مجله ژئوفیزیک ایران، جلد ۲۰، شماره ۱، ۱۴۰۵، صفحه ۷۵-۹۵ (مقاله پژوهشی) DOI: 10.30499/ijg.2025.524794.1699

تخمین موقعیت منبع لرزه‌زایی القایی با استفاده از داده‌های ثبت‌شده در ایستگاه‌های لرزه‌نگاری بر پایه هوش مصنوعی

هومن پارسا[۱] و محمد یاسر رادان[۲*]

[۱] دانش‌آموخته کارشناسی‌ارشد گروه مهندسی عمران، دانشکده فنی و مهندسی، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران
[۲] استادیار، مجتمع دانشگاهی پدافندغیرعامل، دانشگاه صنعتی مالک اشتر، تهران، ایران

(دریافت: ۱۴۰۴/۰۲/۳۱، پذیرش: ۱۴۰۴/۰۵/۰۶)

چکیده

لرزه‌های القایی ناشی از فعالیت‌های انسانی مانند استخراج و تزریق سیالات زیرسطحی می‌توانند یکپارچگی زیرساخت‌های حیاتی را به مخاطره اندازند. این پژوهش می‌تواند بدون نیاز به داده‌های زمین‌شناسی جامع و شبکه لرزه‌نگاری کامل موقعیت منطقه‌ای رومرکز چشمه لرزه‌ای را در چارچوبی چندمرحله‌ای تخمین زند. ابتدا، سیگنال‌های لرزه‌ای با بهره‌گیری از نسبت میانگین سیگنال در پنجره کوتاه‌دوره به میانگین سیگنال در پنجره بلنددوره برای اصلاح جهش‌های ناگهانی خط‌مبنا پیش‌پردازش شد. در ادامه، همبستگی متقابل برای شکل موج‌ها در چهار تاخیر زمانی ۰/۵، ۰/۱، ۰/۰۵ و ۰/۰۱ ثانیه محاسبه و برای استخراج زیرمجموعه بهینه ویژگی‌ها، روش حذف بازگشتی ویژگی به‌کار گرفته شد. ویژگی حاصل به مدل طبقه‌بندی ترکیبی بر پایه میانگین‌گیری احتمال‌ها ارائه شد که ماشین بردار پشتیبان و الگوریتم افزایش گرادیان فوق‌العاده در آن تلفیق شده‌اند. برای بررسی تأثیر نامتوازن بودن داده‌ها، فرایند آموزش مدل بدون و با نمونه‌برداری افزایشی مصنوعی انجام شد. در حالت بدون نمونه‌برداری ، با کاهش گام زمانی به ۰/۱ ثانیه، دقت طبقه‌بندی مدل در تعیین منطقه‌ی رومرکز منبع لرزه‌ای از ۰/۷۳ به ۰/۹۰ افزایش می‌یابد و نتایج پایدارتری را نشان می‌دهد. در مقیاس‌های کمتر، حساسیت مدل به نوسانات نویزی موجب اشباع دقت عملکرد و افزایش اندکی در انحراف‌معیار گردید؛ از این رو، گام ۰/۱ ثانیه به‌عنوان تعادلی مطلوب معرفی شد. با اعمال نمونه‌برداری ، پایداری تخمین‌ها به‌طور کلی بهبود یافته و انحراف‌معیار به‌نحو قابل‌توجهی کاهش یافته است؛ اما در گام‌های بزرگ‌تر از ۰/۰۱ افت جزئی دقت مشاهده شد که به کیفیت نمونه‌های مصنوعی نسبت داده می‌شود. بیشترین بهره‌وری با ترکیب گام ۰/۰۱ ثانیه و نمونه‌برداری حاصل گردید؛ دقت طبقه‌بندی مدل افزایش ۵/۷ درصدی ارتقا یافته است. نتایج نشان می‌دهد که چارچوب پیشنهادی دقت و پایداری خود را حفظ کرده و برای استقرار عملیاتی مناسب است؛ اما در شرایط محدودیت محاسباتی، گام ۰/۱ ثانیه بدون نمونه‌برداری همچنان بهینه‌ترین سازش میان هزینه و دقت طبقه‌بندی را فراهم می‌کند.

کلمه‌های کلیدی: تخمین موقعیت منبع لرزه‌ای، همبستگی متقابل، الگوریتم گرادیان تقویتی، ماشین بردار پشتیبان

* نگارنده رابط: radan@mut.ac.ir

۱ مقدمه

زلزله‌های القایی ناشی از عملیات صنعتی مانند تزریق سیالات، حفاری عمیق یا استخراج منابع زیرسطحی می‌توانند خسارات‌هایی به زیرساخت‌های حساس و حیاتی وارد سازند. هرچند نسبت زلزله‌های القایی به زلزله‌های کاملا طبیعی بسیار کوچک است، اما همین دسته از زلزله‌ها نیز می‌تواند مخرب بوده و اهمیت زیادی در مطالعات مهندسی دارد. به‌منظور واکنش بهینه و پیشگیری از گسترش آسیب، تعیین دقیق محل وقوع این زلزله‌ها از اهمیت ویژه‌ای برخوردار است (السورث، ۲۰۱۳؛ فولجر و همکاران، ۲۰۱۸).

یکی از اهداف اساسی در لرزه‌شناسی، تعیین دقیق مکان منبع رویداد لرزه‌ای در زمین است که به اصطلاح تعیین مکان رویداد نامیده می‌شود. موقعیت‌یابی دقیق رخدادهای لرزه‌ای نقش بسیار مهمی در درک فرایندهای زمین‌شناختی و نیز برآورد خطرات لرزه‌ای دارد. برای این منظور از داده‌های لرزه‌ای ثبت‌شده توسط لرزه‌نگارها استفاده می‌شود که حاوی اطلاعاتی درباره زمان رسید و شکل موج‌های لرزه‌ای در ایستگاه‌های مختلف است. تحلیل این داده‌ها به روش‌های سنتی مبتنی بر محاسبه زمان‌های رسید امواج (P و S) و استفاده از مدل‌های سرعتی زمین انجام می‌گیرد. اما با گسترش شبکه‌های لرزه‌نگاری و افزایش حجم داده‌ها، روش‌های کلاسیک با چالش‌هایی نظیر نیاز به صرف زمان زیاد برای پردازش دستی و عدم توانایی در شناسایی رویدادهای کوچک دارای نویز مواجه شده‌اند (کنگ و همکاران، ۲۰۱۹).

در سال‌های اخیر پیشرفت در الگوریتم‌های هوش مصنوعی و به ویژه یادگیری ماشین فرصت‌هایی تازه‌ای برای پردازش خودکار داده‌های لرزه‌ای فراهم کرده است. یادگیری ماشینی قادر است الگوهای پیچیده و غیرخطی موجود در داده‌ها را که شاید از دید روش‌های سنتی پنهان بمانند، آشکار سازد. مطالعات متعددی نشان داده‌اند که این مدل‌های داده‌محور می‌توانند سیگنال‌های زمین‌لرزه را حتی در سیگنال با نویز بالا را با دقت بالایی تشخیص دهند. دقت بالای این ابزارهای مبتنی بر یادگیری ماشینی در تشخیص فازها و رویدادها، منجر به افزایش تعداد زمین‌لرزه‌های قابل‌شناسایی و بهبود کیفیت داده‌های ورودی برای فرآیند موقعیت‌یابی شده است (راس و همکاران، ۲۰۱۸؛ ژو و بروزا، ۲۰۱۹؛ موسوی و بروزا، ۲۰۲۲).

با توجه به محدودیت‌های کمبود ایستگاه‌ها و گسست شبکه‌های لرزه‌نگاری در برخی مناطق، ترکیب داده‌محوری یادگیری ماشینی و روش‌های کلاسیک می‌تواند بهره‌وری از داده‌های ناقص را به حداکثر برساند. در این چارچوب، ویژگی‌هایی موج و الگوهای رفتاری زمان-فرکانس استخراج شده و توسط مدل‌های یادگیری ماشین برای تخمین مکان چشمه پردازش می‌گردند. نتایج پژوهش‌ها نشان می‌دهد حتی با یک ایستگاه لرزه‌ای نیز می‌توان مکان منبع رویداد را در محدوده‌های چند ده متری تخمین زد. کاربردهای عملی این رویکرد شامل نظارت مرزی برای شناسایی تونل‌زنی‌های پنهانی و حفاظت از تاسیسات صنعتی حساس می‌باشد. همچنین در معادن زیرزمینی، سامانه‌های هوشمند لرزه‌ای می‌توانند وقوع ریزش‌ها ناخواسته را با دقت بالا آشکار کرده و هشدارهای لازم را صادر کنند. این سامانه‌ها افزون بر افزایش ایمنی، باعث بهینه‌سازی عملیات تعمیر و نگهداری و کاهش هزینه‌های ناشی از توقف‌های اضطراری نیز می‌گردند (جکمپودی و همکاران، ۲۰۲۰؛ بیلال و همکاران، ۲۰۲۲). موقعیت‌یابی خودکار و دقیق زمین‌لرزه‌ها می‌تواند به سیستم‌های هشدار سریع و مدیریت بحران نیروی تازه‌ای ببخشد و در کاهش مخاطرات و آسیب‌های ناشی از زمین‌لرزه موثر واقع شود. به بیان دیگر، تلفیق داده‌های لرزه‌ای گسترده با توان یادگیری ماشینی، گامی نوین و پراهمیت در جهت تقویت توان علمی ما در شناسایی و موقعیت‌یابی لرزه‌های زمین محسوب می‌شود (سامادی و همکاران، ۲۰۲۲).

شناسایی محل دقیق تحریک لرزه‌ای در محیط‌های ژئوتکنیکی، نقش کلیدی در تحلیل رفتار زمین، ارزیابی پایداری سازه‌ها و توسعه سامانه‌های پایش زیرسطحی ایفا می‌کند. با این حال، روش‌های متداول در موقعیت‌یابی لرزه‌ای عموما متکی بر دسترسی به شبکه‌های لرزه‌نگاری گسترده، اطلاعات زمین‌شناسی دقیق، و مدل‌های سرعتی معتبر برای امواج لرزه‌ای هستند. این پیش‌نیازها باعث می‌شوند چنین روش‌هایی در محیط‌های دورافتاده، پروژه‌های مقیاس کوچک، یا مناطقی با زیرساخت محدود، غیرقابل اجرا یا بسیار پرهزینه باشند (موسوی و بروزا، ۲۰۲۲). الگوریتم‌های مبتنی بر هوش مصنوعی و یادگیری ماشین این قابلیت را دارند که الگوهای پنهان و پیچیده موجود در داده‌های شکل موج را شناسایی کرده و به صورت سریع، خودکار و دقیق به تخمین موقعیت منبع بپردازند. در مقایسه با رویکردهای کلاسیک، این روش‌ها نه تنها به منابع داده‌ای و سخت‌افزاری کمتری نیاز دارند، بلکه هزینه‌های عملیاتی را نیز به‌شدت کاهش داده و امکان پیاده‌سازی در مقیاس‌های وسیع‌تر را فراهم می‌کنند (پرول و همکاران، ۲۰۱۸).

در پژوهشی که ژانگ و همکاران (۲۰۲۰) انجام دادند، شبکه‌ای عصبی تمام‌کانولوشنالی عمیق با استفاده از داده‌های حاصل از شبکه‌ای متشکل از ۳۰ ایستگاه لرزه‌ای آموزش داده شد تا محل زمین‌لرزه‌های القاشده در اکلاهما به‌طور مستقیم برآورد شود. این مدل یادگیری عمیق بدون نیاز به برداشت دستی فازهای لرزه‌ای ، نگاشت مستقیمی از شکل موج‌های ثبت‌شده به مختصات منبع را فرا گرفت. نتایج نشان داد که روش مبتنی بر شبکه عصبی می‌تواند رویدادهای لرزه‌ای را با میانگین خطایی در حد چند کیلومتر مکان‌یابی کند و فرآیند تعیین محل را نسبت به روش‌های سنتی به‌مراتب سریع‌تر نماید. همچنین یک معماری یادگیری عمیق موسوم به EqConvMixer، با استفاده از سیگنال‌های تک‌ایستگاهی مکان زمین‌لرزه را تخمین می‌زند. این مدل با استخراج ویژگی از شکل موج‌ها، مختصات کانون را مستقیما پیش‌بینی می‌کند و دقتی مشابه روش‌های چندایستگاهی دارد. این نشان می‌دهد که با شبکه‌های عصبی پیشرفته، می‌توان از داده‌های تک‌ایستگاهی برای مکان‌یابی قابل اطمینان زمین‌لرزه‌ها استفاده کرد (السید و همکاران، ۲۰۲۳)

در یک مطالعه کاربردی مرتبط، تکنیک‌های یادگیری ماشین برای مکان‌یابی مستقیم ریزلرزه‌ها در محیط سه‌بعدی به کار گرفته شده است. یک مدل نظارت‌شده مانند الگوریتم جنگل تصادفی بر روی داده‌های رویدادهای شناخته‌شده آموزش داده شد تا مختصات مکانی منابع لرزه‌ای جدید را پیش‌بینی کند، بی‌آنکه نیاز به مدل سرعتی دقیق در زیرسطح باشد. این روش بر روی داده‌های لرزه‌های القاشده آزمایش و مشاهده شد که حتی در محیط‌های پیچیده زیرزمینی نیز قادر به تعیین نسبتا دقیق محل رویدادهای لرزه‌ای است و یافته‌های آن نشان می‌دهد که رویکرد مبتنی بر هوش مصنوعی می‌تواند پایش ریزلرزه‌ها را به‌صورت بلادرنگ ممکن ساخته و مکان منبع آن‌ها را با دقت و پایداری مناسبی برآورد کند (چن و همکاران، ۲۰۲۲).

در جدیدترین پیشرفت این حوزه، سامانه‌های مبتنی بر یادگیری عمیق به‌نام SourceNet توسعه یافته است که امکان تعیین بسیار سریع پارامترهای منبع زمین‌لرزه را فراهم می‌کند. این شبکه عصبی با استفاده از داده‌های لرزه‌ای آموزش دیده تا بلافاصله پس از دریافت اولین موج (P)، مختصات کانون و سایر پارامترهای منبع زمین‌لرزه را تخمین بزند. روش ارائه‌شده بر روی داده‌های ناحیه چانگ‌نینگ چین آزمایش شده و توانسته است مشخصات منبع رویدادهای لرزه‌ای را با دقت مناسب و ظرف زمانی کمتر از ۰/۱ ثانیه استخراج نماید. نتایج این پژوهش نشان می‌دهد که به کارگیری الگوریتم‌های هوش مصنوعی می‌تواند موجب بهبود چشمگیر سرعت و خودکارسازی فرآیند

مکان‌یابی و تعیین ویژگی‌های زمین‌لرزه‌ها شود (زو و همکاران، ۲۰۲۵).

در پژوهش کاربردی اخیر لیونگ و ژو (۲۰۲۴)، به‌منظور تخمین مکان منبع رویداد لرزه‌ای، ابتدا میزان شباهت امواج لرزه‌ای با بهره‌گیری از تکنیک همبستگی متقابل محاسبه گردید. سپس ویژگی‌های استخراج‌شده به‌صورت تأخیرهای زمانی حاصل از این همبستگی، به‌عنوان ورودی به یک مدل یادگیری ماشین از نوع شبکه‌ی عصبی پرسپترون چندلایه قرار گرفت. این مدل با داده‌های شبیه‌سازی‌شده بر پایه‌ی یک مدل سرعت سه‌بعدی آموزش داده شد و بر اساس ورودی‌های مذکور، مختصات منبع ریزلرزه‌ها با دقت بالا تعیین گردید. کاربست این روش بر روی داده‌های واقعی یک سامانه‌ی زمین‌گرمایی تحت پایش، به مکان‌یابی دقیق رویدادهای لرزه‌ای انجامید.

در این پژوهش، با رویکردی داده‌محور و مبتنی بر الگوریتم‌های هوش مصنوعی، چارچوبی برای تخمین موقعیت منطقه‌ای رومرکز چشمه لرزه‌ای در محیط‌های ژئوتکنیکی پیشنهاد می‌شود که علاوه بر دقت بالا، از وابستگی به شبکه‌های لرزه‌نگاری گسترده و داده‌های زمین‌شناسی تفصیلی می‌کاهد. این چارچوب با بهره‌گیری از سیگنال‌های سنسوری محدود و تحلیل مرحله‌ای شباهت موج، سعی دارد در شرایط عملیاتی با دسترسی محدود و در حضور نویز، موقعیت‌یابی قابل‌اطمینانی ارائه دهد. اصلاح جابه‌جایی ناگهانی خط مبنای شکل موج به‌منظور یکنواخت‌سازی سیگنال‌های لرزه‌ای، محاسبه و تشکیل ماتریس شباهت موج در چهار گام زمانی مجزا (با گام‌های زمانی متفاوت برای جابه‌جایی پنجره‌های شکل موج) بر پایه معیار همبستگی متقابل و تخمین موقعیت منبع لرزه‌ای با استفاده از مدل طبقه‌بندی تلفیقی مبتنی بر Soft Voting از مهم‌ترین اهداف پژوهش حاضر به شمار می‌روند.

۲ روش پژوهش

۲-۱ جمع‌آوری و پیش‌پردازش داده‌ها

میدان زمین‌گرمایی چشمه‌های آب‌گرم برادیز در قسمت شمالی ایالت نوادا در غرب ایالات متحده آمریکا، در حوزه حوضه و رشته‌کوه بیسین و رنج قرار دارد. این میدان در حدود ۴۵ کیلومتری شمال‌شرق شهر رینو و میان رشته‌کوه تراکی و چشمه‌های شمالی واقع شده است. این گستره به یک گسل عادی کرانه‌کوهی وابسته است که مسیرهای اصلی جریان سیالات هیدروترمال را فراهم می‌آورد. شبکه لرزه‌نگاری آزمایشگاه برکلی شامل ۱۶ لرزه‌سنج سه‌محوره در جنوب و جنوب‌غرب میدان مستقر شده است (ماتزل و همکاران، ۲۰۱۷؛ رینیش و همکاران، ۲۰۱۸).

شکل ۱. موقعیت جغرافیایی ایستگاه‌ها (نقاط آبی) و رخدادهای القایی (نقاط قرمز) میدان زمین‌گرمایی چشمه‌های آب‌گرم برادیز.

مطالعات نشان می‌دهد که فعالیت استخراج و تزریق سیالات در برادیز مستقیما با وقوع رویدادهای لرزه‌ای کوچک (میکروزلزله‌ها) مرتبط است. به گونه‌ای که دوره‌های پمپاژ مداوم سیال (حجم تولید بلندمدت بالا) با کمبود زلزله ثبت‌شده همراه بوده، در حالی که توقف موقتی پمپاژ با افزایش وقوع میکروزلزله‌ها هم‌زمان است. این نتایج نشان می‌دهد که پمپاژ بلندمدت باعث افزایش تنش مؤثر در سازند و مهار گسل لغزشی می‌شود، اما در زمان توقف پمپاژ تنش مؤثر کاهش یافته و احتمال گسلش لغزشی افزایش می‌یابد. به عبارت دیگر، تغییرات فشار هیدرولیکی سیال در گسل‌های برادیز می‌تواند به‌طور مستقیم زلزله‌های القایی را ایجاد کند (ماتزل و همکاران، ۲۰۱۷؛ رینیش و همکاران، ۲۰۱۸).

در این پژوهش، داده‌های لرزه‌ای القایی میدان زمین‌گرمایی چشمه‌های آب‌گرم برادیز استخراج گردید. این مجموعه شامل ۵۹ رخداد لرزه‌ای القایی با ثبت سه مولفه شتاب (x, y, z) در پنج ایستگاه لرزه‌نگاری است (شکل ۱). داده‌ها در قالب مجموعه آزمایشی میدان چشمه‌های آب‌گرم برادیز ارائه شده است که زیرمجموعه‌ای از اطلاعات موجود در وب‌سایت آزمایشگاه ملی لارنس برکلی برای لرزه‌خیزی القایی به شمار می‌رود.

در جدول ۱، مقادیر حداقل، میانه و حداکثر پارامترهای رومرکز (عرض و طول جغرافیایی)، عمق، بزرگا (Mw؛ مقیاس Moment Magnitude)، گپ آزیموتی و خطای رومرکز (ERH؛ عدم‌قطعیت افقی) را برای ۵۹ رویداد القایی میدان زمین‌گرمایی چشمه‌های آب‌گرم برادیز ارائه شده است.

۲-۲ بستر نرم‌افزاری پردازش داده

کتابخانه (ObsPy) یک بسته متن‌باز به زبان پایتون است که به‌طور ویژه برای خوانش، پردازش و تحلیل داده‌های لرزه‌ای توسعه یافته است. این چارچوب با فراهم‌کردن ساختارهای شیءگرا، امکان یکپارچه‌سازی تمام مراحل زنجیره کاری (از دریافت داده تا تحلیل و نمایش) را در محیطی واحد مهیا می‌کند. در پژوهش حاضر، تمامی مراحل واردسازی داده، پیش‌پردازش و محاسبه شباهت موج با استفاده از این کتابخانه انجام شده است. هر تریس نمایانگر یک رکورد منفرد موج لرزه‌ای همراه با سربرگ شامل زمان مبدأ، نرخ نمونه‌برداری و مشخصات دستگاه است. مجموعه‌ای از تریس‌ها با پوشش زمانی مشترک در یک جریان گروه‌بندی می‌شود تا عملیات دسته‌جمعی مانند برش، هم‌ترازسازی بر آن اعمال گردد. ObsPy توابع متعددی برای فیلترگذاری، تصحیح پاسخ دستگاه، چرخش مؤلفه‌ها، آشکارسازی خودکار رویداد و ترسیم نمودار در اختیار قرار می‌دهد. افزون بر این، از طریق رابط وب‌سرویس می‌توان داده‌ها را مستقیما از مراکز لرزه‌نگاری بین‌المللی دریافت کرد (کریشر و همکاران، ۲۰۱۵).

جدول ۱. مقادیر حداقل، میانه و حداکثر پارامترهای ۵۹ رویداد القایی میدان زمین‌گرمایی چشمه‌های آب‌گرم برادیز.

پارامتر	حداقل	میانه	حداکثر
عرض جغرافیایی (°N)	۳۹/۷۹۱۷	۳۹/۷۹۴۸	۳۹/۸۱۶۴
طول جغرافیایی (°W)	−۱۱۷/۰۰۲۰	−۱۱۸/۹۸۷۳	−۱۱۸/۹۹۴۳
عمق (km)	۰/۱۱۱	۰/۳۱۵	۱/۱۵۳
بزرگا	۰/۲۲	۰/۴۶	۱/۷۳
گپ آزیموتی (°)	۵	۱۴	۲۰
خطای رومرکز (km)	۰/۰۷۸	۰/۱۵۱	۰/۷۳۹

۲-۳ تصحیح خط مبنا

در این پژوهش، یک چارچوب چندمرحله‌ای برای پردازش و تصحیح خط مبنا در داده‌های لرزه‌ای القایی پیشنهاد می‌گردد. هدف اصلی این فرآیند، شناسایی و تصحیح تغییرات ناگهانی خط مبنای سیگنال‌ها است. این جابه‌جایی‌های ناگهانی خط مبنا عمدتا ناشی از عوامل محیطی هستند و در داده‌های این پژوهش ۹٪/۴ از سیگنال‌ها را تحت‌تاثیر قرار داده‌اند. یکی از روش‌های کارآمد برای آشکارسازی تغییرات ناگهانی در سیگنال‌ها، محاسبه نسبت میانگین سیگنال در پنجره کوتاه‌مدت (STA) به میانگین سیگنال در پنجره بلندمدت (LTA) است. برای تعیین بهینه طول پنجره‌های STA و LTA و آستانه نسبت STA/LTA، یک آنالیز حساسیت جامع اجرا شد تا نرخ تشخیص و اخطار کاذب این تغییرات ناگهانی خط مبنا در بهترین سطح قرار گیرد. در این روند، ابتدا داده‌های لرزه‌ای مربوط به هر رویداد استخراج و سپس برای هر ایستگاه لرزه‌نگاری، میانگین مربعی سیگنال در دو پنجره زمانی STA و LTA محاسبه می‌گردد. در نهایت، با محاسبه نسبت STA/LTA می‌توان به‌عنوان معیاری ساده و مؤثر، نقاطی از سیگنال را که دچار تغییرات سریع شده‌اند، به‌طور خودکار شناسایی کرد. در مرحله‌ی تصحیح خط مبنا، با شناسایی نقاط تغییر ناگهانی خط مبنا، هر سیگنال به بخش‌هایی تقسیم می‌شود (هر بخش بین دو نقطه‌ی تغییر ناگهانی قرار دارد و برای پوشش کامل، ابتدا و انتهای سیگنال نیز به‌عنوان نقاط تقسیم در نظر گرفته می‌شوند) و برای هر بخش میانگین خط مبنا محاسبه می‌گردد. سپس، با استفاده از این میانگین، همه بخش‌ها به‌صورت انتقالی بر محور صفر بازمرجع می‌شوند. این کار باعث حذف اثرات تغییرات ناگهانی در خط مبنا خواهد شد.

۲-۴ محاسبه همبستگی متقابل

در این بخش، به‌منظور سنجش همبستگی شکل‌موج‌های ثبت‌شده از رویدادهای مختلف لرزه‌ای، محاسبه جداگانه ماتریس‌های همبستگی متقابل برای هر یک از پنج ایستگاه و سه مولفه (در مجموع ۱۵ ماتریس) انجام شد تا اثرات شرایط محلی حذف شود و شباهت‌ها صرفا بیانگر ویژگی‌های منبع لرزه‌ای باشند.

از آن‌جا که همبستگی متقابل روی کل طول سیگنال محاسبه می‌شود، وجود جابه‌جایی ناگهانی خط مبنا در داده‌ها می‌تواند ضریب همبستگی رویدادهای مرتبط را به‌طور مصنوعی کاهش دهد. اصلاح خط مبنا با حذف این تغییرات ناگهانی، خطای سیستماتیک را برطرف کرده و دقت ضرایب همبستگی را بهبود می‌بخشد.

پیش‌پردازش سیگنال‌ها شامل نرمال‌سازی در بازه ۱- تا ۱ است؛ این نرمال‌سازی اختلاف دامنه‌ها را حذف می‌کند تا محاسبه همبستگی متقابل صرفا بر پایه شباهت شکل‌موج‌ها انجام شود. در گام بعد، برای هر جفت رویداد، سیگنال کوتاه‌تر به‌صورت لغزان روی سیگنال بلندتر قرار گرفت و در هر موقعیت، همبستگی مستقیم با استفاده از کتابخانه ObsPy محاسبه شد. بیشینه ضرایب همبستگی به‌عنوان شاخص شباهت دو رویداد در همان ایستگاه ثبت شده و با انتخاب گام لغزش دقت کافی حفظ و حجم محاسباتی کنترل شد.

برای تضمین بالاترین دقت عددی و اجتناب از تقریب‌های فرکانسی، همبستگی متقابل سیگنال‌ها به‌صورت روش مستقیم برای چهار گام زمانی ۰/۵، ۰/۱، ۰/۰۵ و ۰/۰۱ ثانیه (تماما مضارب صحیح نرخ نمونه‌برداری ۱۰۰ هرتز انتخاب شدند تا در هر گام لغزش هیچ نمونه‌ای از دست نرود) انتخاب شد. این مقادیر پس از انجام تحلیل حساسیت مقدماتی بر روی بازه زمانی ۰/۰۱ تا ۰/۵ ثانیه برگزیده شدند، زیرا گام‌های میانی (مانند ۰/۲ ثانیه) بهبود معنی‌داری در بیشینه ضریب همبستگی ایجاد نکردند اما هزینه محاسباتی را به‌طور چشمگیری افزایش می‌دادند. این انتخاب، پوشش موثر مقیاس‌های زمانی بلندمدت و کوتاه‌مدت را فراهم می‌کند و هم‌زمان بار محاسباتی را در سطح قابل‌قبول حفظ می‌کند.

روش مستقیم مبتنی بر تعریف بنیادی همبستگی است و شامل ضرب نمونه‌ها و سپس جمع زدن نتایج حاصل در حوزه زمان می‌باشد، به طوری که برای دو سیگنال x[n] و y[n] به ترتیب با طول‌های N_x و N_y، تابع همبستگی متقابل کامل به شکل رابطه (۱) تعریف می‌شود.

$$R_{xy}[k] = \sum_{n=0}^{N_x-1} x[n]\ \overline{y[n-k+N_x-1]} \qquad (۱)$$

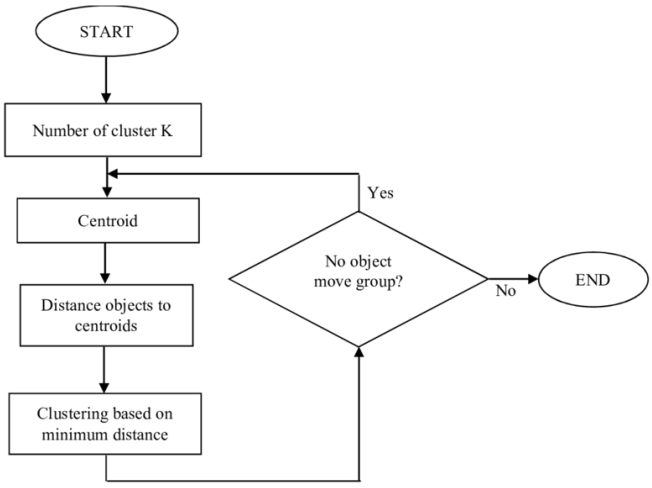
در این روابط R_{xy} [k] ضریب همبستگی متقابل در میزان برد k را نشان می‌دهد.

۲-۵ تفکیک مناطق وقوع زمین‌لرزه با الگوریتم K-میانگین

الگوریتم K-میانگین یکی از روش‌های شناخته‌شده‌ی خوشه‌بندی غیرفعال (یادگیری بدون نظارت) است که هدف آن تقسیم مجموعه‌ای از داده‌های نمونه‌ای به K خوشه‌ی مجزا بر اساس نزدیکی نقاط به مراکز خوشه‌ها است (شکل ۲). در این الگوریتم، داده‌ها به‌گونه‌ای دسته‌بندی می‌شوند که مجموع مربعات فاصله‌ی هر نقطه از مرکز خوشه‌ی مربوطه حداقل شود. به عبارت دیگر، K-میانگین سعی می‌کند تابع هدف رابطه (۲) را کمینه کند.

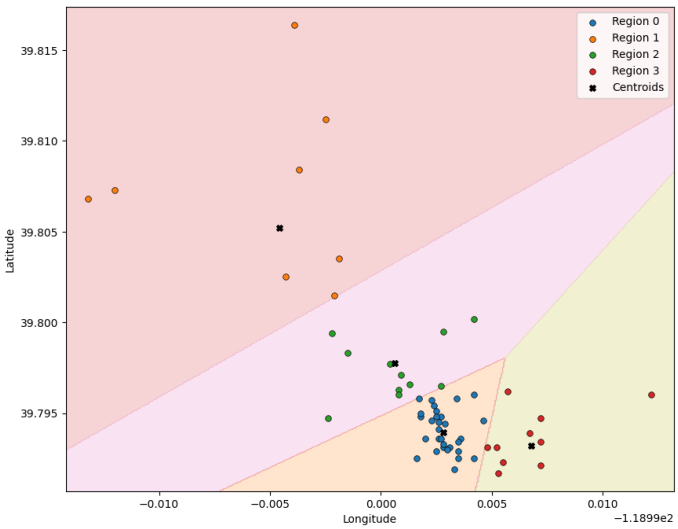
$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \left| x_j - \mu_j \right|^2 \qquad (۲)$$

که در آن C_i نشان‌دهنده خوشه‌ی i ام و μ_i مرکز آن خوشه است. این الگوریتم با ایده‌ی تخصیص مجدد نقاط به نزدیک‌ترین مرکز و به‌روزرسانی مراکز خوشه (میانگین نقاط) طراحی شده است. مراحل اصلی اجرای الگوریتم K-میانگین به‌صورت تکرارشونده انجام می‌شود تا زمانی که شرط همگرایی برآورده شود (جین و هان، ۲۰۱۱).



شکل ۲. فلوچارت اجرای الگوریتم خوشه‌بندی K-میانگین (عبدالرحمن و همکاران، ۲۰۲۱).

به منظور منطقه‌بندی خودکار رویدادهای لرزه‌ای و تهیه برچسب‌های منطقه‌ای رومرکز چشمه لرزه‌ای برای آموزش مدل طبقه‌بندی در این پژوهش، تنها مختصات طول و عرض جغرافیایی هر رویداد به الگوریتم K-میانگین داده شد و خوشه‌ها بر اساس نزدیکی مکانی تعریف گردیدند. در ابتدا تعداد مناطق (خوشه‌ها) مشخص و مراکز اولیه به‌صورت تصادفی تعیین شدند. در ادامه، تخصیص داده‌ها به خوشه‌ها انجام و در هر تکرار، موقعیت هر مرکز با میانگین‌گیری از مختصات نقاط تخصیص‌یافته بازتعیین شد تا متناسب‌ترین مکان با توزیع داده‌ها حاصل گردد.



شکل ۳. منطقه‌بندی رویدادهای لرزه‌ای مبتنی بر خوشه‌بندی K–میانگین

برای تضمین پراکندگی مناسب هر منطقه و جلوگیری از ایجاد خوشه‌های ناخواسته کوچک، الگوریتم چندین بار با مقادیر اولیه متفاوت اجرا شد تا هر خوشه حداقل تعداد رویداد لرزه‌ای موردنظر را شامل شود. پس از رسیدن به ثبات برچسب‌ها، هر رویداد به منطقه‌ای اختصاص یافت که مرکز آن از نظر فاصله اقلیدسی به آن نزدیک‌تر بود.

به دلیل تأمین حداقل تعداد داده‌های لرزه‌ای لازم برای هر ناحیه، فرآیند خوشه‌بندی با استفاده از الگوریتم K–میانگین با عدد خوشه (k=4) اعمال گردید به‌گونه‌ای که ناحیه ۰ شامل ۳۰ رویداد، ناحیه ۱ شامل ۸ رویداد، ناحیه ۲ شامل ۱۱ رویداد و ناحیه ۳ شامل ۱۰ رویداد است (شکل ۳).

۲-۶ مدل‌های هوشمند

۲-۶-۱ روش حذف بازگشتی ویژگی با اعتبارسنجی متقابل (RFECV)

انتخاب زیرمجموعه‌ی بهینه‌ای از ویژگی‌های ورودی برای ساخت مدل‌های یادگیری ماشین، فرایندی بنیادین محسوب می‌شود که با کاهش ابعاد مسئله، دقت و کارایی مدل را بهبود می‌بخشد. یکی از روش‌های رایج در این حوزه، حذف بازگشتی ویژگی (RFE) است که با حذف تدریجی ویژگی‌های کم‌اهمیت، عملکرد مدل بهبود می‌یابد. در روش RFECV، مزیت RFE با اعتبارسنجی متقابل ترکیب می‌شود تا تعداد ویژگی‌های بهینه‌ای مشخص گردد که منجر به بیشینه شدن عملکرد پیش‌بینی مدل می‌گردد (عواد و فریحات، ۲۰۲۳).

در روش RFECV، فرایند انتخاب ویژگی به شکل بازگشتی انجام می‌شود. ابتدا مدل یادگیری ماشین بر روی مجموعه اولیه ویژگی‌ها آموزش داده می‌شود و اهمیت هر ویژگی بر اساس وزن یا ضریب تأثیر آن تعیین می‌گردد. سپس کم‌اهمیت‌ترین ویژگی‌ها حذف شده و مدل مجددا روی داده‌های باقی‌مانده آموزش داده می‌شود تا تأثیر حذف این ویژگی‌ها بر کارایی مدل ارزیابی شود. این فرایند حذف بازگشتی تا زمانی تکرار می‌شود که تعداد ویژگی‌ها به حد نصاب مورد نظر برسد یا بهبود دقت پیش‌بینی متوقف شود. در هر تکرار، با اندازه‌گیری دقت پیش‌بینی مدل و بررسی تغییرات آن نسبت به تعداد ویژگی‌های حذف‌شده، زیرمجموعه‌ی بهینه‌ای از ویژگی‌ها انتخاب می‌گردد که بالاترین دقت مدل را به همراه دارد.روش RFECV کاربردهای متعددی در مسائل طبقه‌بندی و رگرسیون دارد که با کاهش ابعاد، پیچیدگی مدل و خطر بیش‌برازش را کاهش می‌دهد و در کاربردهایی مانند تشخیص ناهنجاری و طبقه‌بندی‌های حساس به کاهش ابعاد، نقش مهمی ایفا می‌کند (عواد و فریحات، ۲۰۲۳).

۲-۶-۲ روش نمونه‌برداری افزایشی مصنوعی از کلاس اقلیت (SMOTE)

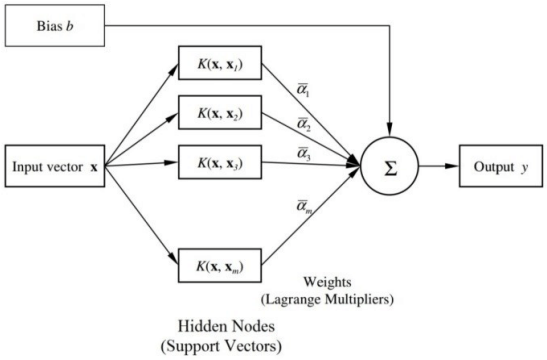
یکی از چالش‌های مهم در یادگیری ماشین، عدم توازن در توزیع کلاس‌های داده است که می‌تواند باعث عملکرد ضعیف مدل‌ها در شناسایی نمونه‌های اقلیت شود. روش SMOTE به‌عنوان یک روش داده‌محور بیش نمونه گیری مطرح شده است. در این روش با تولید مصنوعی نمونه‌های جدید برای کلاس اقلیت، نسبت نمونه‌ها بین کلاس‌ها متعادل می‌شود. بدین صورت که برای هر نمونه اقلیت، همسایه‌های نزدیک آن در فضای ویژگی تعیین شده و سپس با درون‌یابی بین این نمونه‌ها، نمونه‌های مصنوعی جدیدی ایجاد می‌شود. این کار به افزایش تنوع داده‌های اقلیت منجر شده و به مدل یادگیری کمک می‌کند تا بهتر بتواند از روی داده‌های محدود اقلیت تعمیم بیابد. روش SMOTE در بسیاری از حوزه‌های مختلف یادگیری ماشین کاربرد دارد. به‌طور خاص، در مسائلی که عدم توازن کلاس مشاهده می‌شود و شناسایی نمونه‌های اقلیت اهمیت بالایی دارد، از این روش استفاده شده است. کارآیی این روش در این مسائل باعث شده عملکرد الگوریتم‌های طبقه‌بندی در شناسایی صحیح نمونه‌های اقلیت بهبود یابد (التلهان و همکاران، ۲۰۲۵).

روش SMOTE با افزودن تنوع بیشتر به نمونه‌های کلاس اقلیت، توانایی تعمیم مدل یادگیری ماشین را افزایش میدهد. به بیان دیگر، با ایجاد نمونه‌های متنوع‌تر در کلاس اقلیت، فضای ویژگی‌های مربوط به این کلاس گسترده‌تر شده و مدل امکان شناسایی بهینه‌تر نمونه‌های نادر را پیدا می‌کند. همچنین استفاده از SMOTE باعث جلوگیری از حذف احتمالی نمونه‌های مهم کلاس اقلیت می‌شود که این امر در روش‌های ساده‌تر نمونه‌برداری (مانند کم‌نمونه‌برداری) مشکل‌ساز است و می‌تواند منجر به از دست رفتن اطلاعات مهم شود (التلهان و همکاران، ۲۰۲۵).

۲-۶-۳ الگوریتم ماشین بردار پشتیبان (SVM)

ماشین بردار پشتیبان (SVM) الگوریتمی نظارت‌شده برای دسته‌بندی داده‌ها به شمار می‌آید. در این الگوریتم، یک ابرصفحه خطی جداکننده انتخاب می‌شود به نحوی که فاصله (حاشیه) بین نزدیک‌ترین نمونه‌های دو کلاس مختلف به حداکثر برسد. بدین ترتیب تفکیک‌پذیری بین کلاس‌ها افزایش یافته و ریسک طبقه‌بندی نادرست کاهش می‌یابد. همچنین مشخصه مهم SVM این است که تابع تصمیم‌گیری آن تنها بر اساس تعداد اندکی از نمونه‌های آموزشی (بردارهای پشتیبان) تعیین می‌شود (کورتس و وپنیک، ۱۹۹۵).

در فرم اولیه مسئله بهینه‌سازی SVM، پارامترهای مدل (بردار وزن w و بایاس b) به گونه‌ای بهینه می‌شوند که تابع هدف $1/2|w|^2$ مینیمم شود. در این حالت قیدهای خطی $y_i(w^T x_i + b) \geq 1$ برای همه نمونه‌های آموزشی iبرقرار هستند تا هر نمونه در جانبی از ابرصفحه قرار گیرد که با برچسب آن سازگار است (شکل ۴). در مسائل با داده‌های غیرقابل تفکیک خطی معمولا متغیرهای کمکی ξ_i معرفی می‌شوند تا تعدادی از نمونه‌ها مجاز به قرارگیری در ناحیه حاشیه آن باشند. بنابراین قیدها به صورت $y_i(w^T x_i + b) \geq 1 - \xi_i$ با $\xi_i \geq 0$ تنظیم شده و یک ضریب تنظیم‌کننده C برای جلوگیری از خطاهای طبقه‌بندی ناصحیح به تابع هدف افزوده می‌شود. در عمل مسئله بهینه‌سازی با تابع هدف مربعی و قیدهای خطی تبدیل می‌شود که از روش‌های عددی استاندارد قابل حل است (کورتس و وپنیک، ۱۹۹۵).



شکل ۴. فلوچارت معماری ماشین بردار پشتیبان (بوکده و همکاران، ۲۰۱۹).

۲-۶-۴ الگوریتم افزایش گرادیان فوق العاده (XGBoost)

الگوریتم افزایش گرادیان فوق‌العاده (XGBoost) از الگوریتم تقویت درخت گرادیان (GTB) مشتق شده است. در این الگوریتم، بسط تیلور مرتبه دوم بر تابع زیان اعمال می‌شود. برای آموزش، از درخت رگرسیون طبقه‌بندی استفاده شده و سپس جمع وزنی بر روی پیش‌بینی‌ها انجام می‌گیرد (شکل ۵). با تکرار متوالی این فرآیند، یادگیرنده‌ی جدید بر مبنای یادگیرنده‌های قبلی ایجاد می‌گردد. در نتیجه، پس از هر تکرار، مدلی با خطای پیش‌بینی کمتر ساخته می‌شود. برای پیاده‌سازی این الگوریتم، داده‌های آموزشی x_i برای پیش‌بینی متغیر هدف y_i در نظر گرفته می‌شوند که تابع پیش‌بینی به‌صورت رابطه (۳) بیان می‌شود (اسلام و همکاران، ۲۰۲۲؛ رامراج و همکاران، ۲۰۱۶)

$$\hat{y} = \sum_{k=1}^{N} f_k(x_i), \quad f_k \in \mathsf{F} \qquad (۳)$$

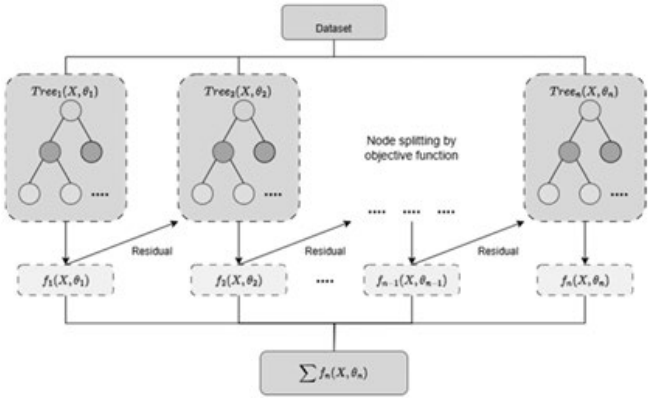
در این رابطه، تعداد درخت‌ها با N مشخص شده و فضای درخت‌ها با $\mathcal{F}$ تعریف گردیده است، به‌طوری که هر تابع f_k از $\mathcal{F}$ انتخاب می‌شود. تابع هدف در تکرار t به‌صورت رابطه (۴) پیشنهاد شده است.

$$\mathrm{obj}^{t} = \sum_{i=1}^{n} l(y_i,\ \hat{y}_i^{\,t-1} + f_t(x_i)) + \Omega(f_t) \qquad (۴)$$

در این رابطه، اختلاف پیش‌بینی $\hat{y}_i^{t-1}$ و مقدار واقعی y_i توسط تابع زیان $l(0,0)$ اندازه‌گیری شده و $\Omega(f_t)$ به‌عنوان ضریب تنظیم‌کننده‌ی پیچیدگی مدل به شکل رابطه (۵) تعریف گردیده است.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda |w|^2 \qquad (۵)$$

که در آن T تعداد برگ‌های درخت، W بردار وزن‌های برگ‌ها، γ پارامتر جریمه‌ی تعداد برگ و λ ضریب جریمه‌ی وزن‌ها می‌باشد.



شکل ۵. فلوچارت معماری الگوریتم افزایش گرادیان فوق‌العاده (هیدیاتوررحمان و هانادا، ۲۰۲۴).

۲-۶-۵ مدل طبقه‌بندی ترکیبی تخمین موقعیت

برای آماده‌سازی داده‌های ورودی مدل، مقادیر همبستگی متقابل به‌دست‌آمده از هر یک از پنج ایستگاه و سه مؤلفه (در مجموع ۱۵ ماتریس) در یک ماتریس واحد تجمیع شد. در این فرایند، شناسه منحصربه‌فرد هر رویداد به‌عنوان محور مشترک به کار گرفته شد تا ضرایب همبستگی همه ترکیب‌ها برای هر رویداد کنار هم قرار گیرند. با بهره‌گیری از الگوهای نام‌گذاری (evXX_STYY_C) اطلاعات مربوط به شناسه رویداد (ev)، ایستگاه (ST) و مولفه شتاب (C) استخراج و سایر ستون‌ها به‌عنوان ویژگی‌های همبستگی مشخص گردیدند.

در ادامه، برای هر جفت رویداد–مولفه، شاخص‌های آماری (میانگین و انحراف معیار ضرایب پنج ایستگاه) محاسبه و به ردیف مربوط به آن رویداد افزوده شد. و با تبدیل ساختار داده‌ها به قالب رویداد–ویژگی، ماتریسی حاصل شد که هر سطر نشان‌دهنده یک رویداد در یک ایستگاه–مولفه (جمعا ۸۸۵ سطر؛ شامل ۵۹ رویداد با ۳ مولفه ثبت شده در ۵ ایستگاه) است و ستون‌های ویژگی آن شامل ضرایب همبستگی متقابل و شاخص‌های آماری متناظر (دو شاخص آماری میانگین و انحراف معیار) می‌باشد. ستون آخر این ماتریس نیز برچسب منطقه‌بندی تولیدشده توسط الگوریتم K–میانگین را نمایش می‌دهد. به‌عنوان نمونه، ردیف‌های مربوط به رویداد ۱ و مولفه x در پنج ایستگاه با شناسه‌های متوالی ev01_ST01_X تا ev01_ST05_X در ماتریس نهایی دیده می‌شوند (شکل ۶). برای تمامی گام‌های زمانی، داده‌های ورودی مدل مطابق با ساختار ماتریس رویداد–ویژگی تهیه می‌شوند.

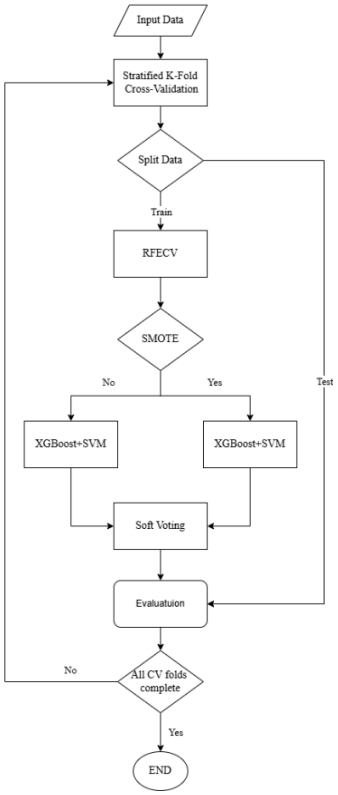
events	ev01	ev02	...	ev59	ev01_mean	ev02_mean	...	ev59_mean	ev01_std	ev02_std	...	ev59_std	region
ev01_ST01_X	1.0000	0.2855	...	0.6859	1.0000	0.7530	...	0.7734	0.0000	0.2819	...	0.1058	0
ev01_ST02_X	1.0000	0.8637	...	0.8287	1.0000	0.7530	...	0.7734	0.0000	0.2819	...	0.1058	0
ev01_ST03_X	1.0000	0.9083	...	0.9252	1.0000	0.7530	...	0.7734	0.0000	0.2819	...	0.1058	0
ev01_ST04_X	1.0000	0.9997	...	0.7576	1.0000	0.7530	...	0.7734	0.0000	0.2819	...	0.1058	0
ev01_ST05_X	1.0000	0.7077	...	0.6697	1.0000	0.7530	...	0.7734	0.0000	0.2819	...	0.1058	0

شکل ۶. نمایی از ماتریس رویداد–ویژگی ورودی مدل پس از تجمیع ضرایب هم‌بستگی متقابل و شاخص‌های آماری متناظر. (ev) شماره رویداد، (ST) شناسه ایستگاه، (X) مولفه شتاب، (mean) میانگین و (std) انحراف معیار ضرایب پنج ایستگاه، و (region) برچسب منطقه‌بندی است.

به منظور سنجش تعمیم‌پذیری، اعتبارسنجی متقابل با ۵ تکرار (در هر مرحله ۴ قسمت برای آموزش و یک قسمت برای آزمایش) به کار گرفته شد؛ در هر تکرار، نمونه‌ها به گونه‌ای تقسیم شدند که توزیع برچسب‌ها در زیرمجموعه‌ها محفوظ بماند. ارزیابی متقاطع بدین ترتیب امکان ارائه برآوردی واقع‌بینانه از عملکرد متوسط مدل و پراکندگی آن را فراهم ساخت و از بیش‌برازش به داده‌های خاص یک تقسیم جلوگیری نمود.

برای انتخاب بهینه زیرمجموعه‌ای از متغیرها وکاهش بعد ویژگی‌ها، فرایند حذف بازگشتی ویژگی‌ها با اعتبارسنجی متقابل (RFECV) بر داده‌های آموزشی به کار گرفته شد. این روش به‌صورت تکراری کم‌اهمیت‌ترین ویژگی‌ها را کنار گذاشت و عملکرد مدل را در هر گام از طریق اعتبارسنجی متقابل لایه‌بندی‌شده بررسی کرد تا در نهایت مجموعه‌ای با بیشترین توان تفکیک استخراج شود.

برای سنجش پیامد متوازن‌سازی، دو پیکربندی مستقل ارزیابی شد: در نخستین پیکربندی، مدل بدون به‌کارگیری SMOTE آموزش داده شد؛ در دومین پیکربندی، تنها داده‌های آموزشی هر تکرار با SMOTE باز نمونه‌برداری گردید تا کمبود نمونه در کلاس‌های اقلیت جبران شود. SMOTE صرفا بر بخش آموزشی اعمال شد تا از نشت اطلاعات به داده‌های آزمایش جلوگیری و برآورد تعمیم‌پذیری بدون سوگیری حفظ شود فلوچارت کامل این مدل طبقه‌بندی در شکل ۷ نمایش داده شده است.



شکل ۷. فلوچارت مدل طبقه‌بندی ترکیبی تخمین موقعیت.

پس از تکمیل مرحله انتخاب ویژگی، یک مدل طبقه‌بندی ترکیبی بر پایه (Soft Voting) ساخته شد.هر یک از مدل‌های ماشین بردار پشتیبان خطی (SVM) و الگوریتم افزایش گرادیان فوق‌العاده (XGBoost) برای هر رویداد احتمال تعلق به هر منطقه را جداگانه پیش‌بینی کرده، سپس احتمال‌های پیش‌بینی‌شده هر کلاس میان دو مدل میانگین‌گیری می‌شود و بر اساس بیشینه احتمال تجمیعی تصمیم‌گیری می‌گردد؛ در این چارچوب، ماشین بردار پشتیبان خطی و الگوریتم افزایش گرادیان فوق‌العاده به‌صورت موازی به کار رفتند. این فرایند برای هر گام زمانی یک‌بار بدون به‌کارگیری SMOTE و یک‌بار پس از اعمال SMOTE بر داده‌های آموزشی اجرا شد.

SVM با استفاده از کرنل خطی و اتکا بر بیشینه‌سازی حاشیه تفکیک، به‌دلیل مقاومت در برابر نویز و کارایی بالا در فضاهای ویژگی با ابعاد زیاد، توانایی ترسیم مرزهای تصمیم ساده و تعمیم‌پذیر را ارائه کند؛ در حالی‌که XGBoost به سبب ساختار درختی و قابلیت مدل‌سازی روابط غیرخطی و برهم‌کنش‌های پیچیده، نقاط ضعف مدل خطی را پوشش می‌دهد. ترکیب احتمالات خروجی این دو مدل در چارچوب رأی‌دهی با احتمال موجب گردید که مزایای تفکیک خطی و قدرت بیان غیرخطی توامان بهره‌برداری شده و واریانس خطا کاهش یابد.

۲-۶-۶ بهینه‌سازی فراپارامترها مدل طبقه‌بندی ترکیبی

به‌منظور تنظیم دقیق فراپارامترهای مدل‌های ماشین بردار پشتیبان خطی وXGBoost ، در گام نخست با انجام آزمون‌های مقدماتی، بازه‌ی اولیه‌ی مقادیر برای هر فراپارامتر تعیین گردید. سپس یک جست‌وجوی شبکه‌ای (Grid Search) دومرحله‌ای همراه با اعتبارسنجی متقابل پنج‌گانه اجرا شد: در مرحله‌ی اول جست‌وجوی اولیه برای محدودسازی دامنه و در مرحله‌ی دوم جست‌وجوی

تفصیلی برای یافتن مقادیر نهایی به کار گرفته شد. معیار انتخاب، میانگین دقت طبقه‌بندی در هر تکرار اعتبارسنجی بود و پس از تعیین ترکیب بهینه، مدل نهایی با همان فراپارامترها روی کل داده‌های آموزشی بازآموزی شد.

در فرایند جست‌وجوی شبکه‌ای، در ماشین بردار پشتیبان، ضریب خطا (C) که میزان جریمه خطاهای طبقه‌بندی را کنترل می‌کند، مقدار ۱/۰ انتخاب گردید تا توازنی مناسب میان سوگیری و واریانس برقرار شود؛ آستانه همگرایی (tol) نیز ۰/۰۰۵ برگزیده شد تا از توقف زودهنگام فرایند بهینه‌سازی جلوگیری گردد. در XGBoost، نسبت نمونه‌برداری ردیف‌ها (subsample) برابر ۰/۶ تعیین شد تا واریانس مدل کاهش یابد، در حالی‌که نسبت نمونه‌برداری ویژگی‌ها (colsample_bytree) بر ۱/۰ ثابت ماند تا کلیه ویژگی‌ها برای ساخت هر درخت در دسترس باشند. تعداد درخت‌ها (n_estimators) به ۳۵ محدود گردید و بیشینه عمق درخت (max_depth) روی ۶ تنظیم شد تا پیچیدگی مدل کنترل شود. نرخ یادگیری (learning_rate) با مقدار ۰/۰۰۱ انتخاب شد تا هر درخت سهم اندکی در مدل نهایی داشته باشد و از بیش‌برازش جلوگیری شود. همچنین، آستانه هرس (gamma) ۰/۲ و جریمه (reg_alpha) ۰/۰۱ برگزیده شدند تا تقسیم گره‌های کم‌اهمیت محدود و در نتیجه از بیش‌برازش ممانعت گردد.

۳ نتایج و بحث

در این بخش ابتدا پارامترهای مورد ارزیابی معرفی شده‌اند. سپس نتایج تحلیل حساسیت مربوط به پارامترهای اصلاح خط مبنا ارائه می‌شوند. در ادامه، نتایج اصلاح خط مبنا گزارش گردیده و در نهایت، نتایج مدل طبقه‌بندی تخمین موقعیت ابتدا برای هر چهار گام زمانی بدون بهره‌گیری از SMOTE و پس از آن با استفاده از SMOTE ارائه و مورد بررسی قرار خواهند گرفت.

۳-۱ پارامترهای مورد ارزیابی

در این زیربخش، پارامترهای ارزیابی تحلیل حساسیت و پارامترهای ارزیابی مدل طبقه‌بندی ترکیبی تخمین موقعیت به‌تفکیک معرفی می‌شوند.

۳-۱-۱ پارمترهای ارزیابی آنالیز حساسیت

برای ارزیابی دقت و قابلیت اطمینان مدل در تشخیص نمونه‌های مثبت و منفی، از دو معیار زیر بهره گرفته می‌شود:

- نرخ مثبت صحیح (TPR): این پارامتر به‌صورت رابطه (۶) تعریف می‌شود.

$$TPR = \frac{TP}{TP + FN} \qquad (۶)$$

که در آن TP تعداد نمونه‌های مثبت به‌درستی شناسایی‌شده و FN تعداد نمونه‌های مثبت به اشتباه تخیص داده شده را نشان می‌دهد.

- نرخ مثبت کاذب (FPR): این پارامتر به‌صورت رابطه (۷) تعریف می‌شود.

$$FPR = \frac{FP}{FP + TN} \qquad (۷)$$

که در آن FP تعداد نمونه‌های منفی به‌اشتباه شناسایی‌شده به‌عنوان مثبت و TN تعداد نمونه‌های منفی به‌درستی شناسایی‌شده است.

۳-۱-۲ پارامترهای ارزیابی مدل طبقه‌بندی ترکیبی تخمین موقعیت

در مسائل دسته‌بندی، هدف پیش‌بینی دسته یا گروهی از رکوردها یا داده‌ها است که به آن‌ها تعلق دارد. در این پژوهش مدل دسته‌بندی چندگانه جهت تفکیک نواحی گوناگون رویدادهای لرزه‌ای به‌کار گرفته شده است. تمامی مفاهیم مطرح‌شده در ادامه تحقیق به‌سادگی قابل تعمیم به سایر مسائل دسته‌بندی چندگانه خواهند بود، مقایسه خروجی مدل با برچسب‌های واقعی به صورت زیر انجام می‌گیرد.

تلاش می‌شود که خطاها (نادرست مثبت و نادرست منفی) به صفر کاهش یابند، اما در عمل این امر محقق نمی‌شود؛ بنابراین، نیاز به مکانیزم‌ها و معیارهایی برای سنجش دقت، صحت و کارایی مدل ایجادشده از داده‌های موجود احساس می‌گردد. این معیارها به‌صورت زیر ارائه می‌شوند:

-دقت طبقه‌بندی، به‌عنوان اولین معیار سنجش، میزان تشخیص درست مدل را نشان می‌دهد و برای مسائل چندکلاسه نیز به‌صورت رابطه (۸) تعریف می‌گرد.

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}} \qquad (۸)$$

که در آن TP_i تعداد نمونه‌هایی است که واقعی آن‌ها کلاس i بوده و به درستی در کلاس i پیش‌بینی شده‌اند، N_{ij} تعداد نمونه‌هایی است که واقعی آن‌ها کلاس i بوده و به کلاس j تخصیص یافته‌اند، و $\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}$ کل تعداد نمونه‌ها را نشان می‌دهد.

-معیار بازخوانی، نسبتی از تعداد نتایج مثبت صحیح به تعداد کل نمونه‌های مثبت واقعی را نشان می‌دهد. برای مسائل چندکلاسه، بازخوانی ماکرو مطابق رابطه (۹) تعریف می‌شود.

$$\text{Recall}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n}\text{Recall}_i$$
$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \qquad (۹)$$

در این رابطه، TP_i تعداد نمونه‌هایی است که واقعا به کلاس iتعلق داشته و به درستی در همان کلاس پیش‌بینی شده‌اند، و FN_i تعداد نمونه‌هایی است که به کلاس i تعلق داشته اما به اشتباه در کلاس‌های دیگر طبقه‌بندی شده‌اند.

-صحت، نسبت تعداد نمونه‌های مثبت به‌درستی شناسایی‌شده به کل نمونه‌های پیش‌بینی‌شده مثبت را نشان می‌دهد و برای مسائل چندکلاسه به‌صورت دقت ماکرو در رابطه (۱۰) تعریف می‌گردد.

$$\text{Precision}_{\text{macro}} = \frac{1}{n}\sum_{i=1}^{n}\text{Precision}_i$$
$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \qquad (۱۰)$$

در این روابط، TP_i تعداد نمونه‌هایی است که واقعی آن‌ها کلاس i بوده و به درستی در همان کلاس پیش‌بینی شده‌اند، و FP_i تعداد نمونه‌هایی است که به اشتباه در کلاس i قرار گرفته‌اند. در صورتی که تعداد پیش‌بینی‌های مثبت نادرست (مؤلفه FP_i) زیاد باشد، مخرج کسر افزایش یافته و مقدار دقت به عددی نزدیک به صفر میل می‌نماید؛ از این رو، نشان‌دهنده کارایی پایین مدل تلقی می‌شود.

با توجه به مطالب فوق، معیارهای بازخوانی و صحت در مقایسه با معیار دقت اولیه، کاربرد وسیع‌تری در حوزه یادگیری ماشین یافته‌اند. به‌جای به‌کارگیری همزمان این دو معیار، پیشنهاد می‌گردد از یک معیار ترکیبی برای ارزیابی الگوریتم‌های دسته‌بندی بهره گرفته شود؛ به این ترتیب، تمرکز بر روی این معیار ترکیبی به‌جای دو معیار مستقل، مطلوب‌تر تلقی می‌گردد. برای مثال، اگر میانگین حسابی بازخوانی و صحت در شرایطی که صحت بالا و بازخوانی پایین (یا بالعکس) باشد، به‌کار گرفته شود، عدد به‌دست‌آمده ممکن است گویای عملکرد واقعی الگوریتم نباشد. از این رو، میانگین هارمونیک به‌عنوان معیاری جایگزین تعریف شده و با گرایش به مقدار کمتر، نقیصه فوق را رفع می‌نماید. این معیار ترکیبی (F1-Score) مطابق رابطه (۱۱) ارائه می‌گردد.

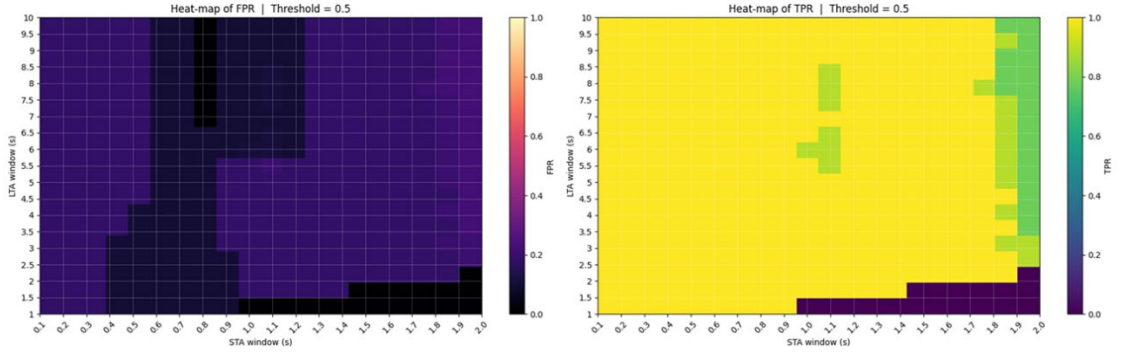
$$F_{1,\text{macro}} = \frac{1}{n}\sum_{i=1}^{n} F_{1,i}$$
$$F_{1,i} = 2\times\frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \qquad (۱۱)$$

که در آن Precision_i و Recall_i هر یک مطابق روابط (۹) و (۱۰) محاسبه می‌گردند. پارامترهای ارزیابی مدل

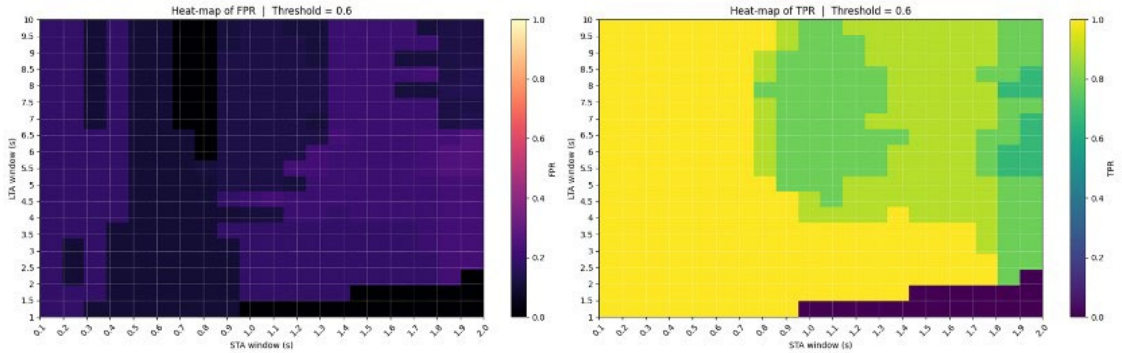
طبقه‌بندی (روابط ۸ تا ۱۱) همگی در بازه ۰ تا ۱ تعریف شده‌اند.

۳-۲ آنالیز حساسیت پارامترهای الگوریتم تصحیح خط مبنا

در این بخش، آنالیز حساسیت پارامترهای الگوریتم تصحیح خط مبنا ارائه می‌گردد. در این راستا، پنجره STA در بازه ۰/۱ تا ۲/۰ با گام ۰/۱، پنجره LTA در بازه ۱/۰ تا ۱۰/۰ با گام ۰/۵ و آستانه نسبت STA/LTA در بازه ۰/۱ تا ۰/۹ با گام ۰/۱ انتخاب شده و یک تحلیل حساسیت شبکه‌ای بر روی آن‌ها صورت پذیرفته است. این بازه‌ها به‌گونه‌ای انتخاب شدند که پوشش کاملی از مقادیر عملیاتی متداول الگوریتم STA/LTA را فراهم آورند و امکان بررسی دقیق تأثیر هر پارامتر بر نرخ شناسایی صحیح و هشدار کاذب را مهیا سازند. در هر ترکیب از پارامترهای فوق، نرخ مثبت صحیح و نرخ هشدار کاذب (برای تشخیص و اصلاح جابه‌جایی ناگهانی خط مبنا) محاسبه شده است. نتایج حاصل فقط برای دو حد آستانه ۰/۵ (شکل ۸-الف) و ۰/۶ (شکل ۸-ب) که با دقت کامل تشخیص داده‌ند نمایش داده شده‌اند.



الف) نتایج حد آستانه ۰/۵

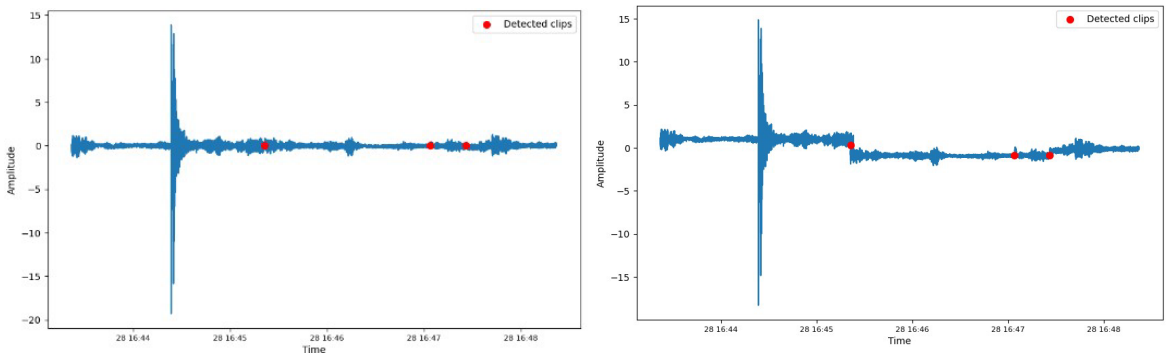


الف) نتایج حد آستانه ۰/۶

شکل ۸. نتایج نرخ مثبت صحیح و نرخ هشدار کاذب برای پارامترهای تصحیح خط مبنا برای حد آستانه ۰/۵ و ۰/۶.

با توجه به نتایج بهترین توازن پارامترها برای تشخیص جهش‌های ناگهانی بدون خطا در حد آستانه ۰/۵ با پنجره STA برابر ۰/۸ و LTA بین ۷ تا ۱۰ و نیز حد آستانه ۰/۶ با پنجره STA برابر ۰/۷ و LTA در بازه ۷ تا ۱۰ و برای حد آستانه ۰/۶ هنگامی که پنجره STA روی ۰/۸ تنظیم گردیده، بازه LTA برابر ۹/۵ تا ۱۰حاصل گردیده است.

پس از شناسایی جهش‌های ناگهانی خط مبنا، اصلاح با کم کردن میانگین‌های هر بازه از یکدیگر، همانند شکل ۹، صورت می‌گیرد.



شکل ۹. تشخیص جهش‌های ناگهانی خط مبنا (سمت راست) و نتایج اصلاح آن (سمت چپ).

۳-۳ نتایج مدل طبقه‌بندی ترکیبی تخمین موقعیت

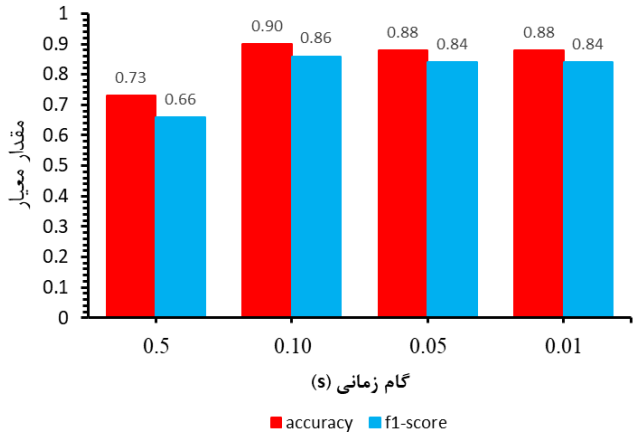
۳-۳-۱ نتایج مدل طبقه‌بندی ترکیبی بدون استفاده از SMOTE

در این بخش، نتایج حاصل از مدل طبقه‌بندی ترکیبی بدون استفاده از SMOTE در گام‌های زمانی ۰/۵، ۰/۱، ۰/۰۵ و ۰/۰۱ با اعتبارسنجی متقابل در ۵ تکرار گزارش شدند. در جدول ۲ نتایج معیارها ارزیابی (میانگین کل برای هر ۵ تکرار اعتبار سنجی متقابل) با انحراف معیار مدل طبقه‌بندی ترکیبی بدون SMOTE ارائه شده‌است. با توجه به جدول ۲ و شکل ۱۰ با کاهش گام زمانی از ۰/۵ به ۰/۱ ثانیه، دقت طبقه‌بندی مدل در تعیین منطقه‌ی رومرکز منبع لرزه‌ای به‌طور معناداری از ۰/۷۳ به ۰/۹۰ افزایش یافت، در حالی که F1-score نیز از ۰/۶۶±۰/۲۵ به ۰/۸۶±۰/۰۷ ارتقاء یافت؛ این امر نشانگر بهبود تفکیک‌پذیری الگوها شکل موج با ثبت تغییرات سریع‌تر شکل موج است. همچنین مشاهده شد که انحراف معیار معیارهای ارزیابی در پنجره ۰/۱ ثانیه به‌میزان چشمگیری کاهش می‌یابد، که پایداری بیشتر برآوردها را دلالت می‌کند.

جدول ۲. نتایج معیارهای ارزیابی مدل طبقه‌بندی ترکیبی بدون SMOTE.

گام زمانی	accuracy	precision	recall	F1-score
۰/۵	۰/۷۳	۰/۷۶ ± ۰/۱۸	۰/۶۳ ± ۰/۳۱	۰/۶۶ ± ۰/۲۵
۰/۱	۰/۹۰	۰/۸۹ ± ۰/۰۲	۰/۸۴ ± ۰/۱۱	۰/۸۶ ± ۰/۰۷
۰/۰۵	۰/۸۸	۰/۸۶ ± ۰/۱۱	۰/۸۳ ± ۰/۱۳	۰/۸۴ ± ۰/۱۲
۰/۰۱	۰/۸۸	۰/۸۶ ± ۰/۱۰	۰/۸۳ ± ۰/۱۳	۰/۸۴ ± ۰/۱۰

افزون بر این، افزایش اندک انحراف معیار F1 در گام‌های زمانی ۰/۰۵ و ۰/۰۱ ثانیه مشاهده گردید که می‌تواند ناشی از حساسیت بالاتر مدل به نوسانات نویزی در مقیاس‌های زمانی بسیار کوتاه باشد؛ به‌عبارتی، تحلیل بیش از حد ریزشکل موج موجب بزرگنمایی اختلالات نویزی و در نتیجه ناپایداری برآوردها شده است.



شکل ۱۰. نمودار مقایسه معیار ارزیابی مدل طبقه‌بندی ترکیبی بدون SMOTE.

بدین‌ترتیب، گام زمانی ۰/۱ ثانیه به‌عنوان پنجره‌ای بهینه تلقی می‌شود که در آن تعادلی مطلوب میان ثبت تغییرات آنی شکل موج و حفظ پایداری نتایج برقرار شده است. در حالی که انتخاب پنجره‌های کوچکتر، جزئیات بیشتری را فراهم می‌آورد، اما با افزایش چشمگیر هزینه‌ی محاسباتی استخراج هم‌بستگی متقابل و تشدید اثرات نویز، ناپایداری نسبی در معیارهای مدل ایجاد می‌گردد؛ بنابراین، افزایش پیچیدگی محاسباتی در گام‌های کمتر از ۰/۱ ثانیه، دیگر نمی‌توان عملکرد بهتر را توجیه کرد.

۳-۳-۲ نتایج مدل طبقه‌بندی ترکیبی با استفاده از SMOTE

در این بخش، نتایج حاصل از مدل طبقه‌بندی ترکیبی با استفاده از SMOTE مربوط به گام‌های زمانی ۰/۵، ۰/۱، ۰/۰۵ و ۰/۰۱ با اعتبارسنجی متقابل در ۵ تکرار ارائه شده‌اند (به‌دلیل نبود همسایه‌های مناسب و فاصله زیاد بین نمونه‌ها و عدم ایجاد داده مصنوعی نویزی SMOTE روی ناحیه‌ی ۱ اجرا نشده است). معیارها (میانگین کل) با انحراف معیار برای داده‌ها آزمایشی برای هر چهار گام زمانی در جدول ۳ و نمودار مقایسه آن با نتایج بدون SMOTE در شکل ۱۱ آورده شده‌اند.

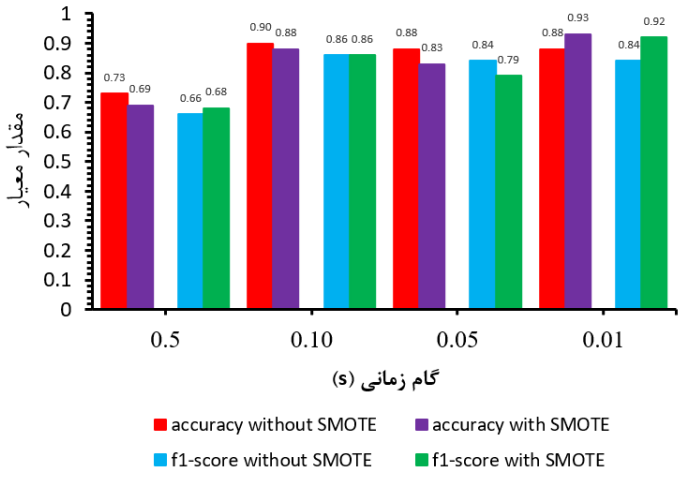
جدول ۳. نتایج معیارهای ارزیابی مدل طبقه‌بندی ترکیبی با SMOTE

گام زمانی	accuracy	precision	recall	F1-score
۰/۵	۰/۶۹	۰/۶۸ ± ۰/۱۴	۰/۶۹ ± ۰/۱۴	۰/۶۸ ± ۰/۱۴
۰/۱	۰/۸۸	۰/۹۲ ± ۰/۰۸	۰/۸۲ ± ۰/۱۱	۰/۸۶ ± ۰/۰۴
۰/۰۵	۰/۸۳	۰/۸۵ ± ۰/۰۹	۰/۷۶ ± ۰/۱۳	۰/۷۹ ± ۰/۰۷
۰/۰۱	۰/۹۳	۰/۹۳ ± ۰/۰۹	۰/۹۱ ± ۰/۰۳	۰/۹۲ ± ۰/۰۵

با توجه به جداول ۳ و ۴ و شکل ۱۱ با اعمال SMOTE و کاهش گام زمانی از ۰/۵ به ۰/۱ ثانیه، دقت طبقه‌بندی آن نسبت به SMOTE کاهش پیدا کرده است و F1-score نسبت به حالت SMOTE تفاوت چندانی نداشته است؛ در عین حال، انحراف معیار آن نسبت به حالت بدون SMOTE تقریبا به نصف کاهش یافت. این امر حاکی از آن است که در پنجره ۰/۱ ثانیه، تخمین پایدارتری فراهم آورد.

در گام ۰/۰۵ ثانیه، دقت طبقه‌بندی مدل و F1-score به ترتیب به ۰/۸۳ و ۰/۷۹±۰/۰۷ رسیدند که نسبت به حالت بدون SMOTE (۰/۸۸ و ۰/۸۴±۰/۱۲) کاهش نشان داد. مشاهده‌ی مذکور بیانگر آن است که در این مقیاس میانی، نمونه‌های مصنوعی تولیدشده تنوع کافی برای شناسایی جزئیات شکل موج را نداشته و در نتیجه، برهم‌کنش آن‌ها با نویز منجر به تضعیف کارایی مدل گردیده است در عین حال تخمین پایدارتری فراهم آورد.

در گام ۰/۰۱ ثانیه، بیشترین بهره‌وری از SMOTE مشاهده شد؛ دقیقا دقت طبقه‌بندی ۰/۹۳ و F1-score معادل ۰/۹۲±۰/۰۵ به دست آمد، در حالی که مقادیر متناظر بدون SMOTE ۰/۸۸ و ۰/۸۴±۰/۱۰ بوده است. این بهبود آشکار بیانگر آن است که در مقیاس زمانی بسیار کوتاه، غنای اطلاعات لحظه‌ای موج به‌درستی توسط SMOTE بازتولید شده و توانسته است نه‌تنها تفکیک‌پذیری الگوها و دقت طبقه‌بندی مدل را افزایش دهد، بلکه نوسانات نویزی را نیز به‌طور مؤثری تعدیل نماید.



شکل ۱۱. نمودار مقایسه معیار ارزیابی مدل طبقه‌بندی ترکیبی با SMOTE و بدون آن.

به‌طور کلی، مشاهده گردید که SMOTE در کنار کاهش گام زمانی، روندی کاهشی بر انحراف معیار شاخص‌ها اعمال نمود و پایداری مدل را به‌ویژه در گام‌های ۰/۱ و ۰/۰۱ ثانیه بهبود بخشید؛ هرچند در گام‌های ۰/۵، ۰/۱ و ۰/۰۵ ثانیه، این هم‌پوشانی به علت کیفیت پایین‌تر نمونه‌های مصنوعی و احتمال هم‌افزایی با نویز، (کم شدن دقت طبقه‌بندی مدل) مزیتی در پی نداشت. از دیدگاه پیچیدگی محاسباتی، افزودن SMOTE موجب افزایش زمان آموزش گردید، اما هزینه‌ی پردازش سیگنال (محاسبه هم‌بستگی متقابل) را تغییر نداد؛ بنابراین، در گام‌های ریز که از پیش هزینه‌ی سیگنال‌پردازی بالا است، بهبود ۵/۷ درصدی دقت طبقه‌بندی مدل و ۹/۵ درصدی F1-score (گام ۰/۰۱ ثانیه) مزیت قابل‌توجهی به حساب آمده است.

بر این اساس، در صورتی که محدودیت محاسباتی چندان بحرانی نباشد، ترکیب گام زمانی ۰/۰۱ ثانیه همراه با SMOTE به‌عنوان پیکربندی برتر پیشنهاد می‌شود؛ اما در شرایط منابع محدود، گام ۰/۱ ثانیه بدون SMOTE همچنان نقطه تعادلی مناسب بین هزینه و دقت طبقه‌بندی محسوب می‌گردد.

۴ نتیجه‌گیری

در این پژوهش، چارچوبی چندمرحله‌ای برای تخمین موقعیت منطقه‌ای رومرکز چشمه لرزه‌ای پیشنهاد و به‌کار گرفته شد. در گام نخست، سیگنال‌های سه‌مولفه شتاب استخراج و با بهره‌گیری از نسبت STA/LTA برای شناسایی و اصلاح جهش‌های ناگهانی خط‌مبنا پیش‌پردازش گردیدند. در ادامه، ماتریس‌های همبستگی متقابل در چهار مقیاس زمانی متفاوت (در هر گام، پنجره‌های زمانی شکل موج با جابجایی‌های زمانی متفاوت ۰/۵، ۰/۱، ۰/۰۵ و ۰/۰۱ ثانیه روی هم منطبق شدند) محاسبه و به‌منظور استخراج زیرمجموعه بهینه ویژگی‌ها روش حذف بازگشتی ویژگی با اعتبارسنجی متقابل (RFECV) به‌کار گرفته شد. بردارهای ویژگی حاصل، به‌عنوان ورودی به یک مدل طبقه‌بندی ترکیبی مبتنی بر Soft Voting شامل طبقه‌بندی‌های SVM و XGBoost ارائه شدند تا احتمال‌های خروجی آن‌ها برای دستیابی به پیش‌بینی‌ای پایدار تلفیق گردد. برای بررسی تأثیر توزیع داده‌ها بر عملکرد، فرآیند آموزش و ارزیابی مدل یک‌بار بدون و بار دیگر با اعمال SMOTE برای متعادل‌سازی نمونه‌های اقلیت انجام شد و کارایی سیستم از نظر دقت مکانیابی و پایداری در شرایط مختلف ارزیابی شد. نتایج به‌دست‌آمده نشان دادند که چارچوب پیشنهادی ضمن حفظ دقت بالا و پاسخ‌دهی مناسب، قابلیت بهره‌برداری عملیاتی را در شرایط با داده‌های محدود فراهم می‌آورد.

یافته‌های پژوهش نشان می‌دهند که پارامترهای بهینه برای تشخیص و اصلاح جهش ناگهانی خط‌مبنا در گام پیش‌پردازش شامل حد آستانه ۰/۵ با پنجره STA برابر ۰/۸ و LTA بین ۷ تا ۱۰ و نیز حد آستانه ۰/۶ با پنجره STA برابر ۰/۷ و LTA در بازه ۷ تا ۱۰ انتخاب شده‌اند، به‌نحوی که تمامی جابه‌جایی‌ها به‌دقت شناسایی و اصلاح گردیده است. علاوه بر این، برای حد آستانه هنگامی که پنجره STA روی ۰/۸ تنظیم گردیده، بازه LTA برابر ۹/۵ تا ۱۰ برگزیده شد تا کامل‌ترین پوشش اصلاحی حاصل شود.

بر اساس نتایج به‌دست‌آمده برای مدل بدون SMOTE ، کاهش گام زمانی از ۰/۵ به ۰/۱ ثانیه منجر به افزایش معنادار دقت طبقه‌بندی مدل در تعیین منطقه‌ی رومرکز منبع لرزه‌ای از ۰/۷۳ به ۰/۹۰ و ارتقای F1-score از ۰/۶۶±۰/۲۵ به ۰/۸۶±۰/۰۷ گردید. همچنین انحراف معیار شاخص‌ها به‌طور چشمگیری کاهش یافت که پایداری بیشتر برآوردها را دلالت می‌کند. در گام‌های زمانی ۰/۰۵ و ۰/۰۱ ثانیه، با وجود اشباع عملکرد، افزایش اندک انحراف معیار مشاهده شد که ناشی از حساسیت بالاتر مدل به نوسانات نویزی در مقیاس‌های بسیار کوتاه قلمداد می‌گردد. بنابراین، گام ۰/۱ ثانیه به‌عنوان نقطه تعادلی مطلوب میان ثبت جزئیات آنی موج و حفظ پایداری نتایج معرفی شده است.

یافته‌ها حاکی‌اند که در صورت اعمال SMOTE ، پایداری تخمین‌ها در تمامی گام‌ها بهبود یافته و انحراف معیار شاخص‌ها تا حدود نیمی کاهش یافته است؛ اما در گام‌های ۰/۵، ۰/۱ و ۰/۰۵ ثانیه افت جزئی دقت طبقه‌بندی مدل مشاهده گردید که ناشی از کیفیت پایین نمونه‌های مصنوعی به‌دلیل نداشتن جزئیات کامل شکل موج در این گام‌ها و تعامل مخرب با نویز بوده است. در مقابل، در گام ۰/۰۱ ثانیه با SMOTE بیشترین بهره‌وری محقق گردید (دقت طبقه‌بندی ۰/۹۳ و F1-score ۹۲/۰۵±۰/۰ در برابر ۰/۸۸ و ۰/۸۴±۰/۱۰ بدون SMOTE و بهبود ۵/۷ درصدی دقت طبقه‌بندی مدل و ۹/۵ درصدی F1-score را دارد)، به‌نحوی که هم تفکیک‌پذیری الگوها افزایش و هم نوسانات نویزی تعدیل شد. بر این اساس، ترکیب گام ۰/۰۱ ثانیه با SMOTE به‌عنوان پیکربندی برتر برای محیط‌های با محدودیت داده‌ای و اولویت پایداری تخمین پیشنهاد می‌شود؛ هرچند در شرایط منابع محاسباتی محدود، گام ۰/۱ ثانیه بدون SMOTE همچنان به‌عنوان گزینه‌ای متوازن از نظر هزینه و دقت طبقه‌بندی توصیه می‌گردد.

منابع

Abdulrahman, L. M., Abdulazeez, A. M., & Hasan, D. A. (2021). COVID-19 world vaccine adverse reactions based on machine learning clustering algorithm. Qubahan Academic Journal, 1(2), 134-140.

Altalhan, M., Algarni, A., & Alouane, M. T. H. (2025). Imbalanced Data problem in Machine Learning: A review. IEEE Access.

Awad, M., & Fraihat, S. (2023). Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. Journal of Sensor and Actuator Networks, 12(5), 67.

Bilal, M. A., Ji, Y., Wang, Y., Akhter, M. P., & Yaqub, M. (2022). Early earthquake detection using batch normalization graph convolutional neural network (bngcnn). Applied Sciences, 12(15), 7548.

Bokde, N., Feijóo, A., Villanueva, D., & Kulat, K. (2019). A review on hybrid empirical mode decomposition models for wind speed and wind power prediction. Energies, 12(2), 254.

Chen, Y., Saad, O. M., Savvaidis, A., Chen, Y., & Fomel, S. (2022). 3D microseismic monitoring using machine learning. Journal of Geophysical Research: Solid Earth, 127(3), e2021JB023842.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Ellsworth, W. L. (2013). Injection-induced earthquakes. science, 341(6142), 1225942.

Elsayed, H. S., Saad, O. M., Soliman, M. S., Chen, Y., & Youness, H. A. (2023). EQConvMixer: A deep learning approach for earthquake location from single-station waveforms. IEEE Geoscience and Remote Sensing Letters, 20, 1-5.

Foulger, G. R., Wilson, M. P., Gluyas, J. G., Julian, B. R., & Davies, R. J. (2018). Global review of human-induced earthquakes. Earth-Science Reviews, 178, 438-514.

Hidayaturrohman, Q. A., & Hanada, E. (2024). Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure. BioMedInformatics, 4(4), 2201-2212.

Islam, M. D., Li, B., Islam, K. S., Ahasan, R., Mia, M. R., & Haque, M. E. (2022). Airbnb rental price modeling based on Latent Dirichlet Allocation and MESF-XGBoost composite model. Machine Learning with Applications, 7, 100208.

Jakkampudi, S., Shen, J., Li, W., Dev, A., Zhu, T., & Martin, E. R. (2020). Footstep detection in urban seismic data with a convolutional neural network. The Leading Edge, 39(9), 654-660.

Jin, X., & Han, J. (2011). K-means clustering. Encyclopedia of machine learning, 563-564.

Kong, Q., Trugman, D. T., Ross, Z. E., Bianco, M. J., Meade, B. J., & Gerstoft, P. (2019). Machine learning in seismology: Turning data into insights. Seismological Research Letters, 90(1), 3-14.

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., & Wassermann, J. (2015). ObsPy: A bridge for seismology into the scientific Python ecosystem. Computational Science & Discovery, 8(1), 014003.

Leong, Z. X., & Zhu, T. (2024). Machine learning-assisted microearthquake location workflow for monitoring the Newberry enhanced geothermal system. Journal of Geophysical Research: Machine Learning and Computation, 1(3), e2024JH000159.

Matzel, E., Zeng, X., Thurber, C., Luo, Y., & Morency, C. (2017, February). Seismic interferometry using the dense array at the Brady geothermal field. In Proceedings of the 42nd Workshop on Geothermal Reservoir Engineering, Stanford, CA, USA (pp. 13-15).

Mousavi, S. M., & Beroza, G. C. (2022). Deep-learning seismology. Science, 377(6607), eabm4470.

ObsPy Development Team. (2024). ObsPy Documentation—Supported File Formats. Retrieved from https://docs.obspy.org

Perol, T., Gharbi, M., & Denolle, M. (2018). Convolutional neural network for earthquake detection and location. Science Advances, 4(2), e1700578.

Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications, 9(40), 651-662.

Reinisch, E. C., Cardiff, M., & Feigl, K. L. (2018). Characterizing volumetric strain at Brady Hot Springs, Nevada, USA using geodetic data, numerical models and prior information. Geophysical Journal International, 215(2), 1501-1513.

Ross, Z. E., Meier, M. A., Hauksson, E., & Heaton, T. H. (2018). Generalized seismic phase detection with deep learning. Bulletin of the Seismological Society of America, 108(5A), 2894-2901.

SAMADI, HAMIDREZA, Kimiaefar, Roohollah, & Hajian, Alireza. (2022). Fast earthquake relocation using ANFIS Neuro-Fuzzy network trained based on the double difference method.

GEOSCIENCES, 32(3 (125)), 93-102. SID. https://sid.ir/paper/1040247/en (inPersian)

Zhang, X., Zhang, J., Yuan, C., Liu, S., Chen, Z., & Li, W. (2020). Locating induced earthquakes with a network of seismic stations in Oklahoma via a deep learning method. Scientific reports, 10(1), 1941.

Zhu, W., & Beroza, G. C. (2019). PhaseNet: a deep-neural-network-based seismic arrival-time picking method. Geophysical Journal International, 216(1), 261-273.

Zuo, K., Zhao, C., & Kuang, W. (2025). SourceNet: A Deep-Learning-Based Method for Determining Earthquake Source Parameters. Bulletin of the Seismological Society of America, 115(2), 379-392.

Iranian Journal of Geophysics, Vol 20, No 1, 2026, P. 4 (Research article)
DOI: 10.30499/ijg.2025.524794.1699

Estimating the location of induced seismic sources using seismic station records and artificial intelligence

Hooman Parsa[1] and Mohammad Yaser Radan[2*]

[1] *M.Sc. in Civil Engineering, Faculty of Engineering, K. N. Toosi University of Technology, Tehran, Iran*
[2] *Assistant Professor, Faculty of Passive Defense, Malek Ashtar University of Technology, Tehran, Iran*

(Received: 21 May 2025, Accepted: 28 July 2025)

Summary

Induced seismic events triggered by human activities such as subsurface fluid extraction and injection can jeopardize the integrity of critical infrastructure. The multistage framework proposed here obviates the need for exhaustive geological models and dense seismic arrays, yet accurately and reliably estimates the regional epicenter location. To derive region-based labels for the supervised classifiers, K-means clustering was first applied to the latitude–longitude coordinates of all recorded events; the resulting cluster assignments were adopted as class labels, providing an objective, data-driven regional segmentation for subsequent training.

In the initial processing stage, three-component seismic recordings were pre-processed by applying the short-term average to long-term average ratio (STA/LTA) to identify and correct abrupt baseline offsets. The cleaned records were then paired to form cross-correlation matrices at four lags (0.5, 0.1, 0.05 and 0.01 s) capturing relative information across multiple temporal scales. Recursive feature elimination with cross-validation (RFECV) extracted the most informative subset of correlation coefficients, substantially reducing dimensionality while preserving discriminative power. These feature vectors drove a probabilistic-averaging (soft-voting) ensemble that couples a support-vector machine (SVM) with an extreme-gradient-boosting (XGBoost) classifier, combining the margin-maximizing strength of SVM with the nonlinear learning capacity of boosted decision trees.

Model development was conducted twice (first on the raw, imbalanced data and then on data balanced with the Synthetic Minority Over-sampling Technique (SMOTE)) to quantify the influence of class imbalance. Without SMOTE, decreasing the correlation-window step from 0.5 s to 0.1 s improved classification accuracy for epicentral region assignment from 0.73 to 0.90 while markedly shrinking the standard deviation of epicentral errors, indicating greater solution stability. Moving to still finer steps (0.05 s and 0.01 s) made the model increasingly sensitive to high-frequency noise, saturating accuracy gains and slightly inflating variance; the 0.1 s lag therefore emerged as an optimal trade-off between resolution and robustness.

With SMOTE, overall stability improved further and error dispersion contracted, yet a modest drop in accuracy appeared at steps coarser than 0.01 s, attributable to the limited representativeness of some synthetic samples. The best performance arose from pairing SMOTE with the 0.01 s step, achieving a classification accuracy of 0.93 in epicentral region assignment, an absolute gain of 5.7% over the non-SMOTE result.

These findings demonstrate that the proposed workflow can deliver accurate, repeatable epicentral estimates in data-limited environments, supporting real-time decision-making without the need for comprehensive subsurface models. Furthermore, where computational resources are constrained, the 0.1 s configuration without SMOTE remains a well-balanced option that combines high classification accuracy with modest processing cost.

Keywords: Seismic event source localization, cross-correlation, XGBoost, SVM

[*]Corresponding author: radan@mut.ac.ir