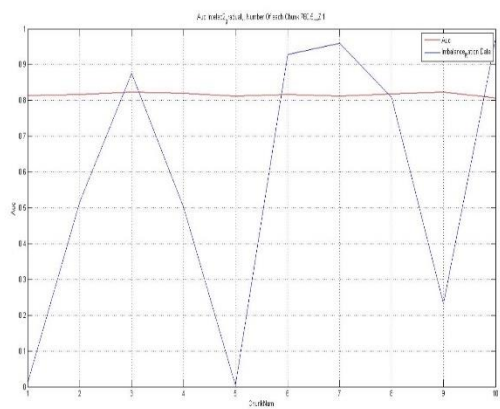
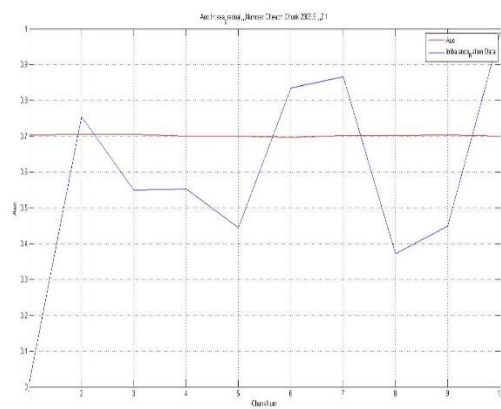
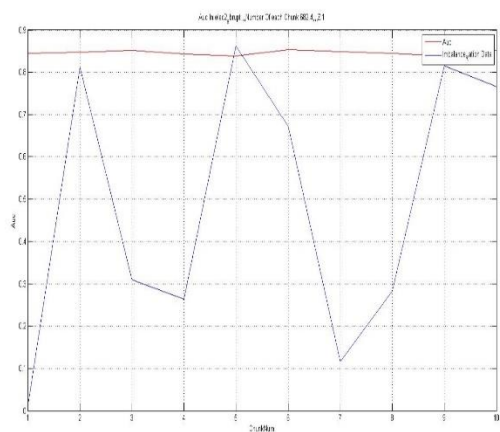
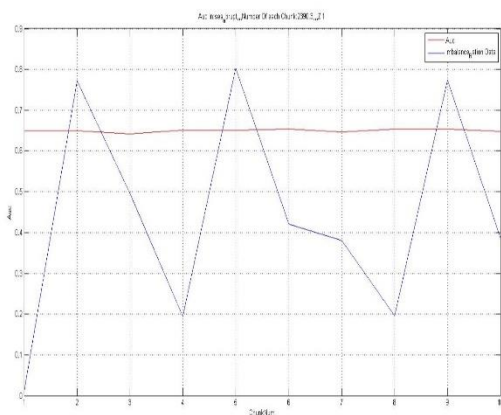
با افزایش نویز، شرط اول ذکر شده نقض می‌شود. در نتیجه خروجی نهایی افت شدیدی خواهد داشت.

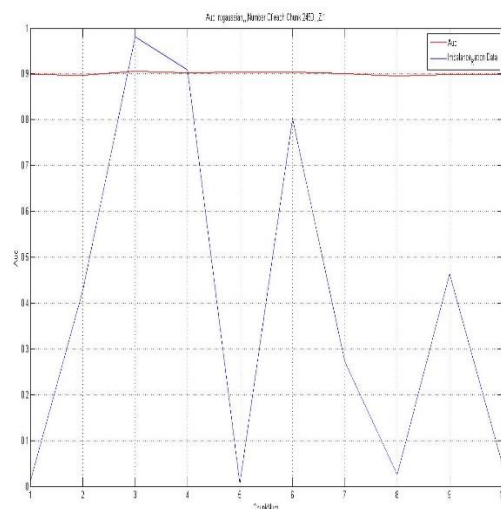
مقدار اندازه‌ی قطعه به صورت آزمون و خطا انتخاب می‌شود. الگوریتم مورد نظر باید با مقادیر مختلف اندازه‌ی قطعه مورد آزمایش و اعتبارسنجی قرار گیرد تا بهترین مقدار انتخاب شود. اندازه‌ی قطعه به طور پویا توسط آزمون‌های فرضیه‌ی آماری انتخاب می‌شود تا طبقه‌بند بر روی قطعه‌ی داده‌ی فعلی به اندازه‌ی کافی پایدار باشد. یکی از دلایل نوسان شدید الگوریتم ACDWM می‌تواند این مورد باشد. همچنین نویز در داده‌ها نیز می‌تواند در نوسان خروجی تأثیرگذار باشد. افزایش نویز سبب تخریب داده‌ها می‌شود و امکان خروجی مناسب را از بین می‌برد.

باتوجه به اینکه نویز وارد شده ممکن است نتایج را بهبود دهد یا از بین ببرد، در نتیجه نوسان خروجی بسیار بالا خواهد بود. از انحراف معیار الگوریتم برای بررسی استحکام استفاده می‌شود.

تمامی الگوریتم‌ها و مجموعه داده‌ها با نرخ‌های عدم توازن مختلف مورد آزمایش و ارزیابی قرار گرفتند. نتایج این ارزیابی به شرح زیر است:

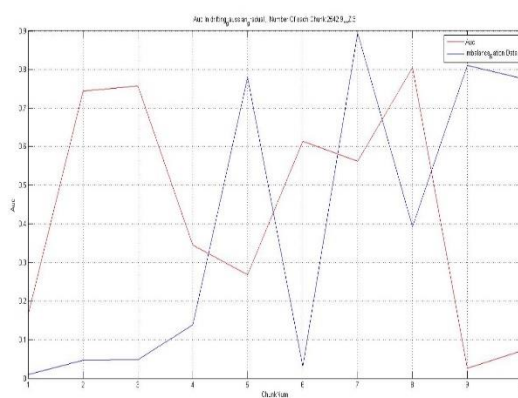
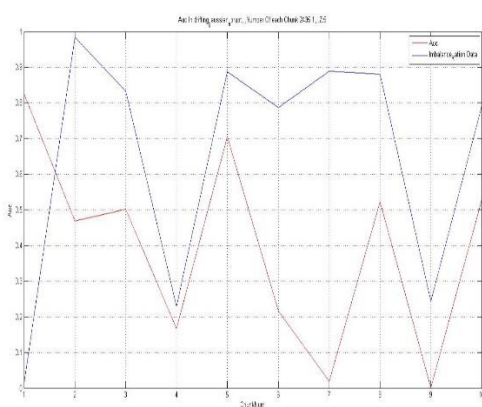


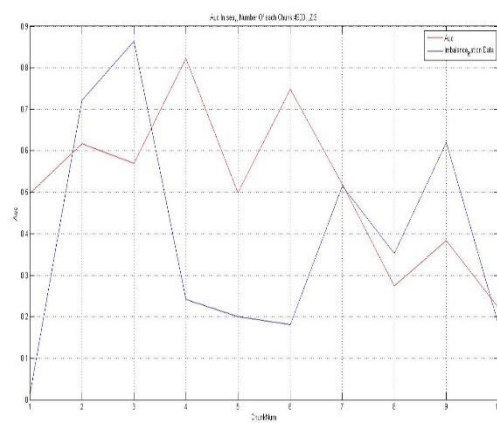
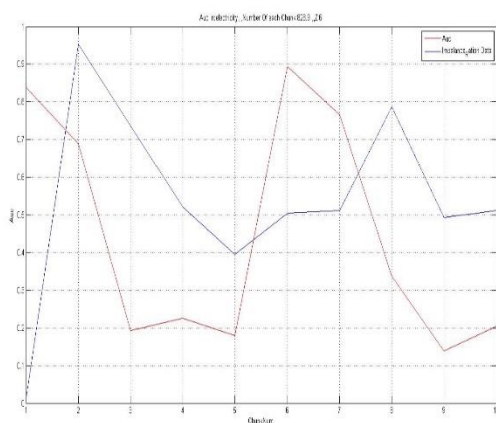
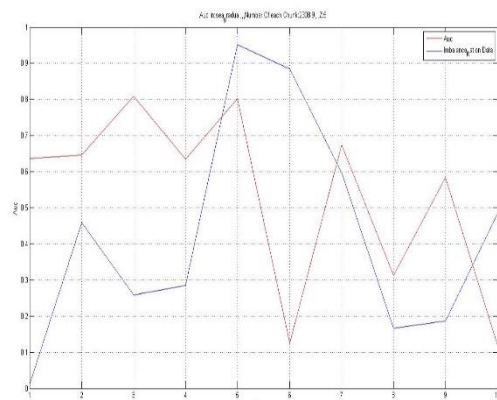
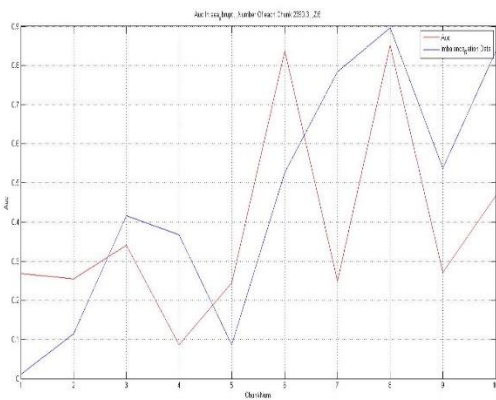
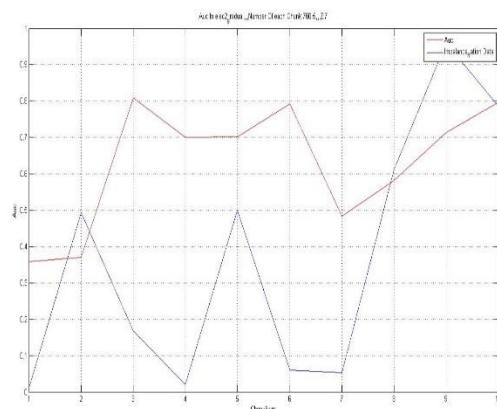
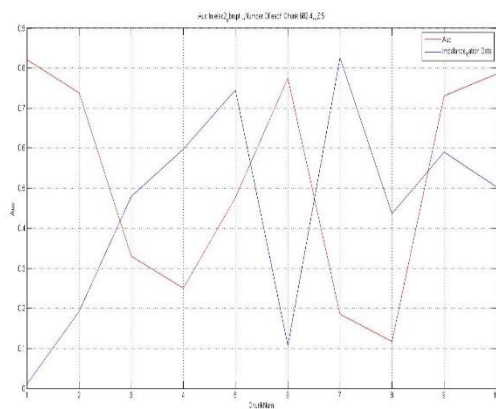


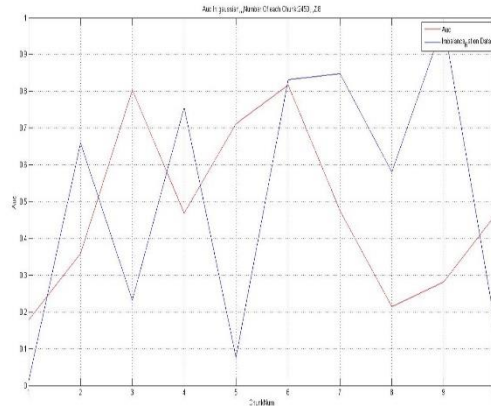


شکل ۴-۱: نمودارهای حاصل از نرخ عدم توازن متغیر با نويز صفر

در نمودارهای فوق، نرخ عدم توازن تغییر داده شده است و میانگین نتایج برای قطعه‌های مختلف ثبت شده است. نتایج نشان داده است در حالتی که نويز صفر است، تغییرات محسوسی در نتایج پایانی با ایجاد تغییر در نرخ عدم توازن ایجاد نمی‌شود. به‌منظور بررسی استحکام و تأثیر آن بر روی نرخ عدم توازن، این آزمایش با درصدهای نويز مختلف انجام شد. نتایج حاصل از نويز ۴۰ درصد در نمودارهای زیر نمایش داده شده است.







شکل ۴-۲: نتایج حاصل از نویز ۴۰ درصد

با اضافه شدن نویز به آزمایش‌های قبلی، نتایج نهایی به شدت تحت تأثیر قرار گرفته است. استحکام یک الگوریتم را می‌توان به دلایل زیر نسبت داد:

(۱) افزایشی بودن: این مفهوم نشان می‌دهد که الگوریتم هیچ نیازی به داده‌های قبلی برای پیش‌بینی ندارد که این امر سبب صرفه‌جویی در ذخیره‌سازی می‌شود. با توجه به این که داده‌های قبلی می‌توانند بر روی نتایج فعلی تأثیرات غیر قابل پیش‌بینی بگذارند، پس عدم استفاده از آن‌ها می‌تواند در استحکام الگوریتم تأثیرگذار باشد. البته این تأثیرگذاری به صورت همیشگی نیست. اگر میزان داده‌های قبلی دخیل در پیش‌بینی، بسیار کم باشد یا داده‌های قبلی بسیار شبیه داده‌های فعلی باشد، افزایشی بودن می‌تواند بر استحکام الگوریتم تأثیر بگذارد.

(۲) تعداد طبقه‌بندها<sup>۱</sup>: تعداد طبقه‌بندهای دخیل در روش رأی‌گیری می‌تواند در استحکام الگوریتم تأثیرگذار باشد. اضافه شدن تعداد طبقه‌بندها، می‌تواند به‌طور کلی نتایج را جابه‌جا کند و تغییرات زیادی در نتیجه رأی‌گیری بگذارد. پس می‌توان این عامل را به‌عنوان یک عامل تأثیرگذار در استحکام الگوریتم مطرح کرد. باید توجه داشت تغییرات در تعداد طبقه‌بندها به میزان زیاد می‌تواند در استحکام الگوریتم‌ها تأثیر بگذارد. تغییرات کم تأثیر زیادی در این زمینه نخواهد داشت.

<sup>۱</sup> Number of classifier



۳) اندازه‌ی قطعه: این پارامتر می‌تواند تعداد طبقه‌بندها را تغییر دهد، زیرا بر روی هر قطعه یک طبقه‌بند در روش گروهی ایجاد می‌شود. با توجه به توضیحات قبلی، تعداد طبقه‌بندها در استحکام الگوریتم تأثیرگذار هستند. در نتیجه با افزایش بیش از حد یا کاهش بیش از حد این متغیر، استحکام الگوریتم دچار تغییر می‌شود. (این مورد باعث تفاوت در نتایج DWMIL و ACDWM می‌شود).

۴) اکثریت وزنی<sup>۱</sup>: رأی‌گیری وزن‌دار بدین مفهوم است که برخی از قطعه‌بندها، درصد رأی بیشتری در خروجی نهایی خواهند داشت. یک قطعه‌بند به‌جای یک رأی می‌تواند چندین رأی را داشته باشد. مشخص است که این عامل استحکام الگوریتم را دچار تغییر می‌کند، زیرا قطعه‌بند با درجه رأی بالاتر می‌تواند قدرت بیشتری در انتخاب خروجی نهایی داشته باشد.

۵) وزن‌دهی پویا<sup>۲</sup>: در این روش برای ساخت هر طبقه‌بند از یک مدل وزن‌دار استفاده می‌شود. بدین صورت که هر مجموعه‌داده‌ی انتخاب‌شده برای هر طبقه‌بند بر مبنای وزن نمونه‌ها تعیین می‌گردد. برای مثال اگر در مجموعه طبقه‌بند اول، نمونه پنجم خروجی درستی نداشت، این نمونه دارای وزن مضاعف شده و در دور بعدی اجرا به تعداد بیشتر تکرار می‌شود. میزان وزن اضافه شده در هر بار تکرار متفاوت است و این امر می‌تواند در استحکام الگوریتم تأثیرگذار باشد.

جدول ۴-۱ میانگین ناحیه‌ی زیر منحنی در همه‌ی قطعه‌ها را برای سناریوی اول نشان می‌دهد. از اعداد این جدول برای محاسبه‌ی انحراف استاندارد استفاده می‌شود.

جدول ۴-۱: نتایج آزمایش‌ها در سناریوی اول

درصد نویز								مجموعه‌داده	الگوریتم
۸۰	۶۰	۴۰	۲۰	۱۵	۱۰	۵	۰		
۰/۶۴	۰/۵۶	۰/۴۸	۰/۶۱	۰/۶۲	۰/۵۸	۰/۶۵	۰/۶۸	Opportunity	ACDWM
۰/۶۱	۰/۶۳	۰/۵۱	۰/۵۲	۰/۵۵	۰/۵۵	۰/۶۱	۰/۶۵	WISDM	

<sup>۱</sup> Weighted majority<sup>۲</sup> Dynamic weighting

۰/۵۵	۰/۵۵	۰/۵۱	۰/۶۱	۰/۵۷	۰/۵۵	۰/۵۹	۰/۶۲	HCI	DWMIL
۰/۴۷	۰/۴۹	۰/۴۸	۰/۴۹	۰/۶۱	۰/۶۳	۰/۶۰	۰/۵۹	skoda	
۰/۶۴	۰/۵۱	۰/۴۸	۰/۶۱	۰/۶۸	۰/۵۱	۰/۵۱	۰/۶۱	Opportunity	
۰/۷۴	۰/۶۳	۰/۵۱	۰/۵۲	۰/۵۵	۰/۵۵	۰/۴۸	۰/۵۶	WISDM	
۰/۶۳	۰/۶۸	۰/۵۱	۰/۵۸	۰/۵۷	۰/۵۱	۰/۶۸	۰/۶۱	HCI	
۰/۵۱	۰/۴۹	۰/۵۶	۰/۵۵	۰/۶۱	۰/۶۳	۰/۶۵	۰/۶۹	skoda	
۰/۶۴	۰/۵۱	۰/۴۸	۰/۶۸	۰/۶۸	۰/۵۱	۰/۶۴	۰/۷۹	Opportunity	DUE
۰/۷۴	۰/۶۳	۰/۵۱	۰/۶۸	۰/۵۵	۰/۶۸	۰/۴۸	۰/۷۴	WISDM	
۰/۶۳	۰/۶۸	۰/۵۱	۰/۵۸	۰/۶۵	۰/۶۴	۰/۶۸	۰/۶۹	HCI	
۰/۵۱	۰/۷۴	۰/۶۸	۰/۵۱	۰/۶۴	۰/۵۸	۰/۵۱	۰/۷۲	skoda	
۰/۶۴	۰/۵۱	۰/۴۸	۰/۵۱	۰/۶۸	۰/۵۹	۰/۶۴	۰/۸۷	Opportunity	KUE
۰/۶۵	۰/۵۱	۰/۷۱	۰/۶۸	۰/۵۵	۰/۴۴	۰/۴۸	۰/۷۹	WISDM	
۰/۵۱	۰/۶۸	۰/۶۴	۰/۶۸	۰/۶۵	۰/۵۱	۰/۴۸	۰/۷۷	HCI	
۰/۶۴	۰/۷۴	۰/۵۱	۰/۴۸	۰/۶۴	۰/۵۸	۰/۵۱	۰/۸۱	skoda	

در جدول ۲-۴ نتایج محاسبه‌ی انحراف معیار برای مقایسه‌ی استحکام الگوریتم‌ها در سناریوی اول برای مجموعه داده‌های تشخیص فعالیت مورد بررسی قرار می‌گیرد. این اعداد تا دقت ۲ رقم اعشار محاسبه می‌شوند. ستون آخر جدول ۲-۴ انحراف معیار الگوریتم را نشان می‌دهد که با گرفتن انحراف معیار از همه‌ی مجموعه داده‌های موجود در جدول ۲-۴ با رابطه‌ی ۳-۱۱ به دست می‌آید. در این فرمول برای محاسبه‌ی انحراف معیار الگوریتم،  $\mu$  میانگین انحراف معیار مجموعه داده‌ها است.

جدول ۲-۴: انحراف معیار نتایج در سناریوی اول بر روی مجموعه داده‌های تشخیص فعالیت

انحراف معیار الگوریتم	انحراف معیار مجموعه داده				الگوریتم
	WISDM	skoda	Opportunity	HCI	
۰/۰۱۱	۰/۰۹	۰/۰۸	۰/۱۰	۰/۰۷	ACDWM
۰/۰۲۳	۰/۰۸	۰/۱۳	۰/۰۸	۰/۰۷	DWMIL
۰/۰۴۳	۰/۱۶	۰/۱۵	۰/۲۱	۰/۰۹	DUE
۰/۰۳۹	۰/۲۴	۰/۲۴	۰/۳۰	۰/۱۹	KUE

الگوریتم‌های ACDWM و DWMIL از DUE و KUE انحراف استاندارد کمتری دارند؛ بنابراین؛ ACDWM و DWMIL روی مجموعه داده‌های بررسی شده استحکام بهتری دارند.

جدول ۳-۴ میانگین ناحیه‌ی زیر منحنی به دست آمده از اجرای الگوریتم‌ها در سناریوی دوم را نمایش می‌دهد.

جدول ۳-۴: نتایج آزمایش‌ها در سناریوی دوم

درصد نویز								مجموعه داده	الگوریتم
۸۰	۶۰	۴۰	۲۰	۱۵	۱۰	۵	۰		
۰/۵۹	۰/۵۶	۰/۵۸	۰/۶۱	۰/۵۸	۰/۵۸	۰/۶۱	۰/۶۸	Opportunity	ACDWM
۰/۶۱	۰/۵۸	۰/۵۱	۰/۵۱	۰/۵۸	۰/۶۲	۰/۶۴	۰/۶۵	WISDM	
۰/۵۱	۰/۵۲	۰/۵۱	۰/۶۱	۰/۶۱	۰/۶۰	۰/۶۱	۰/۶۲	HCI	
۰/۵۲	۰/۴۹	۰/۵۷	۰/۴۹	۰/۵۸	۰/۶۳	۰/۵۶	۰/۵۹	skoda	
۰/۵۹	۰/۵۷	۰/۵۹	۰/۶۷	۰/۵۸	۰/۵۸	۰/۵۹	۰/۶۱	Opportunity	DWMIL
۰/۵۱	۰/۴۹	۰/۴۶	۰/۴۸	۰/۵۲	۰/۶۲	۰/۵۵	۰/۵۶	WISDM	

۰/۵۱	۰/۵۲	۰/۵۵	۰/۵۵	۰/۵۴	۰/۶۱	۰/۶۲	۰/۶۱	HCI	DUE
۰/۵۸	۰/۴۹	۰/۵۶	۰/۵۴	۰/۵۸	۰/۶۵	۰/۶۲	۰/۶۹	skoda	
۰/۵۹	۰/۵۷	۰/۶۲	۰/۶۸	۰/۷۱	۰/۷۸	۰/۷۴	۰/۷۹	Opportunity	
۰/۵۶	۰/۴۹	۰/۴۶	۰/۵۲	۰/۵۲	۰/۶۲	۰/۵۵	۰/۷۴	WISDM	
۰/۵۱	۰/۶۵	۰/۶۳	۰/۵۸	۰/۶۲	۰/۶۵	۰/۶۶	۰/۶۹	HCI	
۰/۵۱	۰/۴۹	۰/۵۶	۰/۵۴	۰/۶۴	۰/۶۵	۰/۶۵	۰/۷۲	skoda	
۰/۵۹	۰/۵۷	۰/۶۲	۰/۷۵	۰/۷۹	۰/۸۱	۰/۸۵	۰/۸۷	Opportunity	KUE
۰/۵۶	۰/۶۴	۰/۵۵	۰/۶۵	۰/۶۸	۰/۷۱	۰/۷۴	۰/۷۹	WISDM	
۰/۵۱	۰/۵۵	۰/۶۵	۰/۶۹	۰/۷۱	۰/۷۴	۰/۷۲	۰/۷۷	HCI	
۰/۶۸	۰/۶۲	۰/۶۸	۰/۷۳	۰/۷۱	۰/۷۵	۰/۸۱	۰/۸۱	skoda	

در جدول ۴-۴ محاسبه‌ی انحراف معیار برای مقایسه‌ی الگوریتم‌ها از جهت استحکام بر روی مجموعه‌داده‌های تشخیص فعالیت در سناریوی دوم نمایش داده می‌شود.

جدول ۴-۴: انحراف معیار نتایج در سناریوی دوم بر روی مجموعه‌داده‌های تشخیص فعالیت

انحراف معیار الگوریتم	انحراف معیار مجموعه‌داده				الگوریتم
	WISDM	skoda	Opportunity	HCI	
۰/۰۱۳	۰/۰۹	۰/۰۶	۰/۰۹	۰/۰۷	ACDWM
۰/۰۳۶	۰/۰۶	۰/۱۳	۰/۰۳	۰/۰۷	DWMIL
۰/۰۴۳	۰/۲۱	۰/۱۶	۰/۱۴	۰/۰۹	DUE
۰/۰۲۹	۰/۱۶	۰/۱۱	۰/۱۹	۰/۱۴	KUE

در جدول ۴-۵ اعداد به دست آمده از اجرای الگوریتم‌ها در سناریوی سوم نمایش داده می‌شود.

جدول ۴-۵: نتایج آزمایش‌ها در سناریوی سوم

درصد نويز								مجموعه داده	الگوریتم
۸۰	۶۰	۴۰	۲۰	۱۵	۱۰	۵	۰		
۰/۵۹	۰/۵۴	۰/۵۹	۰/۵۶	۰/۶۱	۰/۶۱	۰/۶۵	۰/۶۸	Opportunity	ACDWM
۰/۶۱	۰/۵۸	۰/۵۱	۰/۵۱	۰/۵۸	۰/۶۲	۰/۶۴	۰/۶۵	WISDM	
۰/۵۱	۰/۵۲	۰/۵۸	۰/۵۴	۰/۶۱	۰/۶۵	۰/۶۲	۰/۶۲	HCI	
۰/۵۱	۰/۵۴	۰/۴۷	۰/۴۶	۰/۵۱	۰/۵۵	۰/۵۸	۰/۵۹	skoda	
۰/۵۹	۰/۵۴	۰/۶۸	۰/۴۷	۰/۵۲	۰/۵۴	۰/۶۰	۰/۶۱	Opportunity	
۰/۵۸	۰/۴۹	۰/۴۱	۰/۴۵	۰/۵۲	۰/۵۶	۰/۵۵	۰/۵۶	WISDM	DWMIL
۰/۵۹	۰/۵۹	۰/۵۸	۰/۵۶	۰/۵۱	۰/۵۴	۰/۵۶	۰/۶۱	HCI	
۰/۵۱	۰/۵۴	۰/۵۵	۰/۶۸	۰/۶۵	۰/۶۸	۰/۶۶	۰/۶۹	skoda	
۰/۵۹	۰/۶۱	۰/۶۴	۰/۵۸	۰/۶۵	۰/۷۵	۰/۷۲	۰/۷۹	Opportunity	
۰/۵۶	۰/۵۴	۰/۷۱	۰/۶۳	۰/۶۴	۰/۶۵	۰/۷۱	۰/۷۴	WISDM	DUE
۰/۵۹	۰/۵۴	۰/۶۸	۰/۵۶	۰/۵۸	۰/۶۸	۰/۷۱	۰/۶۹	HCI	
۰/۶۸	۰/۶۱	۰/۷۱	۰/۶۸	۰/۵۴	۰/۶۶	۰/۶۶	۰/۷۲	skoda	
۰/۵۹	۰/۵۷	۰/۵۴	۰/۷۱	۰/۶۵	۰/۷۵	۰/۷۴	۰/۸۷	Opportunity	
۰/۵۸	۰/۵۹	۰/۵۸	۰/۵۵	۰/۵۲	۰/۶۱	۰/۶۲	۰/۶۵	WISDM	KUE
۰/۵۱	۰/۵۴	۰/۶۲	۰/۷۷	۰/۶۲	۰/۶۸	۰/۶۵	۰/۷۷	HCI	

skoda	۰/۸۱	۰/۸۲	۰/۸۱	۰/۷۴	۰/۷۲	۰/۶۸	۰/۶۱	۰/۶۸
-------	------	------	------	------	------	------	------	------

نتایج حاصل از محاسبه‌ی انحراف معیار بر روی مجموعه‌داده‌های تشخیص فعالیت در سناریوی سوم در جدول ۴-۶ مشاهده می‌شود.

جدول ۴-۶: انحراف معیار نتایج در سناریوی سوم بر روی مجموعه‌داده‌های تشخیص فعالیت

انحراف معیار الگوریتم	انحراف معیار مجموعه‌داده				الگوریتم
	WISDM	skoda	Opportunity	HCI	
۰/۰۰۷	۰/۰۸	۰/۰۸	۰/۰۹	۰/۰۷	ACDWM
۰/۰۱۴	۰/۰۸	۰/۱۰	۰/۰۸	۰/۰۶	DWMIL
۰/۰۲۷	۰/۱۲	۰/۰۸	۰/۱۵	۰/۰۹	DUE
۰/۰۴۵	۰/۱۳	۰/۱۱	۰/۲۳	۰/۱۶	KUE

#### ۴-۳ نتیجه‌گیری

در این فصل به بیان نتایج نهایی حاصل از آزمایش‌های مختلف پرداختیم. در این آزمایش‌ها از مقادیر مختلف نویز، نرخ عدم توازن و رانش مفهوم استفاده شد تا تأثیرپذیری الگوریتم‌ها از این ۳ معیار مورد بررسی قرار گیرد. در نتیجه دلایل استحکام الگوریتم‌ها با وارد کردن درصدهای متفاوتی از نویز به آن‌ها و بررسی تغییرات معیار ناحیه‌ی زیر منحنی بر اساس نمودارها شرح داده شد. در فصل بعدی به بیان نتیجه‌گیری کلی خواهیم پرداخت.

#### ۴-۳-۱ سناریوی اول

سناریوی اول، کار اصلی ارائه شده در این پایان‌نامه است. نتیجه‌ی آن استحکام بیشتر الگوریتم‌های ACDWM و DWMIL نسبت به DUE و KUE است. انحراف معیار الگوریتم‌های ACDWM و

DWMIL از DUE و KUE بر روی مجموعه داده‌های تشخیص فعالیت در سناریوی اول کمتر است. پس ACDWM و DWMIL استحکام بیشتری نسبت به دو الگوریتم دیگر دارند. علاوه بر این، الگوریتم ACDWM با انحراف معیار کمتر از DWMIL بر روی مجموعه داده‌های تشخیص فعالیت، استحکام بیشتری نسبت به الگوریتم DWMIL دارد.

۲-۳-۴ سناریوی دوم

ACDWM بر روی مجموعه داده‌های تشخیص فعالیت نیز بر اساس انحراف معیار الگوریتم، استحکام بهتری دارد.

۳-۳-۴ سناریوی سوم

الگوریتم‌های ACDWM و DWMIL از DUE و KUE برای مجموعه داده‌های تشخیص فعالیت در سناریوی سوم، انحراف معیار الگوریتم کمتری دارند. پس استحکام ACDWM و DWMIL از دو الگوریتم دیگر بیشتر است. همچنین ACDWM با انحراف معیار الگوریتم کمتر برای مجموعه داده‌های تشخیص فعالیت، بهتر از DWMIL عمل می‌کند.

## فصل پنجم

### نتیجه گیری و پیشنهادها



## ۵-۱ مقدمه

عدم توازن در داده‌ها یکی از حیاتی‌ترین مشکلات در یادگیری ماشینی معاصر است. عدم توازن به تفاوت در نسبت تعداد نمونه‌ها در هر کلاس می‌پردازد، وقتی که برخی از کلاس‌ها به طور قابل توجهی نمونه‌های کمتری دارند، اکثر طبقه‌بندها با کلاس‌های ۱-۰ هدایت می‌شوند. کلاس‌های اقلیت کم نمایش، معمولاً مهم‌تر هستند، بنابراین باید مجموعه داده یا روش یادگیری را تغییر داد تا مرزهای تصمیم متعادل‌تری ایجاد شود که به نفع هیچ یک از کلاس‌ها نباشد. بسیاری از برنامه‌های کاربردی دنیای واقعی، از جمله تشخیص عیب، تشخیص ناهنجاری، تشخیص پزشکی و ... با عدم توازن در داده‌ها روبرو هستند. در این فصل به بیان نتیجه‌گیری کلی پایان‌نامه خواهیم پرداخت.

## ۵-۲ مزایا و معایب

### ۵-۲-۱ مزایا

(۱) قطعه‌بندی جریان داده‌ها، سبب انجام عملیات بر روی قطعه‌های کوچک داده می‌شود، پس افزایش سرعت در این روش محسوس است.

(۲) قطعه‌بندی سبب می‌شود تا تمرکز عملیات بر روی قطعه‌ی کوچکی قرار گیرد و در نهایت با بهره‌گیری از نتایج تمام قطعه‌ها، نتیجه نهایی اعلام می‌گردد. پس روش پیشنهادی دقت را بالا می‌برد.

### ۵-۲-۲ معایب

(۱) اگر نرخ عدم توازن در داده‌ها خیلی بالا باشد، قطعه‌بندی سبب می‌شود تا داده‌های نامتوازن در هر قطعه قرار گیرد و این امر سبب کاهش دقت می‌شود.

(۲) استفاده از روش‌های قطعه‌بندی نیاز به مقداره‌ی برخی پارامترهای اولیه دارد که در نتیجه‌ی نهایی تأثیرگذار است. اگر نتوان مقادیر مناسبی برای این پارامترها انتخاب کرد، نتایج پایانی افت شدیدی خواهد داشت. مقداره‌ی این پارامترها یکی از چالش‌های روش پیشنهادی است.

### ۵-۳ نتیجه‌گیری کلی

در سال‌های اخیر افزایش چشمگیر توانایی در جمع‌آوری مداوم داده‌ها مشاهده شده است. بیشتر این داده‌ها با رسیدن سریع، حجم بالا و طول بی‌نهایت مشخص می‌شوند؛ بنابراین به‌عنوان جریان داده‌ها شناخته می‌شوند. برنامه‌های کاربردی با استفاده از جریان داده‌ها در همه‌جا وجود دارند. تجزیه و تحلیل روند بازار سهام، نظارت بر سلامتی و... نمونه‌هایی از این برنامه‌ها هستند. در وظایف طبقه‌بندی سنتی، داده‌ها ثابت فرض می‌شوند و رانش مفهوم بدون تغییر است. در جریان داده‌ها، رانش مفهوم در طول زمان به دلیل تغییرات در محیط تغییر می‌کند.

به‌عنوان مثال، بازار سهام در نتیجه‌ی اقتصاد، درآمد شرکت‌ها، سیاست‌های پولی دولت و ... روزانه دچار نوسان می‌شود. تغییرات بی‌وقفه در طبقه‌بندی که از داده‌های گذشته به‌صورت موقتی به دست می‌آید، به‌روزرسانی می‌شود. آموزش مجدد مدل به دلیل رانش مفهوم ضروری است.

عدم‌توازن داده‌ها یک مشکل رایج در حوزه‌ی جریان داده‌ها است. در این‌گونه مسائل، داده‌ها می‌توانند عدم‌توازن داشته باشند یا ممکن است در طول زمان تکامل یابند. علاوه بر این، عدم‌توازن داده‌ها همراه با رانش مفهوم، چالش‌های جدید و منحصربه‌فردی را ایجاد می‌کند. نقش‌های کلاس ممکن است تغییر کند و مجموعه‌ی اکثریت تبدیل به اقلیت شود یا کلاس اقلیت به اکثریت مبدل شود. چندین طبقه ممکن است متفاوت از قبل باشند، طبقه‌های جدید ظاهر شود، طبقه‌های قدیمی ناپدید شود یا مشکلات سطح نمونه پدیدار شود. در این موارد نظارت بر هر کلاس برای تغییرات در ویژگی‌های آن کافی نیست، زیرا طبقه‌بند باید نرخ ورودی نمونه‌های جدید در هر کلاس را ردیابی کند.

عدم‌توازن کلاسیک یک مشکل عملی است و معمولاً زمانی رخ می‌دهد که نمونه‌های کمتری در کلاس هدف (کلاس مثبت یا اقلیت) در مقایسه با سایر کلاس‌ها (کلاس منفی یا اکثریت) وجود داشته باشد.

عدم‌توازن کلاس می‌تواند به دلیل ماهیت یک برنامه، محدودیت در جمع‌آوری یک مجموعه داده‌ی نماینده و در نتیجه‌ی مسائل مربوط به هزینه یا حریم خصوصی ایجاد شود. عدم‌توازن، کلاس چالش‌های متعددی را در وظایف یادگیری از جمله توزیع کلاسی ناهنجار، کافی نبودن داده‌ها و مفاهیم پیچیده‌تر ایجاد می‌کند.

از آنجایی که این چالش‌ها با رانش مفهوم در زمینه‌ی جریان‌داده‌ها تداخل می‌کنند، این مشکل حتی شدیدتر و پیچیده‌تر می‌شود. به‌عنوان مثال، در طول یک جریان‌داده، فاصله‌ی زمانی دریافت یک نمونه‌ی مثبت می‌تواند به طور غیرقابل‌پیش‌بینی طولانی باشد؛ بنابراین، جمع‌آوری نمونه‌های مثبت کافی در یک برچسب زمانی برای استنباط بی‌طرفانه‌ی عملکرد واقعی توصیف‌کننده‌ی این کلاس، همیشه دشوار است. علاوه بر این، نمونه‌های مثبت متوالی ممکن است از توزیع‌های خودسرانه‌ی متمایز استخراج شود؛ بنابراین اختلاف توزیع می‌تواند بین هر دو نمونه‌ی مثبتی که هم‌زمان دریافت نشده‌اند، وجود داشته باشد. الگوریتم‌های طراحی‌شده برای طبقه‌بندی جریان‌داده‌های نامتوازن باید چالش‌های ناشی از عدم توازن را در نظر بگیرند، درحالی‌که مجهز به مکانیزم‌های مؤثر برای مدیریت رانش مفهوم هستند.

در این پژوهش با بهره‌گیری از ۴ الگوریتم پرکاربرد (DUE، KUE، DWMIL و ACDWM) در طبقه‌بندی مبتنی بر قطعه‌ی جریان‌داده‌ها، رابطه‌ای به‌منظور مقایسه‌ی استحکام الگوریتم‌ها مطرح گردید. آزمایش‌های انجام‌شده نشان می‌دهد که نرخ عدم توازن برای هر قطعه مقدار ثابتی است. اگر نرخ عدم توازن بیشتر باشد، اندازه‌ی هر گروه طبقه‌بندی بزرگ‌تر می‌شود تا اطمینان حاصل شود که هر نمونه از کلاس اکثریت با احتمال معینی درگیر فرایند آموزش می‌شود.

نرخ عدم‌توازن در روش‌های طبقه‌بند، میزان تأثیر داده‌های نامتوازن را مشخص می‌کند. ایجاد قطعه‌بند بر روی داده‌های نامتوازن بسیار سخت و زمان‌بر است. هر چه این مقدار بیشتر باشد تأثیر داده‌های نامتوازن بیشتر خواهد بود و قطعه‌بند نامتوازن‌تر خواهد شد. درنتیجه پس از اعمال اندازه‌ی قطعه و نرخ عدم‌توازن یکسان در تمامی الگوریتم‌ها، مقدار ناحیه‌ی زیر منحنی در الگوریتم DWMIL تقریباً کمتر از الگوریتم ACDWM شده است.

#### ۴-۵ پیشنهادها

۱- می‌توان نتیجه‌ی الگوریتم‌ها را علاوه بر طبقه‌بندهای دودویی<sup>۱</sup> روی طبقه‌بندهای چند کلاسی<sup>۱</sup> با مجموعه‌داده‌های چند کلاسی برای تشخیص چندین فعالیت بررسی کرد.

<sup>۱</sup> Binary classifier

۲- پیشنهاد می شود عملکرد الگوریتم های قطعه بندی بر روی مجموعه داده های به دست آمده از چندین ساکن در خانه های هوشمند برای تشخیص فعالیت مورد بررسی قرار بگیرد. این الگوریتم ها به تشخیص فعالیت و تشخیص فرد یا افرادی می پردازد که فعالیت را انجام داده اند.

---

<sup>1</sup> Multi classifier

## منابع

- [1] G. Bhavanasi, et al., "Patient activity recognition using radar sensors and machine learning," *Neural. Comput. Appl.*, vol.34, no.18 pp. 16033–16048, Sep. 2022.
- [2] B. Nguyen, et al., "Trends in human activity recognition with focus on machine learning and power requirements," *Mach. Learn*, vol.5, pp. 100072, Sep. 2021.
- [3] F. Gu, et al., "A survey on deep learning for human activity recognition," *ACM Compu. Surv.*, vol.54, no. 177, pp. 1-34, Nov. 2022.
- [4] S. K. Yadav, et al., "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowl. Based. Syst.*, vol.223, pp. 106970, Jul. 2021.
- [5] E. Ramanujam, et al., "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review," *IEEE Sens. J.*, vol.21, pp. 13029 - 13040, Mar. 2021.
- [6] Y. Xu and T. T. Qiu, "Human activity recognition and embedded application based on convolutional neural network," *Artif. Intell.*, vol.1, no. 1, pp. 51-60, Nov. 2021.
- [7] S. Mekruksavanich and A. Jitpattanakul, "Lstm networks using smartphone data for sensor-based human activity recognition in smart homes," *Sens.*, vol.21, no. 5, pp. 1636, Feb. 2021.
- [8] H. H. Tesfamikael, et al., "Simulation of Eye Tracking Control based Electric Wheelchair Construction by Image Segmentation Algorithm," *Journal of Innovative Image Processing (JIIP)*, vol.3, no. 1, pp. 21-35, Apr. 2021.
- [9] H. Abdellahoum, et al., "CSFCM: An improved fuzzy C-Means image segmentation algorithm using a cooperative approach," *Expert Syst. Appl.*, vol.166, pp. 114063, Mar. 2021.
- [10] K. Kowsari, et al., "Text classification algorithms: A survey," *Inf.*, vol.10, no. 4, pp. 150, Apr. 2019.

- 
- [11] V. K. Chauhan, et al., "Problem formulations and solvers in linear SVM: a review" *Artif. Intell. Rev.*, vol.52, no. 2, pp. 803-855, Aug. 2019.
- [12] J. Liang, et al., "Efficient and secure decision tree classification for cloud-assisted online diagnosis services," *IEEE Trans. Dependable Secure Comput.*, vol.18, no. 4, pp. 1632-1644, Jun. 2019.
- [13] N. Alhammad and H. Al-Dossari, "Dynamic Segmentation for Physical Activity Recognition Using a Single Wearable Sensor," *Appl. Sci.*, vol.11, no. 6, pp. 2633, Mar. 2021.
- [14] D. Chen, et al., "Hybrid fuzzy c-means CPD-based segmentation for improving sensor-based multiresident activity recognition" *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11193-11207, Jul. 2021.
- [15] T. Stiefmeier, et al., "Fusion of string-matched templates for continuous activity recognition," in *proc. 11<sup>th</sup>. IEEE Int. Sym. Wrbl. Co.*, Boston, 2007, pp. 41-44
- [16] J. R. Kwapisz, et al. "Activity recognition using cell phone accelerometers," *SIGKDD Explor.*, vol. 12, no. 2, pp. 74-82, Dec. 2010.
- [17] K. Forster, et al., "Unsupervised classifier self-calibration through repeated context occurrences: Is there robustness against sensor displacement to gain?," in *proc. Int. Sym. Wrbl. Co.*, Linz, 2009, pp. 77-84
- [18] M. Shoaib, et al., "Towards physical activity recognition using smartphone sensors," in *proc. 10<sup>th</sup>. international conference on ubiquitous intelligence and computing and 10<sup>th</sup>. international conference on autonomic and trusted computing*, Vietri sul Mare, 2013, pp. 80-87
- [19] N. Dua, et al. "Multi-input CNN-GRU based human activity recognition using wearable sensors," *Computing.*, vol. 103, no. 7, pp. 1461-1478, Mar. 2021.
- [20] H. Qian, et al. "Weakly-supervised sensor-based activity segmentation and recognition via learning from distributions," *Artif. Intell.*, vol. 292, pp. 103429, Mar. 2021.
- [21] P. Fryzlewicz, "Wild binary segmentation for multiple change-point detection," *Ann. Stat.*, vol. 42, no. 6, pp. 2243-2281, Nov. 2014.

- 
- [22] Y. Guédon, "Exploring the latent segmentation space for the assessment of multiple change-point models," *Comput. Stat.*, vol. 28, no. 6, pp. 2641-2678, Dec. 2013.
- [23] Z. Harchaoui and O. Cappé, "Retrospective mutiple change-point estimation with kernels," in *proc. IEEE/SP 14<sup>th</sup>. Workshop on Statistical Signal Processing*, Washington, 2007, pp. 768–772
- [24] R. Killick, et al. "Optimal detection of changepoints with a linear computational cost," *J. Am. Stat. Assoc.*, vol. 107, no. 500, pp. 1590-1598, Dec. 2012.
- [25] O. Banos, et al. "Window size impact in human activity recognition," *Sensors*, vol. 14, no. 4, pp. 6474-6499, Mar. 2014.
- [26] Z. Harchaoui, et al., "Kernel change-point analysis," in *proc. Adv. Neural. Inf. Process. Syst.*, A Coruna, 2019, pp. 1-5
- [27] E. Keogh, et al. "An online algorithm for segmenting time series," in *proc. IEEE International Conference on Data Mining*, San Jose, 2001, pp. 289-296
- [28] D. S. Matteson and N. A. James "A nonparametric approach for multiple change point analysis of multivariate data," *J. Am. Stat. Assoc.*, vol. 109, no. 505, pp. 334-345, Jun. 2013.
- [29] G. Rigaiil, "Pruned dynamic programming for optimal multiple change-point detection," *ArXiv.org*, vol. 1, pp. 1-9, Apr. 2010.
- [30] G. Rigaiil, "A pruned dynamic programming algorithm to recover the best segmentations with 1 to  $K_{\max}$  change-points," *J. Soc. fr. stat.*, vol. 156, no. 4, pp. 180-205, Aug. 2015.
- [31] W. Zhang, et al. "Laparotomy cryoablation in rabbit VX2 pancreatic carcinoma," *Pancreas.*, vol. 46, no. 3, pp. 288-295, Mar. 2017.
- [32] Lu. Yang, et al. "Adaptive Chunk-Based Dynamic Weighted Majority for Imbalanced Data Streams With Concept Drift," *IEEE. Trans. Neural. Netw. Learn. Syst.*, vol. 31, no. 8, pp. 2764 - 2778, Aug. 2020.

- [33] L. Zeng, et al. "Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm," *Knowl. Based. Syst.*, vol. 195, no. 11, pp. 105694, May. 2020.
- [34] Lu. Yang, et al., "Dynamic Weighted Majority for Incremental Learning of Imbalanced Data Streams with Concept Drift," in *proc. 26<sup>th</sup>. International Joint Conference on Artificial Intelligence*, Melbourne, 2017, pp. 2393–2399
- [35] Lu. Yang, et al. "Kappa Updated Ensemble for Drifting Data Stream Mining," *Machine Language*, vol. 195, no. 1, pp. 175-218, Jan. 2020.
- [36] S. Ishii and D. Ljunggren, " A Comparative Analysis of Robustness to Noise in Machine Learning Classifiers," B.S. thesis, School. Elect. Eng. Comp. Sci., KTH Royal Inst. Tech., Stockholm, SW, 2021.



## واژه‌نامه‌ی فارسی به انگلیسی

Robustness	استحکام.....
Deoxyribonucleic acid (DNA)	اسیددی‌اکسی‌ریبونوکلیک.....
Ten-fold cross validation	اعتبارسنجی متقابل ده برابری.....
Incremental	افزایشی.....
Weighted majority	اکثریت وزنی.....
Palm alphabet	الفبای کف دست.....
Activity transition	انتقال فعالیت.....
Standard deviation	انحراف استاندارد.....
Chunk-size	اندازه‌ی قطعه.....
Shoulder abduction	انقباض شانه.....
Thresholding	آستانه‌گذاری.....
Spinal cord injuries (SCI)	آسیب‌های نخاعی.....
Histogram statistics	آمار هیستوگرام.....
Dynamic programming (DP)	برنامه‌نویسی پویا.....
Magnitude	بزرگی.....
Maximum	بیشینه.....
Overfit	بیش‌برازش.....

Entropy	بی‌نظمی
Sliding window	پنجره‌ی کشویی
Radial basis function (RBF)	تابع پایه‌ی شعاعی
Optical character recognition (OCR)	تشخیص حروف نوری
Activity recognition (AR)	تشخیص فعالیت
Human activity recognition (HAR)	تشخیص فعالیت‌های انسانی
Change point detection (CPD)	تشخیص نقاط تغییر
Drill	تمرینی
Random forest	جنگل تصادفی
External rotation (ER)	چرخش خارجی
Internal rotation (IR)	چرخش داخلی
Multi-resident	چند ساکن
Single nucleotide polymorphism (SNP)	چندشکلی تک‌نوکلئوتیدی
Wearable sensors	حسگرهای پوشیدنی
Motion sensors	حسگرهای حرکت
Door sensors	حسگرهای در
Temperature sensors	حسگرهای دما
Object sensors	حسگرهای شیء
Ambient sensors	حسگرهای محیطی

---

---

Switch	حسگر سوئیچ
Localization	حسگر محلی سازی
Elbow flexion (EF)	خم شدن آرنج
Shoulder flexion	خم شدن شانه
Clustering	خوشه‌بندی
K-means	خوشه‌بندی سخت سنتی
Fuzzy c-means	خوشه‌بندی فازی
Raw data	داده‌ی خام
Accuracy	دقت
Precision	دقت، درستی
Brain-computer interface (BCI)	رابط مغز و رایانه
Incremental drift	رانش افزایشی
Gradual drift	رانش تدریجی
Virtual drift	رانش مجازی
Concept drift	رانش مفهوم
Recurring drift	رانش مکرر
Abrupt drift (sudden drift)	رانش ناگهانی
Real drift	رانش واقعی
Ambient assisted living (AAL)	زندگی به کمک محیط

---



---

Gyroscope	ژیروسکوپ
Global positioning system (GPS)	سامانه‌ی موقعیت‌یاب جهانی
Recurrent neural network (RNN)	شبکه‌ی عصبی بازگشتی
Convolutional neural network (CNN)	شبکه‌ی عصبی کانولوشن
Accelerometers (acceleration sensors)	شتاب‌سنج (حسگرهای شتاب)
Radio frequency identification (RFID)	شناسایی فرکانس رادیویی
Smoothe	صاف کردن
Convolve	ضرب ستاره‌ای
Binary classifier	طبقه‌بند دودویی
Multi classifier	طبقه‌بند چند کلاسی
Classification	طبقه‌بندی
Data stream classification	طبقه‌بندی جریان داده‌ها
K-nearest neighbors (KNN)	طبقه‌بندی نزدیک‌ترین همسایگان
Non- stationary	غیر ایستا
Moving average filter (MAF)	فیلتر میانگین متحرک
Chunk	قطعه
Segmentation	قطعه‌بندی
Dynamic segmentation	قطعه‌بندی پویا
Binary segmentation (BS)	قطعه‌بندی دودویی

---



---

Wild binary segmentation (WBS)	قطعه‌بندی دودویی وحشی
Peak	قله
Minority class	کلاس اقلیت
Majority class	کلاس اکثریت
Underfit	کم‌برازش
Minimum	کمینه
Ensemble	گروهی
Extension	گسترش
Elbow extension (EE)	گسترش آرنج
Missing	گم‌شده
Support measure machine (SMM)	ماشین اندازه‌گیری پشتیبان
Support vector machine (SVM)	ماشین بردار پشتیبان
False positive (FP)	مثبت کاذب
True positive (TP)	مثبت واقعی
Dataset	مجموعه داده
Centroid	مرکز خوشه
Benchmark	معیار
Magnetometer	مغناطیس‌سنج
Erroneous	مقادیر اشتباه

---



---

Receiver operating characteristic curve (ROC curve)	منحنی مشخصه‌ی عملکرد گیرنده.....
False negative (FN)	منفی کاذب.....
True negative (TN)	منفی واقعی.....
Mean	میانگین.....
Area under the ROC curve (AUC)	ناحیه‌ی زیر منحنی.....
Imbalance ratio	نرخ عدم توازن.....
False positive rate (FP rate)	نرخ مثبت کاذب.....
True positive rate (TP rate)	نرخ مثبت واقعی.....
False negative rate (FN rate)	نرخ منفی کاذب.....
True negative rate (TN rate)	نرخ منفی واقعی.....
Weakly-supervised	نظارت ضعیف.....
Noise	نویز.....
Class noise	نویز کلاس.....
Gaussian noise	نویز گاوسی.....
Attribute noise	نویز ویژگی.....
Inertial measurement units (IMU)	واحدهای اندازه‌گیری اینرسی.....
Ground-truth	واقعی.....
Dynamic weighting	وزن‌دهی پویا.....
Random fourier feature (RFF)	ویژگی فوریه‌ی تصادفی.....

Pearson Product-moment correlation (PMC) ..... همبستگی محصول-لحظه‌ی پیرسون

Machine learning (ML) ..... یادگیری ماشین

## واژه‌نامه‌ی انگلیسی به فارسی

Abrupt drift (sudden drift).....	رانش ناگهانی
Accelerometers (acceleration sensors).....	شتاب‌سنج (حسگرهای شتاب)
Accuracy.....	دقت
Activity recognition (AR).....	تشخیص فعالیت
Activity transition.....	انتقال فعالیت
Ambient assisted living (AAL) .....	زندگی با کمک محیط
Ambient sensors .....	حسگرهای محیطی
Attribute noise.....	نویز ویژگی
Area under the ROC curve (AUC).....	ناحیه‌ی زیر منحنی
Benchmark.....	معیار
Binary classifier .....	طبقه‌بند دودویی
Binary segmentation (BS) .....	قطعه‌بندی دودویی
Brain-computer interface (BCI) .....	رابط مغز و رایانه
Centroid.....	مرکز خوشه
Change point detection (CPD) .....	تشخیص نقاط تغییر
Chunk.....	قطعه
Chunk-size.....	اندازه‌ی قطعه



Class noise.....	نویز کلاس
Classification.....	طبقه‌بندی
Clustering.....	خوشه‌بندی
Concept drift.....	رانش مفهوم
Convolutional neural network (CNN).....	شبکه‌ی عصبی کانولوشن
Convolve.....	ضرب ستاره‌ای
Dataset.....	مجموعه داده
Data stream classification... ..	طبقه‌بندی جریان داده‌ها
Deoxyribonucleic acid (DNA)...	اسیددی اکسی‌ریبونوکلئیک
Door sensors.....	حسگرهای در
Drill.....	تمرینی
Dynamic programming (DP).....	برنامه‌نویسی پویا
Dynamic segmentation.....	قطعه‌بندی پویا
Dynamic weighting.....	وزن‌دهی پویا
Elbow extension (EE).....	گسترش آرنج
Elbow flexion (EF).....	خم شدن آرنج
Erroneous .....	مقادیر اشتباه
Ensemble.....	گروهی
Entropy.....	بی‌نظمی

---



---

External rotation (ER).....	چرخش خارجی
Extension.....	گسترش
False negative rate (FN rate) .....	نرخ منفی کاذب
False negative (FN) .....	منفی کاذب
False positive rate (FP rate) .....	نرخ مثبت کاذب
False positive (FP) .....	مثبت کاذب
Fuzzy c-means.....	خوشه‌بندی فازی
Gaussian noise.....	نویز گاوسی
Global positioning system (GPS).....	سامانه‌ی موقعیت‌یاب جهانی
Gradual drift.....	رانش تدریجی
Ground-truth.....	واقعی
Gyroscope.....	ژیروسکوپ
Histogram statistics.....	آمار هیستوگرام
Human activity recognition (HAR) .....	تشخیص فعالیت‌های انسانی
Imbalance ratio.....	نرخ عدم‌توازن
Incremental.....	افزایشی
Incremental drift.....	رانش افزایشی
Inertial measurement units (IMU).....	واحدهای اندازه‌گیری اینرسی
Internal rotation (IR) .....	چرخش داخلی

K-means .....	خوشه‌بندی سخت سنتی
K-nearest neighbors (KNN).....	طبقه‌بندی نزدیک‌ترین همسایگان
Localization.....	حسگر محلی سازی
Machine learning (ML) .....	یادگیری ماشین
Magnetometer.....	مغناطیس‌سنج
Magnitude.....	بزرگی
Majority class.....	کلاس اکثریت
Maximum.....	بیشینه
Mean.....	میانگین
Minimum.....	کمینه
Minority class.....	کلاس اقلیت
Missing.....	گم‌شده
Moving average filter (MAF) .....	فیلتر میانگین متحرک
Multi-resident.....	چند ساکن
Noise.....	نویز
Motion sensors .....	حسگرهای حرکت
Multi classifier .....	طبقه‌بند چند کلاسی
Non-stationary .....	غیر ایستا
Object sensors.....	حسگرهای شیء

---



---

Optical character recognition (OCR).....	تشخیص حروف نوری.....
Overfit.....	بیش‌برازش.....
Palm alphabet.....	الفبای کف دست.....
Peak.....	قله.....
Pearson product-moment correlation (PMC) .....	همبستگی محصول-لحظه‌ی پیرسون.....
Precision.....	دقت، درستی.....
Radial basis function (RBF) .....	تابع پایه‌ی شعاعی.....
Random forest.....	جنگل تصادفی.....
Random fourier feature (RFF) .....	ویژگی فوریه‌ی تصادفی.....
Raw data.....	داده‌ی خام.....
Real drift.....	رانش واقعی.....
Receiver operating characteristic curve (ROC curve)....	منحنی مشخصه‌ی عملکرد گیرنده .....
Recurrent neural network (RNN).....	شبکه‌ی عصبی بازگشتی.....
Recurring drift.....	رانش مکرر.....
Radio frequency identification (RFID).....	شناسایی فرکانس رادیویی.....
Robustness.....	استحکام.....
Segmentation.....	قطعه‌بندی.....
Shoulder abduction.....	انقباض شانه.....
Shoulder flexion.....	خم شدن شانه.....

---



---

Single channel.....	تک کانال
Single frame.....	تک فریم
Single nucleotide polymorphism (SNP).....	چندشکلی تک‌نوکلئوتیدی
Single segmentation.....	قطعه‌بندی منفرد
Sliding window.....	پنجره‌ی کشویی
Smoothe.....	صاف کردن
Spinal cord injuries (SCI) .....	آسیب‌های نخاعی
Standard deviation .....	انحراف استاندارد
Support measure machine (SMM) .....	ماشین اندازه‌گیری پشتیبان
Support vector machine (SVM) .....	ماشین بردار پشتیبان
Switch.....	حسگر سوئیچ
Temperature sensors.....	حسگرهای دما
Ten-fold cross validation.....	اعتبارسنجی متقابل ده برابری
Thresholding.....	آستانه‌گذاری
True negative rate (TN rate) .....	نرخ منفی واقعی
True negative (TN) .....	منفی واقعی
True positive rate (TP rate) .....	نرخ مثبت واقعی
True positive (TP) .....	مثبت واقعی
Underfit.....	کم‌برازش

Valley.....	دره
Virtual drift.....	رائش مجازی
Weakly-supervised.....	نظارت ضعیف
Wearable sensors.....	حسگرهای پوشیدنی
Weighted majority.....	اکثریت وزنی
Wild binary segmentation (WBS) .....	قطعه‌بندی دودویی وحشی

## Abstract

In recent years, data stream segmentation has become a very important topic in machine learning science, as more and more data can be treated as data streams. Most of this data is characterized as high speed, non-stationary data distribution and unlimited length, examples of which can be seen in activity detection, credit card transactions and telecommunications. Concept drift occurs when the distribution of data streams changes over time. Concept drift requires the adjustment of classifiers to adapt to new conditions. The type of changes can be classified as abrupt, gradual, incremental and recurring concept drift. Real-world applications are always a combination of different types of concept drifts. In this thesis, by using 4 commonly used methods in the chunk-based classification of data streams (ACDWM, DWMIL, DUE and KUE), a relationship is made in order to compare the robustness of the algorithms with each other against noise and also considering the imbalance ratio and concept drift is presented. The mentioned algorithms were examined and tested with 4 datasets. By examining the graphs obtained from entering different amounts of noise, imbalance ratio and concept drift to the data, an efficient relationship has been obtained.

**Keywords:** data stream segmentation, activity detection, noise, imbalance ratio, concept drift



**Amirkabir University of Technology**  
**(Tehran Polytechnic)**

**Computer Engineering Department**  
**MSc Thesis**

**robustness in algorithms using segmentation for  
activity recognition**

**By**  
**Ali Amani**

**Supervisor**  
**Dr. Mohammadreza Razzazi**

**Shahrivar 1401**