دوره ۱۶، شماره ۴۶، بهار ۱۴۰۵، صفحه ۳۱-۴۸

نشریه روش های تحلیلی و عددی در مهندسی معدن

تارنمای نشریه: http://anm.yazd.ac.ir/

مقاله پژوهشی

# کاربرد الگوریتم‌های نمونه‌گیری در طبقه‌بندی داده‌های ژئوشیمیایی نامتوازن: مطالعه موردی؛ داده‌های ژئوشیمیایی برگه ۱:۱۰۰۰۰۰ قاین

**حمید گرانیان[1*]**

**۱- گروه مهندسی معدن، دانشگاه صنعتی بیرجند، بیرجند، ایران**

(دریافت: ۳۰ دی ۱۴۰۳، بازنگری: ۲۸ فروردین ۱۴۰۴، پذیرش: ۰۶ خرداد ۱۴۰۴)

## چکیده

**داده‌های ژئوشیمیایی ماهیت نامتوازن (یعنی تعداد نمونه‌ها با عیار کم یا کلاس زمینه زیاد و تعداد نمونه‌ها با عیار بالا یعنی کلاس آنومالی کم) دارند. طبقه‌بندی این داده‌ها، منجر به ایجاد مدلی اریب‌دار (کم شدن احتمال تعلق نمونه‌های جدید به کلاس‌هایی با نمونه‌های کمتر) همراه با کاهش دقت و صحت مدل خواهد شد. در این مقاله، سه دسته الگوریتم نمونه‌گیری افزایشی، نمونه‌گیری کاهشی و نمونه‌گیری ترکیبی برای متوازن‌سازی داده‌ها معرفی شده است. همچنین عملکرد این الگوریتم‌ها بر روی داده‌های ژئوشیمیایی رسوبات آبراهه‌ای برگه قاین توسط دو روش طبقه‌بندی ماشین بردار پشتیبان و شبکه عصبی مصنوعی بررسی شده است. نتایج نشان می‌دهند که متوازن‌سازی داده‌ها (رساندن نسبت تعداد نمونه‌های کلاس زمینه به کلاس آنومالی از ۶٫۳ با ۱) می‌توان افزایش قابل‌توجهی در کمیت سنجه‌های ماتریس درهم‌ریختگی مثل صحت، حساسیت، وضوح، دقت، امتیاز-F، مقدار-F، میانگین-G و سطح زیر منحنی ROC (به میزان ۱۰ تا ۵۰ درصد) و کاهش حدود ۱۰ درصدی در سنجه خطا ایجاد نماید. به‌طوری‌که الگوریتم‌های نمونه‌گیری افزایشی، ترکیبی و کاهشی به ترتیب بالاترین عملکرد را دارند. همچنین نقشه‌های آنومالی‌های ژئوشیمیایی مدل‌سازی شده توسط الگوریتم‌های متوازن‌سازی در منطقه موردمطالعه نشان می‌دهد که این مدل‌ها می‌توانند ضمن افزایش وسعت آنومالی‌های ژئوشیمیایی، همپوشانی خوبی بین این آنومالی‌ها با واحدهای سنگی حاوی کانی‌سازی برقرار نمایند. در این خصوص، الگوریتم‌های نمونه‌گیری افزایشی و سپس الگوریتم نمونه‌گیری ترکیبی از عملکرد بالاتری برخوردار هستند؛ بنابراین پیشنهاد این مقاله استفاده از الگوریتم‌های متوازن‌سازی داده‌های (به‌کارگیری الگوریتم‌های نمونه‌گیری افزایشی و سپس الگوریتم‌های نمونه‌گیری ترکیبی) قبل از طبقه‌بندی داده‌های اکتشافی است.**

## کلمات کلیدی

**الگوریتم SMOTE، الگوریتم ADASYN، الگوریتم RUS، الگوریتم OSS، الگوریتم SMOTE-Tomek، الگوریتم ADASYN-CNN، برگه قاین.**

*عهده‌دار مکاتبات: h.geranian@birjandut.ac.ir

DOI: 10.22034/ANM.2025.22666.1661

## ۱- مقدمه

پردازش داده‌های اکتشافی مستلزم به‌کارگیری تکنیک‌هایی است که بتواند ساختار داده‌ها را شناسایی و مدل‌سازی نماید. طبقه‌بندی و خوشه‌بندی از الگوریتم‌های داده‌کاوی هستند که قادر به استخراج این ساختارها می‌باشند. طبقه‌بندی یک روش یادگیری ماشین نظارت‌شده است که هدف آن برآورد یک مدل از داده‌ها با چند کلاس بر اساس روابط آماری بین آن‌ها و سپس محاسبه احتمال تعلق یک نمونه جدید به کلاس‌ها مختلف است [۱]. در این روش نوع کلاس هر نمونه و تعداد کل کلاس‌های از قبل مشخص است. همچنین در این روش نیاز به تعدادی نمونه آموزشی جهت ساختن مدل (۵۰ تا ۸۰ درصد نمونه‌ها) و تعدادی نمونه آزمایشی (۲۰ تا ۵۰ درصد نمونه‌ها) برای بررسی صحت‌سنجی مدل ساخته‌شده است [۱، ۲].

روش‌های طبقه‌بندی متنوع بوده و در تحلیل داده‌های اکتشافی (خصوصاً ژئوشیمیایی) به‌طور گسترده استفاده شده است. مدل‌سازی آلودگی‌های محیط زیستی در اطراف معدن انگوران و کانی‌زایی در کانسار طلای ساری گونای توسط روش تحلیل تمایز [۳، ۴]، شناسایی الگوهای پراکندگی ژئوشیمیایی عناصر نادر خاکی در اطراف معدن گل‌گهر و طبقه‌بندی سنگ‌ها بر اساس خواص ژئوشیمیایی آن‌ها توسط روش درخت تصمیم [۵، ۶]، پتانسیل‌یابی طلا در کانسار طلای اپی‌ترمال ساری گونای و شناخت الگوهای ژئوشیمیایی و کانی‌شناسی در کانسارهای طلای هیدروترمال توسط الگوریتم تئوری بیزین [۷، ۸]، شناسایی لیتولوژی ژئوشیمیایی و تعیین رابطه‌ی کمی بین آلتراسیون‌ها و آنومالی‌های ژئوشیمیایی کانسار طلای تنورچه به کمک روش ماشین بردار پشتیبان [۹، ۱۰]، طبقه‌بندی سنگ‌ها به کمک داده‌های ژئوشیمیایی و فتوگرافی و پیش‌بینی کانی‌سازی چندفلزی همراه با طلا به کمک داده‌های ژئوشیمیایی و زمین‌شناسی توسط روش جنگل تصادفی [۱۱، ۱۲]، تعیین رابطه بین عناصر ردیاب و زون‌های متالورژی به‌منظور شناسایی محدوده‌های مستعد کانی‌سازی و ارزیابی خواص ژئوشیمیایی زغال‌سنگ توسط روش رگرسیون لجستیک [۱۳، ۱۴]، مدل‌سازی داده‌های ژئوشیمیایی به‌منظور پیش‌بینی کانی‌سازی در استرالیای غربی و گناباد توسط روش شبکه عصبی [۱۵، ۱۶]، شناسایی آنومالی‌های ژئوشیمیایی چند عنصری و تلفیق الگوریتم نظارت‌شده و نیمه نظارت‌شده به‌منظور تعیین آنومالی ژئوشیمیایی توسط روش k-نزدیک‌ترین همسایگی [۱۷، ۱۸] و تعیین نقشه پتانسیل معدنی کانی‌سازی سرب-روی با سنگ میزبان کربناته در زون ارومیه-دختر و تهیه نقشه‌های بازماند عیار طلا به کمک روش مجموعه‌ی طبقه‌بندی کننده‌های گرادیان بالا (XGboost)[1] [۱۹، ۲۰] نمونه‌هایی از این کاربردها می‌باشند.

برآورد مدلی با دقت بالا و پیش‌بینی احتمال تعلق نمونه‌های جدید به کلاس‌های مختلف با صحت بالا در کلیه‌ی روش‌های طبقه‌بندی ذکرشده، منوط به متوازن بودن نمونه‌های در داده‌های آموزشی است. متوازن بودن نمونه‌ها، به معنی برابر بودن توزیع تعداد نمونه‌ها در کلاس‌های مختلف است. درحالی‌که در عمل با داده‌های نامتوازن[2] روبر هستیم. به‌طور مثال در داده‌های ژئوشیمیایی، تعداد نمونه‌ها با عیار پایین (داده‌های کلاس زمینه) بسیار زیاد بوده در مقابل، تعداد نمونه‌ها با عیار بالا (داده‌های کلاس آنومالی) کم می‌باشند. طبقه‌بندی با داده‌های نامتوازن باعث ایجاد مدلی اریب‌دار، کاهش دقت مدل و کم شدن احتمال تعلق نمونه‌های جدید به کلاس‌هایی با نمونه‌های کمتر می‌شود [۲۱-۲۳]؛ بنابراین متوازن کردن نمونه‌ها قبل از ساخت مدل طبقه‌بندی، یک پیش‌شرط برای پردازش داده‌ها است. ازآنجاکه این نکته کمتر موردتوجه متخصصین علوم زمین قرارگرفته است، هدف این مقاله معرفی الگوریتم‌های متوازن کردن داده‌ها و بررسی تأثیر آن‌ها در کاربرد روش‌های طبقه‌بندی به‌منظور انتخاب بهترین الگوریتم است.

برای طبقه‌بندی داده‌های نامتوازن سه رویکرد روش‌های نمونه‌گیری از داده‌ها، تعدیل روش‌های طبقه‌بندی و روش‌های یادگیری هزینه-حساسیت پیشنهادشده است [۲۴، ۲۵]. در رویکرد اول با افزایش مصنوعی، یا کاهش نمونه‌ها و یا ترکیب آن‌ها، داده‌های کلاس‌ها متوازن می‌گردد [۲۴، ۲۶، ۲۷]. در روش‌های گروه دوم با استفاده از الگوریتم-های تلفیقی همچون bagging، boosting، جنگل تصادفی یا جنگل ایزوله عملکرد پیش‌بینی‌کننده را در مجموعه داده‌های نامتوازن افزایش می‌دهند [۲۸، ۲۹، ۳۰]. در حالی که در رویکرد سوم، هزینه مدل یادگیری را برای طبقه‌بندی نادرست نقاط کلاس دارای نمونه‌ی کمتر را افزایش می‌دهند تا عملکرد کل سیستم بهبود یابد [۳۱، ۳۲، ۳۳]. مطالعات

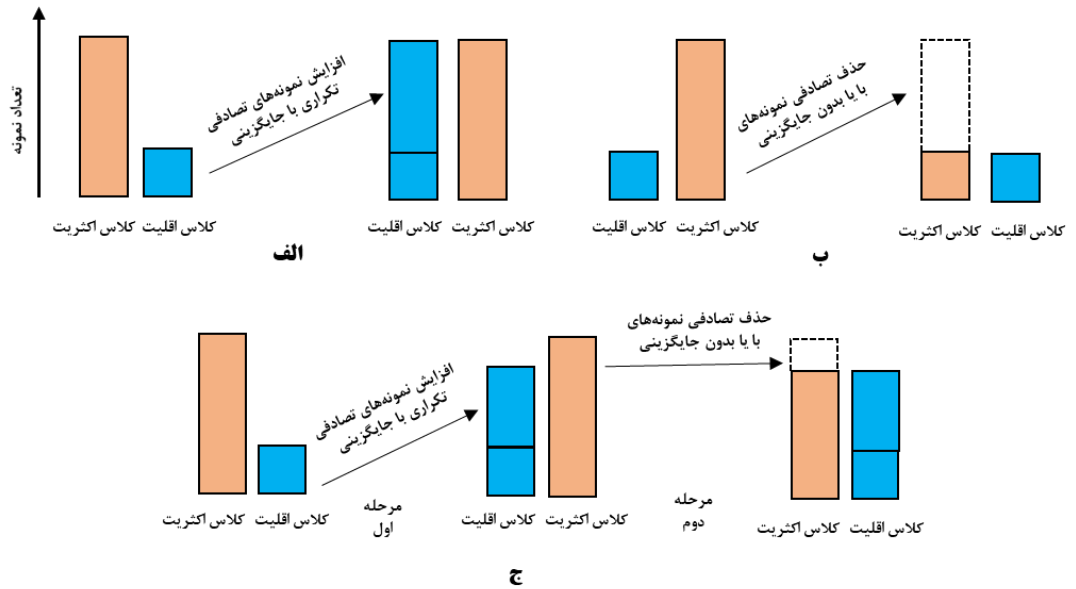
نشان داده است که الگوریتم‌های رویکرد اول به دلیل داشتن روش‌های متفاوت برای هر دسته داده‌ی نامتوازن با شرایط خاص، با داده‌های واقعی سازگارتر بوده و نتایج بهتری را به همراه داشته است [۲۵، ۲۶، ۲۷، ۳۴]. لذا در این مقاله، به معرفی الگوریتم‌های رویکرد اول پرداخته شده است. برای این منظور از داده‌های اکتشافی ورقه ۱:۱۰۰۰۰۰ قاین در استان خراسان جنوبی و روش‌های طبقه‌بندی ماشین بردار پشتیبان و شبکه عصبی مصنوعی استفاده شده است تا تأثیر متوازن‌سازی داده‌ها بر تفسیر داده‌های ژئوشیمیایی و تشخیص آنومالی‌های ژئوشیمیایی بررسی شود.

## ۲- الگوریتم‌های متوازن کردن داده‌ها

در یک مجموعه داده‌ی آموزشی دو کلاسه با $n$ نمونه، اگر تعداد نمونه‌های کلاس اول ($n_1$) بیشتر از تعداد نمونه‌های کلاس دوم ($n_2$) باشد به‌نحوی‌که $n_1 \gg n_2$ و $n=n_1+n_2$، با طبقه‌بندی یک دسته داده‌ی چولگی‌دار و نامتوازن روبر هستیم؛ بنابراین کلاس اول، کلاس اکثریت[۳]، کلاس دوم، کلاس اقلیت[۴] و به نسبت $\frac{n_1}{n_2}$، نسبت عدم توازن[۵] گفته می‌شود. نسبت عدم توازن می‌تواند از ۴:۱ تا ۱۰۰:۱ و یا حتی بیشتر تغییر نماید. با افزایش نسبت عدم توازن، میزان اریب‌دار شدن مدل طبقه‌بندی افزایش‌یافته و دقت کلاسه‌بندی کاهش می‌یابد. درنتیجه برآورد احتمال تعلق نمونه‌های جدید به کلاس اقلیت به‌شدت کاهش می‌یابد. در حالی که هنگام طبقه‌بندی یک مجموعه داده‌ی نامتوازن، کلاس اقلیت معمولاً از اهمیت بیشترین برخوردار است. این نکته بدان معنی است که مهارت یک مدل در پیش‌بینی صحیح برچسب یا احتمال نمونه‌ای برای کلاس اقلیت از نمونه‌ای برای کلاس اکثریت مهم‌تر است؛ بنابراین کاهش احتمال تعلق نمونه به کلاس اقلیت هزینه مدل‌سازی را افزایش می‌دهد [۲۱، ۳۵]. برای حل این مشکل سه راه‌حل نمونه‌گیری افزایشی[۶]، نمونه‌گیری کاهشی[۷] و روش ترکیبی ارائه شده است که شکل ۱ اصول و فرآیند کار آن‌ها را نشان می‌دهد. در ادامه مقاله، این الگوریتم‌ها معرفی شده‌اند.

### ۲-۱- الگوریتم‌های نمونه‌گیری افزایشی

در این روش با اضافه کردن تعدادی نمونه به کلاس اقلیت، تعداد نمونه‌های آن افزایش‌یافته تا برابر تعداد نمونه‌های کلاس اکثریت شود؛ بنابراین در این روش، تعداد کل نمونه‌های داده‌های آموزشی افزایش می‌یابد (شکل ۱- الف). برای اجرای این روش، الگوریتم‌های نمونه‌گیری افزایشی تصادفی (ROS)[۸]، تکنیک نمونه‌گیری افزایشی مصنوعی (SMOTE)[۹]، نمونه‌گیری افزایشی مرزی (BOS)[۱۰] و نمونه‌گیری مصنوعی تطبیقی (ADASYN)[۱۱] برای کلاس اقلیت معرفی شده‌اند. مزیت این روش کاهش میزان اریب شدن مدل و افزایش دقت آن است. در مقابل از معایب آن می‌توان به افزایش تعداد نمونه‌ها بدون در نظر گرفتن نمونه‌های کلاس اکثریت که منجر به ایجاد نمونه‌های مبهم می‌شود، بالا رفتن احتمال بیش‌برازشی و افزایش زمان پردازش اشاره نمود [۲۱، ۳۵، ۳۶].




**شکل ۱: فرآیند روش‌های نمونه‌گیری افزایشی (الف)، نمونه‌گیری کاهشی (ب) و روش ترکیبی (ج) برای متوازن کردن داده‌ها [۲۲].**

**۱-۱-۲- الگوریتم SMOTE**

الگوریتم نمونه‌گیری افزایشی مصنوعی برای کلاس اقلیت (SMOTE) یکی از پرکاربردترین رویکردها در این زمینه است که توسط Chawla و همکارانش در سال ۲۰۰۲ ارائه‌شده است [۳۷]. در این الگوریتم تعداد نمونه‌های کلاس اقلیت توسط گام‌های زیر افزایش می‌یابد [۲۱، ۳۷]:

۱- یک بردار متغیر از اختلاف بین $k$ نزدیک‌ترین نمونه‌های اطراف یکی از نمونه‌های کلاس اقلیت تشکیل می‌شود.

۲- این بردار در یک عدد تصادفی بین ۰ تا ۱ ضرب می‌شود.

۳- مقدار به‌دست‌آمده به مقدار بردار متغیر اصلی (نمونه‌ی اولیه کلاس اقلیت) اضافه‌شده تا نمونه‌ای جدید به دست آید.

این فرآیند تا رسیدن تعداد نمونه‌های کلاس اقلیت به کلاس اکثریت ادامه می‌یابد. در سال‌های اخیر مدل‌های تعدیل‌یافته بسیاری برای این الگوریتم معرفی شده است که از آن جمله می‌توان به MSMOTE، WSMOTE، SASMOTE، DSMOTE، ESMOTE و GBSMOTE و غیره اشاره کرد [۳۸-۴۲].

**۲-۱-۲- الگوریتم ADASYN**

الگوریتم نمونه‌گیری افزایشی مصنوعی تطبیقی برای کلاس اقلیت (ADASYN) یک مدل تعدیل‌یافته الگوریتم قبلی است که توسط He و همکارانش در سال ۲۰۰۸ ارائه شده است [۴۳]. در این الگوریتم نمونه‌های جدید در مناطقی از فضای ویژگی بیشتر تولید می‌شوند که چگالی نمونه‌های اقلیت کمتر است؛ بنابراین داده‌های مصنوعی بیشتری برای نمونه‌های کلاس اقلیت تولید می‌شوند که یادگیری آن‌ها در مقایسه با سایر نمونه‌ها سخت‌تر است. مراحل این الگوریتم عبارت‌اند از [۴۳، ۴۴]:

۱- محاسبه تعداد نمونه‌هایی مصنوعی که بایستی برای کلاس اقلیت تولید شود از رابطه (۱):

$$G = (n_1 - n_2) \times \beta \qquad (۱)$$

که $\beta \in \{0\,,1\}$ پارامتر تعیین‌کننده شدت توازن بوده و در حالت توازن کامل $\beta = 1$ است.

۲- پیدا کردن $k$ نزدیک‌ترین نمونه‌های اطراف هر یک از نمونه‌های کلاس اقلیت و محاسبه نسبت زیر:

$$r_i = {\Delta_i}/{k} \quad i = 1, 2, \dots, n_2 \qquad (۲)$$

که $\Delta$ تعداد نمونه‌های نزدیک هر نمونه در کلاس اقلیت است و $r_i \in \{0\,,1\}$.

۳- نرمالیزه کردن نسبت فوق توسط رابطه (۳):

$$\hat{r}_i = {r_i}/{\sum_{i=1}^{n_2} r_i} \quad i = 1, 2, \dots, n_2 \qquad (۳)$$

که $\hat{r}_i$ توزیع چگالی است زیرا $\sum \hat{r}_i = 1$.

۴- محاسبه تعداد نمونه‌های مصنوعی که بایستی برای هر نمونه‌ی کلاس اقلیت تولید شود، توسط رابطه (۴):

$$g_i = \hat{r}_i \times G \qquad (۴)$$

۵- برای هر نمونه‌ی کلاس اقلیت $g_i$ نمونه همانند الگوریتم SMOTE تولید می‌شود.

در سال‌های اخیر برای این الگوریتم نیز مدل‌های تعدیل‌یافته‌ای از قبیل ADASYN-N، ADASYN-KNN و ADASYN-LOF پیشنهاد شده است [۴۵، ۴۶]

## ۲-۲- الگوریتم‌های نمونه‌گیری کاهشی

در این روش با حذف کردن تعدادی از نمونه‌های کلاس اکثریت، تعداد آن‌ها کاهش یافته تا برابر تعداد نمونه‌های کلاس اقلیت گردد (شکل ۱-ب). برای کم کردن تعداد نمونه‌ها می‌توان از راه‌حل گروه‌بندی کردن نمونه‌ها یا حذف آن‌ها و یا ترکیب این دو استفاده نمود. برای اجرای این روش، الگوریتم‌های نمونه‌گیری کاهشی تصادفی (RUS)[۱۲]، قانون نزدیک‌ترین همسایگی متراکم شده (CNN)[۱۳]، نمونه‌گیری کاهشی نزدیک گم شده (NMUS)[۱۴]، نمونه‌گیری کاهشی پیوندهای Tomek (TLUS)[۱۵]، قانون نزدیک‌ترین همسایگی اصلاح‌شده (ENN)[۱۶]، انتخاب یک‌طرفه (OSS)[۱۷] و قانون تمیزکردن همسایگی (NCR)[۱۸] ارائه شده است. کاهش اریب‌دار شدن مدل طبقه‌بندی برای کلاس اقلیت و کاهش زمان پردازش از مزایایی روش نمونه‌گیری کاهشی است. همچنین از داست دادن اطلاعات، اریب‌دار شدن نمونه‌های کلاس اکثریت و خطای تجزیه‌وتحلیل از معایب این روش محسوب می‌شود [۲۱، ۳۴].

**۱-۲-۲- الگوریتم RUS**

الگوریتم نمونه‌گیری کاهشی تصادفی (RUS) شامل انتخاب تصادفی نمونه‌هایی از کلاس اکثریت برای حذف آن‌ها از مجموعه‌ی داده آموزشی است. این فرآیند را می‌توان تا زمان توزیع متوازن کلاس‌های موردنظر (مانند تعداد

نمونه‌های مساوی برای هر کلاس) تکرار کرد. برای اجرای این الگوریتم می‌توان به هر نمونه یک شماره نسبت داد و با انتخاب تصادفی شماره‌ی نمونه‌ها، آن‌ها را از کلاس اکثریت حذف نمود.

این رویکرد ممکن است برای آن دسته از مجموعه داده‌هایی مناسب‌تر باشد که در آن تعداد کافی نمونه برای ساختن مدل طبقه‌بندی با دقت قابل‌قبول وجود داشته باشد. حذف احتمالی نمونه‌هایی از کلاس اکثریت که ممکن است برای تطبیق و ساختن مرز تصمیم‌گیری در مدل طبقه‌بندی مهم و یا حیاتی باشند؛ از محدودیت‌های کاربرد این الگوریتم محسوب می‌شود. در این الگوریتم هیچ راهی برای شناسایی یا حفظ نمونه‌های خوب یا غنی از اطلاعات در کلاس اکثریت وجود ندارد [۲۱، ۲۲].

**۲-۲-۲- الگوریتم OSS**

الگوریتم انتخاب یک‌طرفه (OSS) توسط Kubat و Matwin در سال ۱۹۹۷ ارائه‌شده است [۵۱]. در این الگوریتم با حذف نمونه‌های نویز، خط-مرزی و اضافی و حفظ نمونه‌های ایمن، تعداد نمونه‌های کلاس اکثریت کاهش می‌یابد. الگوریتم آن شامل گام‌های زیر است [۵۱، ۵۲]:

۱- مجموعه داده‌های آموزشی اولیه $S$ در نظر گرفته می‌شود.

۲- زیرمجموعه‌ای از داده‌های آموزشی تحت عنوان $C$ ساخته می‌شود که شامل تمامی نمونه‌های کلاس اقلیت و یک نمونه‌ی تصادفی از کلاس اکثریت باشد (یعنی $C \subset S$).

۳- نمونه‌های مجموعه $S$ توسط قانون ۱-نزدیک‌ترین همسایگی (1-NN) با استفاده از نمونه‌های زیرمجموعه $C$، طبقه‌بندی می‌شود. سپس کلاس نمونه‌های به‌دست‌آمده با کلاس اولیه مقایسه شده و تمامی نمونه‌های دارای طبقه‌بندی نادرست به زیرمجموعه $C$ منتقل می‌گردد. در این حالت، زیرمجموعه $C$ حاصل نسبت به مجموعه $S$ سازگار و کوچک‌تر خواهد بود (یعنی نمونه‌های اضافی کلاس اکثریت از آن حذف شده است).

۴- از زیرمجموعه حاصل از گام ۳، نمونه‌های نویز و خط-مرزی کلاس اکثریت توسط الگوریتم پیوندهای Tomek حذف می‌گردد. زیرمجموعه به‌دست‌آمده متوازن خواهد بود.

اگر دو نمونه‌ی $x$ و $y$ از دو کلاس متفاوت باشند و فاصله‌ی آن‌ها را با d(x,y) نشان دهیم، زوج نمونه‌های (x,y) دارای پیوند Tomek خواهند بود، اگر هیچ نمونه‌ای همچون $z \in S$ وجود نداشته باشد که $d(x,z) < d(x,y)$ و یا $d(y,z) < d(x,y)$. نمونه‌های دارای پیوند Tomek نمونه‌های نویز یا روی خط-مرزی هستند.

## ۳-۲- الگوریتم‌های ترکیبی

در این روش از ترکیب هم‌زمان روش‌های نمونه‌گیری افزایشی و کاهشی استفاده می‌شود. به‌طوری‌که ابتدا با روش نمونه‌گیری افزایشی تعداد نمونه‌های کلاس اقلیت افزایش‌یافته و سپس با روش نمونه‌گیری کاهشی تعداد نمونه‌های کلاس اکثریت کاهش می‌یابد تا توازن بین نمونه‌های کلاس‌ها ایجاد گردد (شکل ۱-ج). هر یک از الگوریتم‌های ذکرشده در بخش‌های قبلی می‌تواند برای دو روش نمونه‌گیری افزایشی و کاهشی استفاده شود؛ بنابراین تعداد الگوریتم‌های این روش بسیار زیاد خواهد بود. در این روش از اصلاح الگوریتم‌های طبقه‌بندی نیز می‌توان استفاده کرد. نمونه‌گیری ترکیبی مزایای نمونه‌گیری افزایشی و کاهشی را باهم ترکیب می‌کند و امکان متعادل کردن مجموعه داده‌ها را فراهم می‌کند. درعین‌حال از دست دادن اطلاعات و خطرات بیش‌ازحد برازش را نیز کاهش می‌دهد. در مقابل، این روش پیچیدگی بالایی دارد و نیازمند یافتن تعادل بین دو روش نمونه‌گیری است که باعث افزایش سختی طراحی و بهینه‌سازی می‌شود. همچنین ابرپارامترهای جدیدی در این روش معرفی می‌گردند که نیاز به تنظیم پارامتر و انتخاب مدل بیشتری دارد؛ بنابراین هزینه آزمایش و محاسبات را افزایش می‌دهد. علاوه بر این، ممکن است به دلیل وابستگی به توزیع داده‌ها و ویژگی‌های متغیرها، سازگاری محدودی با مجموعه داده‌های مختلف داشته باشد [۲۱، ۴۷].

**۱-۳-۲- الگوریتم SMOTE-Tomek**

این الگوریتم اولین بار توسط Batista و همکارانش در سال ۲۰۰۳ معرفی شده است [۴۳]. این روش ترکیبی از توانایی الگوریتم SMOTE برای تولید داده‌های مصنوعی برای کلاس اقلیت و توانایی الگوریتم پیوندهای Tomek برای حذف داده‌هایی است که به‌عنوان پیوندهای Tomek از کلاس اکثریت شناسایی می‌شوند (نمونه‌هایی از داده‌ها از کلاس اکثریت که نزدیک‌ترین فاصله را با داده‌های کلاس اقلیت دارند). مراحل این دو الگوریتم قبلاً شرح داده‌شده است.

**۲-۳-۲- الگوریتم ADASYN-CNN**

الگوریتم ADASYN-CNN که در این مقاله معرفی می‌شود، ترکیبی از نمونه‌گیری افزایشی برای کلاس اقلیت، توسط الگوریتم ADASYN و سپس نمونه‌گیری کاهشی برای کلاس اکثریت، توسط الگوریتم CNN است. فرآیند الگوریتم ADASYN در بخش قبلی به آن اشاره شد. در الگوریتم CNN که اولین بار توسط Hart در سال ۱۹۶۸ ارائه‌شده است به دنبال یافتن یک زیرمجموعه سازگار (زیرمجموعه‌ای که تمام نمونه‌های آن توسط قانون نزدیک‌ترین همسایگی به‌درستی طبقه‌بندی شوند) از داده‌های اولیه است [۵۴]. این کار توسط مراحل زیر صورت می‌گیرد [۵۴،۴۷]:

۱- اولین نمونه در زیرمجموعه $S$ قرار می‌گیرد.

۲- نمونه دوم توسط قانون نزدیک‌ترین همسایگی با استفاده از زیرمجموعه $S$ طبقه‌بندی می‌شود. اگر نمونه دوم به‌درستی طبقه‌بندی شود، در زیرمجموعه $G$ قرار می‌گیرد؛ در غیر این صورت در زیرمجموعه $S$ قرار داده می‌شود.

۳- گام دوم برای کلیه نمونه‌ها صورت می‌گیرد. سپس این فرآیند برای زیرمجموعه $G$ نیز انجام می‌شود تا در یکی از دو حالت زیر الگوریتم متوقف گردد.

الف) زیرمجموعه $G$ تمام شود؛ یعنی همه‌ی اعضای آن به $S$ منتقل شده‌اند (در این صورت، زیرمجموعه سازگار شامل کل نمونه‌های مجموعه اصلی است) و یا

ب) یک دور کامل بر روی زیرمجموعه $G$ بدون هیچ‌گونه انتقال به زیرمجموعه $S$ انجام شود.

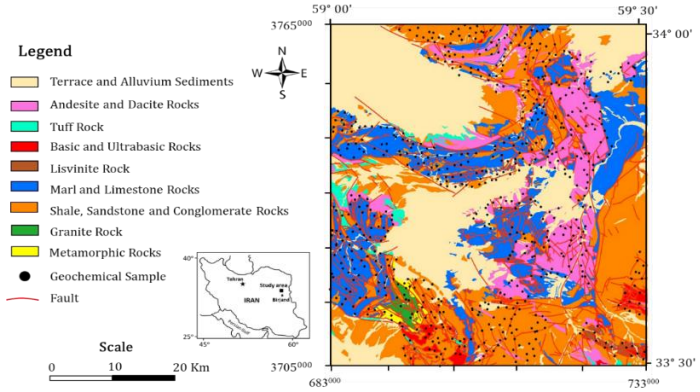
۴- زیرمجموعه نهایی $S$ به‌عنوان زیرمجموعه سازگار در نظر گرفته شده و زیرمجموعه $G$ کنار گذاشته می‌شود.

## ۳- زمین‌شناسی و داده‌های منطقه موردمطالعه

برگه ۱:۱۰۰،۰۰۰ قاین در استان خراسان جنوبی و در فاصله حدود ۱۰۰ کیلومتری شمال شهر بیرجند قرار دارد (شکل ۲). این منطقه به لحاظ زون‌های زمین‌شناسی ساختاری ایران جزو بلوک لوت محسوب می‌شود. بلوک لوت یکی از خرد قاره‌های ایران مرکزی است که فعالیت‌های گسل‌های امتداد لغز دو طرف بلوک لوت به همراه فرورانش بلوک افغان به زیر بلوک لوت باعث تشکیل سنگ‌های کالک-آلکالن در بخش‌های وسیعی از مرکزی و شمالی این بلوک شده است. به‌طوری‌که بر روی پی‌سنگ‌های دگرگونی پرکامبرین و ژوراسیک پسین، سنگ‌های ولکانیکی، ولکانوکلاستیک، ساب ولکانیکی و توده‌های نفوذی از کرتاسه پسین تا کواترنره قابل‌مشاهده است [۴۸].

شکل ۲ نقشه زمین‌شناسی این برگه را نشان می‌دهد. قدیم‌ترین واحدهای سنگی منطقه سنگ‌های دگرگونی گنایس، شیست و کوارتزیت و توده گرانیتی با سن پروتوزوئیک هستند که در بخش جنوب غربی برگه رخنمون دارند. سنگ‌های رسوبی شیل، ماسه‌سنگ و آهک ژوراسیک بیشترین رخنمون را در برگه دارند که در بخش‌های غربی و شمالی برگه دیده می‌شوند. سایر واحدهای سنگی رسوبی از قبیل کنگلومرا، شیل و ماسه‌سنگ در شمال و جنوب برگه به همراه بخشی از توف‌ها منطقه متعلق به کرتاسه بالایی هستند. واحدهای سنگی رسوبی به همراه توف‌های در شرق برگه نیز متعلق به دور پالئوسن می‌باشند. واحد سنگی لیستونیتی قرارگرفته در منتهی‌الیه بخش جنوب شرقی برگه، دارای سن کرتاسه میانی تا بالایی است. سنگ‌های بازی و فوق‌بازی برگه از دو بخش تشکیل شده است. واحدهای رخنمون‌دار در بخش جنوب شرقی برگه اولترابازیک‌های کرتاسه هستند. درحالی‌که واحد رخنمون دار در بخش جنوبی تا جنوب غربی برگه، سنگ‌های بازیک پلیستوسن می‌باشند. سنگ‌های آذرین درونی و بیرونی حد واسط در برگه در دو بخش سنگ‌های آندزیتی پورفیری پالئوسن-ائوسن که بیشترین رخنمون را دارد و سنگ‌های آندزیتی-بازالتی و آندریتی-دیوریتی الیگوسن-میوسن قابل‌مشاهده است. حدود نیمه از برگه را نیز رسوبات آبرفتی و پادگانه به همراه رسوبات دشت‌های سیلابی تشکیل می‌دهد. از برگه قاین ۶۵۲ نمونه از رسوبات آبراهه‌ای برداشت شده است. نمونه‌ها برای ۶۳ عنصر به روش ICP-OES تجزیه شده‌اند (نمونه‌برداری و تجزیه شیمیایی توسط کارشناسان سازمان زمین‌شناسی و اکتشافات معدنی کشور صورت گرفته است). شکل ۲ موقعیت و پراکندگی این نمونه‌ها را نشان می‌دهد. در این مقاله، پس از حذف عناصر دارای مقادیر سنسورد، از نتایج تجزیه شیمیایی ۲۷ عنصر باقیمانده استفاده شده است که جدول ۱ پارامترهای آمار توصیفی آن‌ها را نشان می‌دهد. مقادیر چولگی و کشیدگی عناصر نشان می‌دهند که تابع توزیع عناصر Al، Fe و تا حتی Si از توزیع نرمال، سایر عناصر اصلی نزدیک به نرمال و بقیه عناصر از توزیع لاگ‌نرمال تبعیت می‌کنند. مقایسه میانگین عناصر

در جدول ۱ با عدد کلارک نیز نشان می‌دهد که نمونه‌ها برای عناصر Al، Ba، Co، Cu، Fe، K، Mg، Mn، Na، Ni، P، Si، Ti و V تهی‌شدگی و برای عناصر As، B، Bi، Ca، Cr، La، Li، Pb، Sb، Sr، Y، Zn و Zr غنی‌شدگی دارند. همچنین نسبت انحراف معیار به میانگین برای کلیه عناصر کمتر از ۱ است؛ بنابراین می‌توان استنباط نمود که عناصر در مقادیر بزرگ‌تر ازآنچه برای آن‌ها انتظار بوده است تمرکزیافته و ازاین‌رو احتمال وجود آنومالی در داده‌ها وجود دارد [۵۵].




**شکل ۲: نقشه ساده‌شده زمین‌شناسی ورقه ۱:۱۰۰٫۰۰۰ قاین به همراه موقعیت نمونه‌های ژئوشیمیایی (سازمان زمین‌شناسی کشور)**

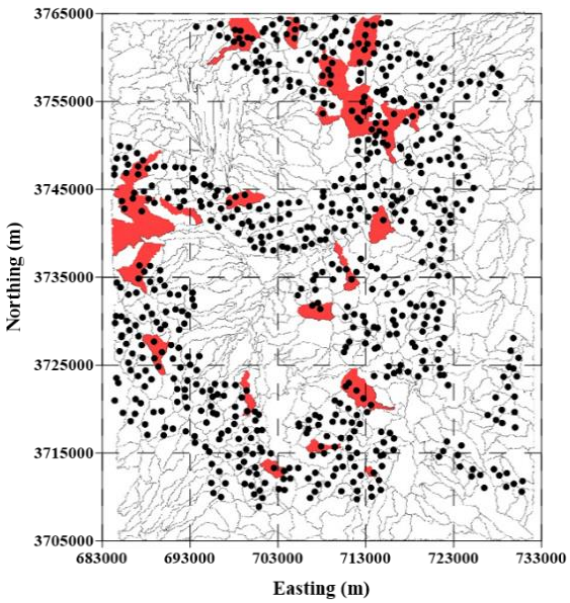
**جدول ۱: پارامترهای آمار توصیفی داده‌های ژئوشیمیایی ورقه قاین**

| متغیر (واحد) | میانگین | انحراف معیار | کمینه | میانه | بیشینه | چولگی | کشیدگی |
|---|---|---|---|---|---|---|---|
| *Al (%)* | ۳٫۷۷ | ۰٫۴۵ | ۲٫۵۹ | ۳٫۷۶ | ۶٫۱۹ | ۰٫۷۱ | ۱٫۸۶ |
| *As (ppm)* | ۱۰٫۵۵ | ۸٫۲۱ | ۱۰٫۰۰ | ۱۰٫۰۰ | ۱۷۴٫۰۰ | ۱۷٫۹۷ | ۳۳۱٫۰۵ |
| *Ba (ppm)* | ۳۸۸٫۰۳ | ۷۳٫۳۱ | ۲۴۴٫۰۰ | ۳۸۲٫۰۰ | ۹۷۴٫۰۰ | ۲٫۷۲ | ۱۵٫۴۲ |
| *B (ppm)* | ۸۲٫۴۵ | ۲۴٫۵۵ | ۳۱٫۰۰ | ۷۸٫۰۰ | ۲۷۵٫۰۰ | ۲٫۶۹ | ۱۳٫۰۷ |
| *Bi (ppm)* | ۵٫۰۸ | ۰٫۴۲ | ۵٫۰۰ | ۵٫۰۰ | ۹٫۰۰ | ۶٫۱۹ | ۴۲٫۶۴ |
| *Ca (%)* | ۱۷٫۰۴ | ۳٫۴۸ | ۶٫۰۰ | ۱۶٫۸۶ | ۳۰٫۵۷ | ۰٫۳۵ | ۰٫۸۴ |
| *Co (ppm)* | ۱۴٫۶۲ | ۴٫۴۷ | ۵٫۰۰ | ۱۴٫۰۰ | ۴۰٫۰۰ | ۱٫۵۱ | ۶٫۱۷ |
| *Cr (ppm)* | ۱۳۹٫۷۱ | ۵۵٫۰۳ | ۶۶٫۰۰ | ۱۳۰٫۰۰ | ۶۰۸٫۰۰ | ۴٫۴۴ | ۲۵٫۹۶ |
| *Cu (ppm)* | ۲۴٫۹۵ | ۶٫۱۳ | ۱۲٫۰۰ | ۲۴٫۰۰ | ۷۳٫۰۰ | ۲٫۱۴ | ۹٫۹۶ |
| *Fe (%)* | ۳٫۱۶ | ۰٫۵۱ | ۱٫۸۹ | ۳٫۱۱ | ۵٫۹۴ | ۰٫۸۰ | ۱٫۹۰ |
| *K (%)* | ۱٫۸۰ | ۰٫۳۰ | ۰٫۹۱ | ۱٫۸۳ | ۲٫۷۴ | ۰٫۰۳ | −۰٫۱۸ |
| *La (ppm)* | ۱۱۶٫۶۶ | ۲۴٫۲۷ | ۱۰۰٫۰۰ | ۱۰۳٫۰۰ | ۲۳۶٫۰۰ | ۱٫۷۶ | ۲٫۹۶ |
| *Li (ppm)* | ۵۰٫۵۶ | ۳٫۲۰ | ۵۰٫۰۰ | ۵۰٫۰۰ | ۸۴٫۰۰ | ۷٫۳۳ | ۵۹٫۲۴ |
| *Mg (%)* | ۱٫۹۷ | ۰٫۶۰ | ۱٫۲۷ | ۱٫۸۱ | ۷٫۶۰ | ۴٫۳۶ | ۲۷٫۲۴ |
| *Mn (%)* | ۰٫۰۹ | ۰٫۰۱ | ۰٫۰۷ | ۰٫۰۹ | ۰٫۲۲ | ۲٫۵۳ | ۱۷٫۳۰ |
| *Na (%)* | ۱٫۴۶ | ۰٫۳۲ | ۰٫۵۲ | ۱٫۴۱ | ۲٫۸۲ | ۰٫۴۸ | ۰٫۵۱ |
| *Ni (ppm)* | ۷۲٫۸۸ | ۵۶٫۱۸ | ۲۵٫۰۰ | ۶۳٫۰۰ | ۶۴۱٫۰۰ | ۵٫۹۸ | ۴۱٫۰۸ |
| *P (%)* | ۰٫۰۶ | ۰٫۰۱ | ۰٫۰۲ | ۰٫۰۵ | ۰٫۱۰ | ۰٫۶۲ | ۰٫۹۱ |
| *Pb (ppm)* | ۳۱٫۳۶ | ۵٫۵۹ | ۱۰٫۰۰ | ۳۱٫۰۰ | ۶۲٫۰۰ | ۰٫۲۹ | ۱٫۷۴ |
| *Sb (ppm)* | ۳۰٫۹۹ | ۳٫۰۳ | ۳۰٫۰۰ | ۳۰٫۰۰ | ۷۰٫۰۰ | ۵٫۸۹ | ۵۲٫۵۷ |
| *Si (%)* | ۲۶٫۶۵ | ۱٫۶۸ | ۲۰٫۰۷ | ۲۶٫۷۹ | ۳۲٫۵۳ | −۰٫۴۰ | ۱٫۰۲ |
| *Sr (ppm)* | ۵۷۹٫۸۷ | ۱۵۰٫۹۴ | ۳۵۹٫۰۰ | ۵۴۸٫۵۰ | ۱۸۷۳٫۰۰ | ۳٫۶۷ | ۲۱٫۳۵ |
| *Ti (%)* | ۰٫۴۵ | ۰٫۰۷ | ۰٫۲۳ | ۰٫۴۵ | ۰٫۶۵ | ۰٫۲۰ | −۰٫۰۵ |
| *V (ppm)* | ۱۰۰٫۸۲ | ۱۶٫۶۱ | ۴۵٫۰۰ | ۱۰۳٫۰۰ | ۱۶۰٫۰۰ | −۰٫۱۶ | ۰٫۴۸ |
| *Y (ppm)* | ۵۳٫۰۶ | ۹٫۹۸ | ۵۰٫۰۰ | ۵۰٫۰۰ | ۱۸۵٫۰۰ | ۶٫۱۹ | ۵۷٫۹۷ |
| *Zn (ppm)* | ۶۹٫۸۶ | ۲۷٫۶۱ | ۲۵٫۰۰ | ۶۵٫۵۰ | ۲۲۷٫۰۰ | ۱٫۶۶ | ۴٫۶۹ |
| *Zr (ppm)* | ۲۱۱٫۸۵ | ۵۲٫۴۷ | ۹۹٫۰۰ | ۲۰۳٫۰۰ | ۴۹۶٫۰۰ | ۱٫۱۵ | ۲٫۵۵ |

## ۴- پردازش داده‌های ژئوشیمیایی

با توجه به ماهیت ترکیبی یا بسته بودن داده‌های ژئوشیمیایی (داشتن مجموع مقادیر ثابت ۱ یا ۱۰۰ درصد)، از روش‌های انتقال نسبت لگاریتمی (نسبت لگاریتمی افزایشی: alr، نسبت لگاریتمی میان مرکز: clr و نسبت لگاریتمی ایزومتریک: ilr) می‌توان برای تبدیل آن‌ها از سیستم بسته به باز استفاده کرد. در این مقاله به دلیل حفظ بعد داده‌ها، از روش clr استفاده شده است. در مرحله بعدی و برای کلاسه‌بندی نمونه‌های ژئوشیمیایی، روش آنومالی ژئوشیمیایی مرکب جمعی بکار رفته است. برای این منظور ابتدا داده‌های استانداردشده (انتقال داده‌ها به یک بازه مشخص مثلاً صفر تا یک)، عیار جمعی محاسبه و سپس داده‌های جامعه آنومالی از زمینه به روش فرکتال عیار-مساحت از هم تفکیک گردیده است [۴۹، ۵۰]. مطابق شکل ۳ که در آن موقعیت آنومالی‌های مرکب جمعی بر اساس حوزه آبریز هر نمونه به نمایش گذاشته شده است، از ۶۵۲ نمونه، ۸۹ نمونه در محدوده‌ی آنومالی‌ها قرار دارند؛ بنابراین ۵۶۳ نمونه متعلق به جامعه زمینه (کلاس اکثریت) و ۸۹ نمونه متعلق به جامعه آنومالی (کلاس اقلیت) خواهد بود. درنتیجه نسبت عدم توازن داده‌ها برابر ۶٫۳ است.

برای طبقه‌بندی یا کلاسه‌بندی، داده‌ها به‌صورت تصادفی به دو بخش داده‌های آموزشی (حدود ۸۰ درصد داده‌ها؛ یعنی ۴۵۰ نمونه‌ی کلاس اکثریت بعلاوه ۷۱ نمونه‌ی کلاس اقلیت و مجموعاً ۵۲۱ نمونه) و داده‌های آزمایشی (حدود ۲۰ درصد داده‌ها؛ یعنی ۱۱۳ نمونه‌ی کلاس اکثریت بعلاوه ۱۸ نمونه‌ی کلاس اقلیت و مجموعاً ۱۳۱ نمونه) تفکیک شده است. طبقه‌بندی بر روی داده‌های به روش ماشین بردار پشتیبان[۱۹] (SVM) و شبکه عصبی مصنوعی[۲۰] (ANN) به‌عنوان دو روش شناخته‌شده و دارای اعتبار بالا انجام شده است. جدول ۲ نتایج طبقه‌بندی داده‌های اولیه (داده‌های نامتوازن) را با دو روش یادشده نشان می‌دهد که به کمک نرم‌افزار Statistica 14 صورت گرفته است. در روش SVM از تابع کرنل RBF با مقادیر $\gamma = 0.04$ و $c = 7$ استفاده شده است. بهینه‌سازی پارامترهای طبقه‌بندی نیز به روش 10-Fold Cross-Validation صورت گرفته است. مطابق داده‌های جدول ۲، صحت کلی طبقه‌بندی داده‌های آموزشی ۸۸٫۹ درصد و داده‌های آزمایشی ۸۰٫۹ درصد است.




**شکل ۳: موقعیت آنومالی‌های ژئوشیمیایی مرکب جمعی (نواحی قرمز) و نمونه‌های ژئوشیمیایی در ورقه قاین**

جدول ۲ نتایج طبقه‌بندی داده‌های نامتوازن با شبکه عصبی مصنوعی و با استفاده از الگوریتم MLP را نشان می‌دهد. در این روش از دو لایه پنهان با ۲۱ نورون، تابع فعال‌سازی لایه پنهان از نوع Gaussian، تابع فعال‌سازی لایه خروجی Softmax و تابع خطا از نوع آنتروپی برای مدل کردن داده‌ها استفاده شده است. مقدار صحت کلی طبقه‌بندی برای داده‌های آموزشی و آزمایشی به ترتیب برابر ۸۹٫۱ و ۸۵٫۵ درصد و برای هر کلاس نیز در جدول آمده است.

از داده‌های جدول ۲ دو نکته را می‌توان نتیجه گرفت. نکته اول اینکه، صحت طبقه‌بندی داده‌های کلاس زمینه در روش SVM نسبت به روش ANN کمی بالاتر است. در مقابل صحت طبقه‌بندی داده‌های کلاس آنومالی در روش ANN کمی بیشتر از روش SVM است؛ بنابراین می‌توان گفت که روش ANN نسبت به روش SVM به داده‌های نامتوازن کمی مقاوم‌تر است. نکته دوم اینکه، علیرغم بالا بودن متوسط صحت طبقه‌بندی در هر دو روش (حدود ۸۷ درصد)، صحت طبقه‌بندی داده‌های کلاس اقلیت خیلی پایین است؛ بنابراین برای داشتن طبقه‌بندی با صحت قابل‌قبول و کاهش احتمال برآورد نمونه‌های کلاس اقلیت، نیاز به متوازن کردن داده‌های آموزشی خواهد بود.

**جدول ۲: نتایج طبقه‌بندی داده‌های نامتوازن به دو روش SVM و ANN**

| | | کلاس‌های پیش‌بینی‌شده | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | روش *SVM* | | | | روش *ANN* | | | |
| | | داده‌های آموزشی | | داده‌های آزمایشی | | داده‌های آموزشی | | داده‌های آزمایشی | |
| | | زمینه | آنومالی | زمینه | آنومالی | زمینه | آنومالی | زمینه | آنومالی |
| کلاس‌های واقعی | زمینه | ۴۲۸ | ۲۲ | ۹۸ | ۱۵ | ۴۲۶ | ۲۴ | ۱۰۲ | ۱۱ |
| | آنومالی | ۳۶ | ۳۵ | ۱۰ | ۸ | ۳۳ | ۳۸ | ۸ | ۱۰ |
| | صحت (٪) | ۹۵٫۱ | ۴۹٫۳ | ۸۶٫۷ | ۴۴٫۴ | ۹۴٫۷ | ۵۳٫۵ | ۹۰٫۳ | ۵۵٫۵ |

در ادامه و برای متوازن کردن داده‌ها، از شش الگوریتم شرح داده‌شده در بخش ۲ مقاله استفاده‌شده است. این محاسبات در نرم‌افزار MATLAB و نرم‌افزار داده‌کاوی Keel صورت گرفته است. تعداد نمونه‌های الگوریتم SMOTE و ADASYN برابر ۹۰۰ عدد (۴۵۰ نمونه کلاس زمینه و ۴۵۰ نمونه کلاس آنومالی)، تعداد نمونه‌های الگوریتم RUS و OSS برابر ۱۴۲ عدد (۷۱ نمونه کلاس زمینه و ۷۱ نمونه کلاس آنومالی) و تعداد نمونه‌های الگوریتم SMOTE-Tomek و ADASYN-CNN برابر ۷۲۰ عدد (۳۶۰ نمونه کلاس زمینه و ۳۶۰ نمونه کلاس آنومالی) محاسبه و انتخاب گردیده است. مدل‌سازی این داده‌های متوازن به دو روش SVM و ANN انجام‌شده و مدل به‌دست‌آمده برای طبقه‌بندی مجدد داده‌های آزمایشی بکار رفته است که جدول ۳ نتایج آن‌ها را نشان می‌دهد. در روش طبقه‌بندی SVM بالاترین صحت کلی با الگوریتم SMOTE و برابر ۹۰٫۸ درصد و کمترین صحت کلی با الگوریتم RUS و برابر ۷۷٫۱ درصد به‌دست‌آمده است. در حالی که در روش طبقه‌بندی ANN بیشترین و کمترین صحت کلی به ترتیب با الگوریتم‌های ADASYN و RUS و برابر ۹۷٫۷ و ۸۴٫۷ درصد برآورد شده است. جدول ۳ صحت طبقه‌بندی هر کلاس و برای هر الگوریتم را به تفکیک نشان می‌دهد.

داده‌های جدول ۳ نشان می‌دهند که اولاً، روش طبقه‌بندی ANN از صحت کلی بالاتری نسبت به روش SVM برخوردار است. ثانیاً، الگوریتم‌های نمونه‌گیری افزایشی بالاترین صحت کلی طبقه‌بندی را دارند و الگوریتم‌های نمونه‌گیری ترکیبی و الگوریتم‌های نمونه‌گیری کاهشی به ترتیب در جایگاه‌های بعدی قرار دارند. ثالثاً، اختلاف معنی‌داری که بین صحت طبقه‌بندی کلاس‌های اقلیت و اکثریت در داده‌های نامتوازن وجود داشت (داده‌های جدول ۲)، در داده‌های متوازن وجود ندارد (داده‌های جدول ۳) و حتی در بعضی الگوریتم، صحت کلاس‌بندی داده‌های آزمایشی کلاس اقلیت از کلاس اکثریت بیشتر نیز شده است.

**جدول ۳: نتایج طبقه‌بندی داده‌های آزمایشی با مدل‌های متوازن به دو روش SVM و ANN**

| | | کلاس‌های پیش‌بینی‌شده | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| نوع داده | | *SMOTE* | | | | *ADASYN* | | | |
| | | روش *SVM* | | روش *ANN* | | روش *SVM* | | روش *ANN* | |
| | | زمینه | آنومالی | زمینه | آنومالی | زمینه | آنومالی | زمینه | آنومالی |
| کلاس‌های واقعی | زمینه | ۱۰۲ | ۱۱ | ۱۱۱ | ۲ | ۱۰۰ | ۱۳ | ۱۱۲ | ۱ |
| | آنومالی | ۱ | ۱۷ | ۲ | ۱۶ | ۱ | ۱۷ | ۳ | ۱۶ |
| | صحت (٪) | ۹۰٫۳ | ۹۴٫۴ | ۹۸٫۲ | ۸۹٫۹ | ۸۸٫۵ | ۹۴٫۴ | ۹۹٫۱ | ۸۹٫۹ |
| نوع داده | | *RUS* | | | | *OSS* | | | |
| کلاس‌های واقعی | زمینه | ۸۸ | ۲۵ | ۹۵ | ۱۸ | ۸۹ | ۲۴ | ۹۹ | ۱۴ |
| | آنومالی | ۵ | ۱۳ | ۲ | ۱۶ | ۵ | ۱۳ | ۲ | ۱۶ |
| | صحت (٪) | ۷۷٫۹ | ۷۲٫۲ | ۸۴٫۱ | ۸۹٫۹ | ۷۸٫۸ | ۷۲٫۲ | ۸۷٫۶ | ۸۹٫۹ |
| نوع داده | | *SMOTE-Tomek* | | | | *ADASYN-CNN* | | | |
| کلاس‌های واقعی | زمینه | ۹۳ | ۲۰ | ۱۰۷ | ۶ | ۱۰۱ | ۱۲ | ۱۱۰ | ۳ |
| | آنومالی | ۱ | ۱۷ | ۱ | ۱۷ | ۱ | ۱۷ | ۲ | ۱۶ |
| | صحت (٪) | ۸۲٫۳ | ۹۴٫۴ | ۹۴٫۷ | ۹۴٫۴ | ۹۸٫۳ | ۹۴٫۴ | ۹۷٫۳ | ۸۹٫۹ |

## ۵- مقایسه الگوریتم‌های متوازن‌سازی

تحلیل داده‌های ژئوشیمیایی ورقه قاین نشان داد که نتایج طبقه‌بندی داده‌های نامتوازن با داده‌های متوازن کاملاً متفاوت است. همچنین نتایج الگوریتم‌های متوازن‌سازی داده‌ها نیز کمی با یکدیگر تفاوت دارند. در ادامه و برای مقایسه نتایج از دو ابزار سنجه‌های ماتریس درهم‌ریختگی و بررسی نقشه مرتبط با داده‌های اکتشافی منطقه موردمطالعه استفاده شده است.

### ۵-۱- سنجه‌های ماتریس درهم‌ریختگی

استفاده از ماتریس درهم‌ریختگی (ماتریس داده‌های جدول ۲ و ۳) و سنجه‌هایی که بر اساس داده‌های این ماتریس تعریف می‌شوند، یکی از روش‌های ارزشیابی کارایی مدل‌های طبقه‌بندی است. اگر TP و FN به ترتیب تعداد نمونه‌های جامعه آنومالی درست و نادرست طبقه‌بندی‌شده باشد، FP و TN نیز به ترتیب تعداد نمونه‌های جامعه زمینه نادرست و درست طبقه‌بندی‌شده، P تعداد نمونه‌های جامعه آنومالی و N تعداد نمونه‌های جامعه زمینه در نظر گرفته شود، سنجه‌های کمی ارزیابی مدل طبقه‌بندی را می‌توان مطابق جدول ۴ تعریف نمود.

**جدول ۴: سنجه‌های کمی ارزیابی مدل‌های طبقه‌بندی [۳۹،۲۳]**

| سنجه | پارامتر | فرمول |
|---|---|---|
| صحت | *AC* | $\frac{TP+TN}{P+N}$ |
| خطا | *ER* | $\frac{FP+FN}{P+N}$ |
| حساسیت | *S or R* | $\frac{TP}{P}$ |
| وضوح | *SP* | $\frac{TN}{N}$ |
| دقت | *P* | $\frac{TP}{TP+FP}$ |
| امتیاز-*F* | *F-Score* | $\frac{2\times P\times S}{P+S}$ |
| مقدار-*F* | *F-Value* | $\frac{(1+\beta^2)\times P\times S}{\beta^2\times P+S}$ |
| میانگین-*G* | *G-Mean* | $\sqrt{S\times SP}$ |
| سطح زیر منحنی | *AUC* | $\frac{1}{2}(1+\frac{TP}{TP+FN}-\frac{FP}{FP+TN})$ |

جدول ۵ مقادیر سنجه‌های دو ماتریس درهم‌ریختگی داده‌های نامتوازن و متوازن (جدول‌های ۲ و ۳) را برای داده‌های آزمایشی نشان می‌دهد. مقادیر صحت (AC) و خطا (ER) نشان می‌دهند که متوازن کردن داده‌ها، حدود ۱۰ درصد صحت را افزایش و همین مقدار خطا را کاهش داده است (به‌استثناء الگوریتم‌های نمونه‌گیری کاهشی). همچنین روش طبقه‌بندی ANN از صحت بالاتر و خطای کمتر نسبت به روش طبقه‌بندی SVM برخوردار است. به‌طوری‌که بالاترین صحت و کمترین خطا متعلق به روش طبقه‌بندی ANN با الگوریتم متوازن‌سازی ADASYN است.

مقادیر حساسیت (S) و وضوح (SP) در جدول ۵ به ترتیب نشان‌دهنده‌ی صحت طبقه‌بندی کلاس زمینه و آنومالی است. داده‌های حساسیت این جدول نشان می‌دهد اگرچه متوازن‌سازی در بعضی مواقع صحت طبقه‌بندی کلاس زمینه را افزایش داده است ولی این افزایش معنی‌دار نیست. در مقابل مقادیر وضوح نشان‌دهنده‌ی افزایش قابل‌ملاحظه‌ای در طبقه‌بندی داده‌های متوازن کلاس آنومالی است. به‌طوری‌که این افزایش تا حدود ۵۰ درصد نیز رسیده است. بیشترین مقدار صحت طبقه‌بندی کلاس زمینه (S) متعلق به روش طبقه‌بندی ANN با الگوریتم متوازن‌سازی ADASYN و برابر ۰٫۹۹۱ است و بالاترین مقدار صحت طبقه‌بندی کلاس آنومالی (SP) برابر ۰٫۹۴۴ است که در چندین الگوریتم متوازن‌سازی و هر دو روش طبقه‌بندی خصوصاً روش SVM به‌دست‌آمده است. مقادیر دقت (P) در جدول ۵ نشان‌دهنده‌ی نسبت تعداد نمونه‌های کلاس زمینه است که به‌درستی طبقه‌بندی شده‌اند. داده‌های این سنجه نشان می‌دهد که متوازن‌سازی توانسته است، دقت طبقه‌بندی را بین ۹-۶ درصد افزایش دهد. بالاترین دقت برابر ۰٫۹۹۱ و به روش طبقه‌بندی ANN با الگوریتم متوازن‌سازی SMOTE-Tomek به‌دست‌آمده است. دو سنجه F-Score و F-Value نیز ترکیب یا میانگین هارمونیک دو سنجه دقت و حساسیت (یا Recall) را نشان می‌دهند. این سنجه‌ها تمامیت طبقه‌بندی درست داده‌های کلاس زمینه هستند. داده‌های جدول ۵ نشان‌دهنده‌ی افزایش این دو سنجه با متوازن‌سازی داده‌ها است. در این دو سنجه نیز همانند سنجه‌های قبلی تأثیر الگوریتم‌های نمونه‌گیری کاهشی کمتر است. به‌طوری‌که بالاترین مقدار توسط الگوریتم ADASYN-CNN به‌دست‌آمده است.

**جدول ۵: مقادیر سنجه‌های ماتریس درهم‌ریختگی داده‌های آزمایشی**

| روش | | نوع داده | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | نامتوازن | متوازن | | | | | |
| | | | *SMOTE* | *ADASYN* | *RUS* | *OSS* | *SMOTE-Tomek* | *ADASYN-CNN* |
| *SVM* | *AC* | ۰٫۸۰۹ | ۰٫۹۰۸ | ۰٫۸۹۳ | ۰٫۷۷۱ | ۰٫۷۷۹ | ۰٫۸۳۹ | ۰٫۹۰۱ |
| | *ER* | ۰٫۱۹۱ | ۰٫۰۹۲ | ۰٫۱۰۷ | ۰٫۲۲۹ | ۰٫۲۲۱ | ۰٫۱۶۱ | ۰٫۰۹۹ |
| | *S* | ۰٫۸۶۷ | ۰٫۹۰۳ | ۰٫۸۸۵ | ۰٫۸۸۵ | ۰٫۷۷۹ | ۰٫۸۲۳ | ۰٫۹۸۳ |
| | *SP* | ۰٫۴۴۴ | ۰٫۹۴۴ | ۰٫۹۴۴ | ۰٫۹۴۴ | ۰٫۷۲۲ | ۰٫۹۴۴ | ۰٫۹۴۴ |
| | *P* | ۰٫۹۰۷ | ۰٫۹۹۰ | ۰٫۹۹۰ | ۰٫۹۴۶ | ۰٫۹۴۷ | ۰٫۹۸۹ | ۰٫۹۹۰ |
| | *F-Score* | ۰٫۸۸۷ | ۰٫۹۴۵ | ۰٫۹۳۵ | ۰٫۹۱۴ | ۰٫۸۵۵ | ۰٫۸۹۸ | ۰٫۹۸۶ |
| | *F-Value* | ۰٫۸۹۹ | ۰٫۹۷۱ | ۰٫۹۶۷ | ۰٫۹۳۳ | ۰٫۹۰۸ | ۰٫۹۵۱ | ۰٫۹۸۹ |
| | *G-Mean* | ۰٫۶۲۰ | ۰٫۹۲۳ | ۰٫۹۱۴ | ۰٫۹۱۴ | ۰٫۷۵۰ | ۰٫۹۳۴ | ۰٫۹۶۳ |
| | *AUC* | ۰٫۶۵۶ | ۰٫۹۲۳ | ۰٫۹۱۵ | ۰٫۷۵۰ | ۰٫۷۵۵ | ۰٫۸۸۴ | ۰٫۹۱۹ |
| *ANN* | *AC* | ۰٫۸۵۵ | ۰٫۹۶۹ | ۰٫۹۷۷ | ۰٫۸۴۷ | ۰٫۸۷۸ | ۰٫۹۴۷ | ۰٫۹۶۲ |
| | *ER* | ۰٫۱۴۵ | ۰٫۰۳۱ | ۰٫۰۲۳ | ۰٫۱۵۳ | ۰٫۱۲۲ | ۰٫۰۵۳ | ۰٫۰۳۸ |
| | *S* | ۰٫۹۰۳ | ۰٫۹۸۲ | ۰٫۹۹۱ | ۰٫۸۴۱ | ۰٫۸۷۶ | ۰٫۹۴۷ | ۰٫۹۷۳ |
| | *SP* | ۰٫۵۵۵ | ۰٫۸۹۹ | ۰٫۸۹۹ | ۰٫۸۹۹ | ۰٫۸۹۹ | ۰٫۹۴۴ | ۰٫۸۹۹ |
| | *P* | ۰٫۹۲۷ | ۰٫۹۸۲ | ۰٫۹۷۴ | ۰٫۹۷۹ | ۰٫۹۸۰ | ۰٫۹۹۱ | ۰٫۹۸۲ |
| | *F-Score* | ۰٫۹۱۵ | ۰٫۹۸۲ | ۰٫۹۸۲ | ۰٫۹۰۵ | ۰٫۹۲۵ | ۰٫۹۶۹ | ۰٫۹۷۷ |
| | *F-Value* | ۰٫۹۲۲ | ۰٫۹۸۲ | ۰٫۹۷۷ | ۰٫۹۴۸ | ۰٫۹۵۷ | ۰٫۹۸۲ | ۰٫۹۸۱ |
| | *G-Mean* | ۰٫۷۰۸ | ۰٫۹۳۹ | ۰٫۹۴۳ | ۰٫۸۷۰ | ۰٫۸۸۷ | ۰٫۹۴۵ | ۰٫۹۳۵ |
| | *AUC* | ۰٫۷۲۹ | ۰٫۹۳۶ | ۰٫۹۱۲ | ۰٫۸۶۵ | ۰٫۸۸۲ | ۰٫۹۴۶ | ۰٫۹۳۱ |

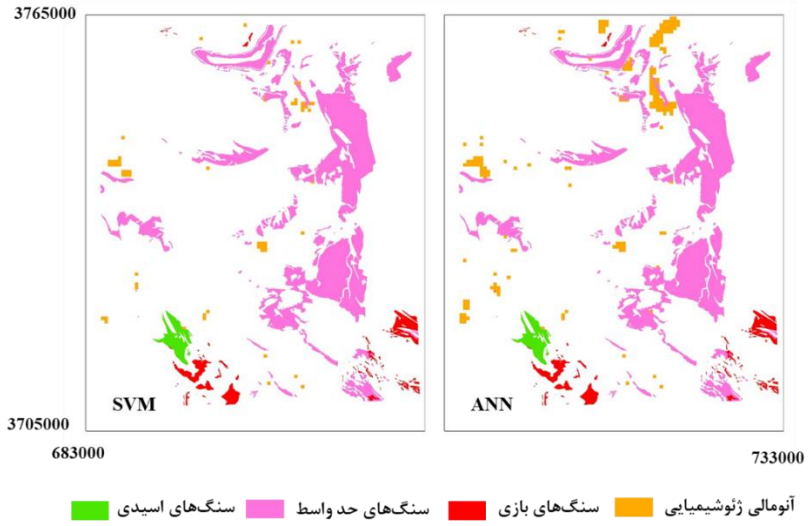
سنجه G-Mean نشان‌دهنده‌ی دقت کلی طبقه‌بندی هم‌زمان کلاس‌های زمینه و آنومالی است و یک ارزیابی متعادلی از عملکرد روش طبقه‌بندی را ارائه می‌دهد. داده‌های جدول ۵ نشان می‌دهند که یک افزایش حدود ۳۰ درصد در کمیت این سنجه در داده‌های متوازن شده نسبت به داده‌های نامتوازن به وجود آمده است. بالاترین مقدار این سنجه برابر ۰٫۹۶۳ است که در روش طبقه‌بندی SVM با الگوریتم ADASYN-CNN به‌دست‌آمده است.

سنجه AUC سطح زیر منحنی ROC[۲۱] را نشان می‌دهد که در بهترین حالت مقدار آن می‌توان برابر ۱ باشد. داده‌های جدول ۵ نشان‌دهنده‌ی افزایش قابل‌ملاحظه‌ای (حدود ۲۰ درصدی) در کمیت این سنجه با متوازن کردن داده‌ها است. بیشترین مقدار این سنجه با روش طبقه‌بندی ANN و الگوریتم SMOTE-Tomek به‌دست‌آمده است که برابر ۰٫۹۴۶ است.

سنجه‌های ماتریس درهم‌ریختگی در بازه [۱، ۰] تغییر می‌کنند. در حالت ایده‌آل و بهترین وضعیت، مقدار آن‌ها برای کلیه سنجه‌ها برابر یک و برای سنجه خطا برابر صفر خواهد بود؛ بنابراین از داده‌های جدول ۵ می‌توان نتیجه گرفت که متوازن‌سازی داده‌ها توانسته است مقدار کلیه سنجه‌ها را افزایش و مقدار سنجه خطا را کاهش دهد. همچنین سه نکته کلی به‌دست‌آمده از نتایج جدول ۳ برای داده‌های این جدول نیز برقرار است. درمجموع می‌توان الگوریتم ADASYN-CNN در روش طبقه‌بندی SVM و الگوریتم SMOTE در روش طبقه‌بندی ANN را به‌عنوان بهترین روش (به دلیل داشتن بالاترین مقدار مجموع سنجه‌ها) پیشنهاد نمود.

## ۲-۵- نقشه‌های مرتبط با داده‌های اکتشافی

در ادامه از مدل‌های به‌دست‌آمده از داده‌های متوازن برای تهیه نقشه‌ی آنومالی‌های ژئوشیمیایی مرکب در منطقه موردمطالعه استفاده شده است. برای این منظور عیار هر یک از عناصر در سلول‌های ۵۰۰×۵۰۰ متری به کمک روش عکس مجذور فاصله برآورد گردید. تعداد کل سلول‌ها در برگه قاین، ۱۲۲۲۱ عدد است و ماتریس داده‌ها ۲۷×۱۲۲۲۱ خواهد بود. در ابتدا نقشه‌ی آنومالی‌های ژئوشیمیایی منطقه موردمطالعه به کمک داده‌های نامتوازن برآورد گردید که شکل ۴ نقشه‌ی این آنومالی‌ها را به دو روش طبقه‌بندی SVM و ANN نشان می‌دهد.

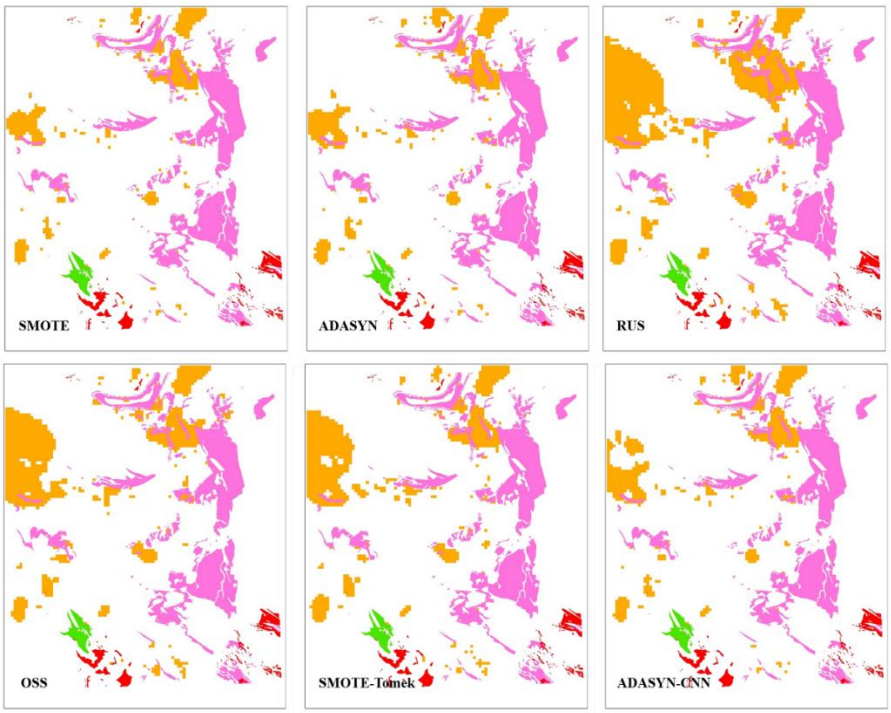



**شکل ۴: نقشه‌ی آنومالی‌های ژئوشیمیایی مرکب برگه قاین برآورده شده با داده‌های نامتوازن به همراه موقعیت واحدهای سنگی آذرین**

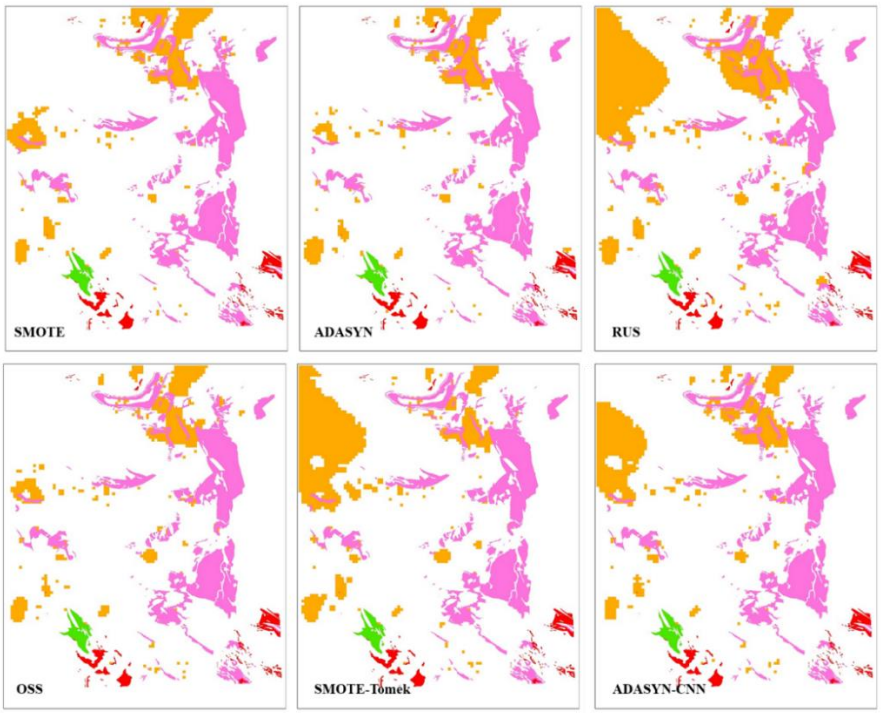
مساحت محدوده‌ی آنومالی‌ها در نقشه SVM-۴ حدود ۱۶ کیلومترمربع و برای شکل ANN-۴ نیز ۵۰٫۲۵ $Km^2$ است. شکل ۴ نشان می‌دهد که موقعیت آنومالی‌های ژئوشیمیایی برآورد شده، همپوشانی کمی با واحدهای سنگی آذرین در منطقه موردمطالعه دارد. مساحت کم آنومالی‌ها می‌تواند یکی از این دلایل این مسئله باشد. هرچند در بخش شمالی برگه (خصوصاً در شکل ANN-۴)، بخشی از آنومالی‌ها با سنگ‌های آذرین حد واسط همپوشانی نشان می‌دهند.

شکل ۵ نقشه‌های آنومالی‌های ژئوشیمیایی در برگه قاین را نشان می‌دهد که به روش طبقه‌بندی SVM و با کمک مدل‌های تهیه‌شده توسط شش الگوریتم متوازن‌سازی داده‌ها حاصل شده است. وسعت آنومالی‌ها در این نقشه‌ها نسبت به داده‌های نامتوازن (شکل ۴) افزایش قابل‌ملاحظه‌ای یافته است. به‌طوری بیشترین وسعت آنومالی‌ها متعلق به الگوریتم RUS با ۳۵۳٫۲۵ $Km^2$ و کمترین وسعت متعلق به الگوریتم SMOTE با ۱۳۵ $Km^2$ است. وسعت آنومالی‌ها در سایر الگوریتم‌ها ADASYN، OSS، SMOTE-Tomek و ADASYN-CNN نیز به ترتیب ۱۵۹٫۲۵، ۲۸۲٫۷۵، ۲۵۰٫۷۵ و ۱۶۸٫۷۵ کیلومترمربع است. افزایش زیاد وسعت آنومالی‌های ژئوشیمیایی در سه شکل RUS-۵، OSS-۵ و SMOTE-Tomek-۵ بیشتر متعلق به آنومالی شمال-غربی برگه است که همپوشانی کمی با سنگ‌های آذرین دارد و بیشتر در داخل سنگ‌های رسوبی قرار دارد. در مقابل، آنومالی‌های شمالی برگه در شکل‌های SMOTE-۵، ADASYN-۵ و ADASYN-CNN-۵ همپوشانی بالای با سنگ‌های آذرین حد واسط نشان می‌دهند. اگرچه افزایش وسعت آنومالی به‌دست‌آمده می‌تواند احتمال کشف کانی‌سازی را افزایش می‌دهد؛ ولی در مقابل می‌تواند باعث افزایش هزینه‌های اکتشافی در فازهای بعدی نیز گردد. لذا تأیید این نکته نیازمند مطالعات و برداشت‌های صحرایی است. درمجموع، آنومالی‌های شکل‌های SMOTE-۵ و ADASYN-CNN-۵ بیشتر درصد همپوشانی را با سنگ‌های آذرین در برگه موردمطالعه دارند.

شکل ۶ طبقه‌بندی داده‌های کل برگه قاین را توسط مدل‌های به‌دست‌آمده از داده‌های متوازن و به روش ANN نشان می‌دهد. وسعت آنومالی‌های ژئوشیمیایی به‌دست‌آمده توسط الگوریتم‌های SMOTE، ADASYN، RUS، OSS، SMOTE-Tomek و ADASYN-CNN به ترتیب ۱۷۲٫۷۵، ۱۴۶٫۲۵، ۳۸۴٫۲۵، ۱۵۷، ۳۱۲٫۲۵ و ۲۴۵٫۲۵ کیلومترمربع هستند. بیشترین وسعت آنومالی‌ها در این روش نیز همانند روش SVM توسط الگوریتم متوازن‌سازی RUS به‌دست‌آمده که در بخش شمال غربی برگه خود را نشان می‌دهند و با سنگ‌های آذرین همپوشانی ندارد. این نتیجه با داده‌های جدول ۵ مطابقت دارد. بالاترین درصد همپوشانی آنومالی‌ها با واحدهای سنگی آذرین متعلق به شکل‌های SMOTE-۶ و ADASYN-۶ است؛ بنابراین، به لحاظ تطابق موقعیت آنومالی‌های شکل‌های ۵ و ۶ با شکل ۳ و همپوشانی با سنگ‌های آذرین می‌توان الگوریتم‌های نمونه‌گیری افزایشی (SMOTE و ADASYN) و سپس الگوریتم‌های ترکیبی (ADASYN-CNN) را معرفی نمود.




**شکل ۵: نقشه‌ی آنومالی‌های ژئوشیمیایی مرکب برآورده شده با داده‌های متوازن به روش طبقه‌بندی SVM (راهنما همانند شکل ۴)**




**شکل ۶: نقشه‌ی آنومالی‌های ژئوشیمیایی مرکب برآورده شده با داده‌های متوازن به روش طبقه‌بندی ANN (راهنما همانند شکل ۴)**

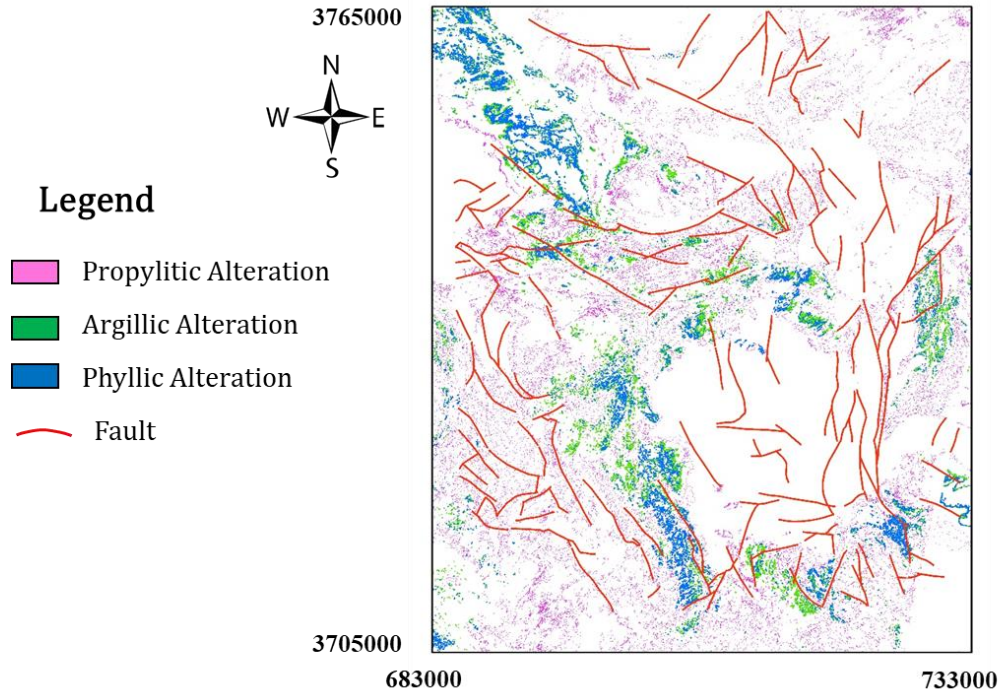
در ادامه، ارتباط آنومالی‌های ژئوشیمیایی با آلتراسیون‌ها و گسل‌های اصلی موردبررسی قرار گرفته است که می‌تواند به صحت آنومالی‌ها کمک نماید. شکل ۷ پراکندگی آلتراسیون‌های هیدروترمال و گسل‌های اصلی را در ورقه قاین نشان می‌دهد. آلتراسیون‌ها به کمک روش نسبت‌گیری باندی بر روی تصویر آستر منطقه موردمطالعه به دست آمده است. برای این منظور ابتدا تصحیح اتمسفری و هندسی بر روی تصویر صورت گرفته و سپس از نسبت‌های باندی ۸/(۹+۷) برای تعیین کانی‌های اپیدوت، کلریت و کربنات‌های مرتبط با آلتراسیون پروپیلیتیک، نسبت باندی ۵/(۶+۴) برای

تعیین کانی‌های کائولینیت و مونت‌موریلونیت مرتبط با آلتراسیون آرژیلیک و نسبت باندی ۶/(۷+۵) برای تعیین کانی‌های مسکویت و ایلیت مرتبط با آلتراسیون فیلیک استفاده شده است [۵۶، ۵۷].

آلتراسیون‌های فیلیک بخش میانی و جنوبی ورقه قاین، نشان داده‌شده در شکل ۷، در مرز بین سنگ‌های آذرین و واحدهای رسوبی قرار دارند. ازآنجاکه این آلتراسیون در نزدیکی توده کانی‌سازی به وجود می‌آید، بنابراین احتمال کانی‌سازی در آن‌ها نسبت به سایر بخش‌های ورقه می‌تواند بیشتر باشد. آلتراسیون‌های آرژیلیک و فیلیک در کل ورقه همپوشانی بالای با هم دارند. وجود این آلتراسیون‌ها در بخش شمال غربی ورقه و همپوشانی آن‌ها با رسوبات آواری همان طور که قبلاً نیز ذکر شد، نمی‌تواند با کانی‌سازی مرتبط باشد؛ بنابراین روش‌های متوازن‌سازی که آنومالی‌های ژئوشیمیایی بزرگی در این بخش از ورقه نشان می‌دهند (همانند الگوریتم‌های نمونه‌گیری کاهشی)، نمی‌تواند از اعتبار برخوردار باشد.

در مقابل، آلتراسیون پروپیلیتیک در کل ورقه قاین پراکنده بوده و همپوشانی خوب با واحدهای رسوبی بخصوص سنگ‌های آهکی نشان می‌دهد. مقایسه شکل ۷ با شکل‌های ۵ و ۶، نشان‌دهنده‌ی همپوشانی آلتراسیون‌های پروپیلیتیک با آنومالی‌های ژئوشیمیایی در بخش شمال‌شرقی دارد. بهترین همپوشانی آنومالی‌های ژئوشیمیایی متعلق به الگوریتم‌های نمونه‌گیری افزایشی (SMOTE و ADASYN) و سپس الگوریتم‌های ترکیبی (از قبیل ADASYN-CNN) است.

گسل‌های اصلی ورقه قاین در شکل ۷ را می‌توان ازلحاظ راستا به سه بخش تقسیم نمود. گسل‌های شمالی-جنوبی که بیشتر در بخش شرقی و میانی ورقه قرار دارند و با آلتراسیون‌ها و آنومالی‌های ژئوشیمیایی ارتباط ندارند. گسل‌های شمال‌غربی- جنوب‌شرقی در بخش جنوب‌غربی ورقه می‌توانند با آنومالی‌های ژئوشیمیایی دارای همپوشانی با سنگ‌های اسیدی و بازی مرتبط باشند که در کلیه الگوریتم‌های شکل‌های ۵ و ۶ قابل‌مشاهده است. در حالی که گسل‌های شمال‌شرقی- جنوب‌غربی بیشتر در بخش شمال‌شرقی ورقه قرار دارند و با آنومالی‌های ژئوشیمیایی الگوریتم‌های نمونه‌گیری افزایشی و ترکیبی همپوشانی دارند. البته تأیید مطالب ذکر نشده نیازمند مطالعات صحرایی است.




شکل ۷: پراکندگی آلتراسیون‌های هیدروترمال و گسل‌های اصلی در ورقه قاین

## ۶- نتیجه‌گیری

داده‌های اکتشافی (خصوصاً داده‌های ژئوشیمیایی) ماهیت نامتوازن دارند. طبقه‌بندی با داده‌های نامتوازن باعث ایجاد مدلی اریب‌دار، کاهش دقت مدل و کم شدن احتمال تعلق نمونه‌های جدید به کلاس‌های اقلیت (که از حساسیت بالاتری برخوردار است) می‌شود. در این مقاله، سه دسته الگوریتم نمونه‌گیری افزایشی (SMOTE و ADASYN)، نمونه‌گیری کاهشی (RUS و OSS) و نمونه‌گیری ترکیبی (SMOTE-Tomek و ADASYN-CNN) برای متوازن‌سازی داده‌ها معرفی گردید و کاربرد آن‌ها توسط دو روش طبقه‌بندی SVM و ANN بر روی داده‌های ژئوشیمیایی رسوبات آبراهه‌ای برگه قاین بررسی شد. نتایج نشان داد که روش طبقه‌بندی ANN به دلیل استفاده از لایه‌های پنهان و نورون‌های زیاد از کارآیی بالاتری نسبت به روش SVM برخوردار است. نتایج این مقاله نشان داد که متوازن‌سازی داده‌ها می‌توان افزایش قابل‌ملاحظه‌ای در کلیه سنجه‌های ماتریس درهم‌ریختگی و کاهش قابل‌توجه در سنجه خطا ایجاد نماید (کمیت سنجه‌های ماتریس درهم‌ریختگی مثل صحت، حساسیت، وضوح، دقت، امتیاز-F، مقدار-F، میانگین-G و سطح زیر منحنی به میزان ۱۰ تا ۵۰ درصد و کاهش حدود ۱۰ درصدی در سنجه خطا). به‌طوری‌که الگوریتم‌های نمونه‌گیری افزایشی بالاترین مقدار سنجه‌های ماتریس درهم‌ریختگی را دارند و الگوریتم‌های نمونه‌گیری ترکیبی و الگوریتم‌های نمونه‌گیری کاهشی به ترتیب در جایگاه‌های بعدی قرار دارند. همچنین نقشه‌های آنومالی‌های ژئوشیمیایی مدل‌سازی شده توسط الگوریتم‌های متوازن‌سازی در برگه قاین نشان دادند که این مدل‌ها می‌توانند ضمن افزایش وسعت آنومالی‌ها، همپوشانی خوبی بین آنومالی‌ها با واحدهای سنگی حاوی کانی‌سازی برقرار نمایند. نقشه پراکندگی آلتراسیون‌های هیدروترمال (پروپیلیتیک، آرژیلیک و فیلیک) و گسل‌های اصلی در ورقه قاین نیز نشان می‌دهد که آنومالی‌های شمال‌شرقی ورقه در اولویت اول احتمال کانی‌سازی و آنومالی‌های جنوب‌غربی در اولویت دوم قرار دارند. مقایسه آنومالی‌های ژئوشیمیایی به‌دست‌آمده از الگوریتم‌های متوازن‌سازی با شواهد زمین‌شناسی منطقه موردمطالعه نشان می‌دهد که نقشه‌های پراکندگی آنومالی‌های ژئوشیمیایی به‌دست‌آمده از الگوریتم‌های نمونه‌گیری افزایشی (SMOTE و ADASYN) و سپس الگوریتم ترکیبی (ADASYN-CNN) در این زمینه از عملکرد بالاتری برخوردار بودند؛ بنابراین متوازن‌سازی داده‌ها قبل از طبقه‌بندی پیشنهاد این مقاله است که می‌تواند باعث افزایش دقت، صحت و کارایی مدل طبقه‌بندی شود. همچنین استفاده از الگوریتم‌های نمونه‌گیری افزایشی و سپس الگوریتم‌های نمونه‌گیری ترکیبی پیشنهاد دیگر این مقاله است.

## سپاسگزاری

از سازمان زمین‌شناسی و اکتشافات معدنی کشور به خاطر استفاده از داده‌های اکتشافی منطقه موردمطالعه و از آقایان دکتر مصطفی سبزه‌کار و دکتر اسماعیل هداوندی به خاطر راهنمایی‌های ارزشمندشان در استفاده از نرم‌افزار Keel تشکر و قدردانی می‌گردد.

## مراجع

[1] Zaki, M.J. and Meira, W. (2020). Data Mining and Machine Learning: Fundamental Concepts and Algorithms, Cambridge University Press, New York, 777 P.

[2] Cerulli, G. (2023). Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata, Springer Cham, 391 P.

[3] Moradzadeh, A., Zare, M., Kamkar Rouhani, A. and Doulati Aredehjan, F. (2019). Classification of environmental geochemical data using discriminant analysis and neural network in carbonate-sulfide waste dumps of lead and zinc mines. Journal of Mining Engineering 14(44): 12-25 [In Persian].

[4] Geranian, H., Tabatabaei, S.H., Asadi, H.H. and Carranza, E.J.M. (2016). Application of discriminant analysis and support vector machine in mapping gold potential areas for further drilling in the Sari-Gunay gold deposit, NW Iran. Nat. Resour. Res. 25: 145–159.

[5] Zaremotlagh, S. and Hezarkhani, A. (2017). The use of decision tree induction and artificial neural networks for recognizing the geochemical distribution patterns of LREE in the Choghart deposit, Central Iran. Journal of African Earth Sciences 128: 37-46.

[6] Degtyareva. K., Kukartseva, O., Tynchenko, V., Mariupolskiy, T. and Pereverzev, D. (2024). Analysis of geochemical characteristics of rocks

[17] Chen, Y., Zhao, Q. and Lu, L. (2022). Combining the outputs of various k-nearest neighbor anomaly detectors to form a robust ensemble model for high-dimensional geochemical anomaly detection. Journal of Geochemical Exploration 231(1):106875.

[18] Chen, Y. and Lu, L. (2023). The Anomaly Detector, Semi-supervised Classifier, and Supervised Classifier Based on K-Nearest Neighbors in Geochemical Anomaly Detection: A Comparative Study. Math. Geosci. 55: 1011–1033.

[19] Parsa, M. (2021). A data augmentation approach to XGboost-based mineral potential mapping: An example of carbonate-hosted Zn-Pb mineral systems of Western Iran. Journal of Geochemical Exploration 228: 106811.

[20] Ibrahim, B., Majeed, F., Ewusi, A. and Ahenkorah, I. (2022). Residual geochemical gold grade prediction using extreme gradient boosting. Environmental Challenges 6: 100421.

[21] Brownlee, J. (2021). Imbalanced Classification with Python, Machine Learning Mastery, 463 P.

[22] Wongvorachan, T., He, S. and Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information 14: 54.

[23] Han, J., Kamber, M. and Pei, J. (2022). Data mining: concepts and techniques, 4th Edition, Morgan Kaufmann, 752 P.

[24] Kashyap, J., and Gulati, P. (2020). Hybrid Resampling Technique to Tackle the Imbalanced Classification Problem. 10.21203/rs.3.rs-36578/v1.

[25] Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., and Japkowicz, N. (2024). The class imbalance problem in deep learning. Machine Learning 113: 4845–4901.

[26] Khushi, M., Shaukat, K., Mahboob Alam, T., Hameed, I.A., Uddin, S., and Luo, S. (2021). A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access 9: 109960-109975.

[27] Altalhan, M., Algarni, A., and Turki-Hadj Alouane, M. (2025). Imbalanced Data Problem in Machine Learning: A Review. IEEE Access 13: 13686-13699.

[28] Liu, L., Wu, X., Li, S., Tan, S., and Bai, Y. (2022). Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. BMC Medical Informatics and Decision Making volume 22: Article number: 82.

using machine learning methods. E3S Web of Conferences 583, 01007.

[7] Geranian, H., Tabatabaei, S.H. and Asadi, H.H. (2013). Application of classifiers based on Bayes decision theory in gold potential mapping in Sari Gunay epithermal gold deposit. Geochemistry Journal 1(4): 347-355 [In Persian].

[8] Ziaii, M., Abedi, A. and Ziaei, M. (2009). Geochemical and mineralogical pattern recognition and modeling with a Bayesian approach to hydrothermal gold deposits. Applied Geochemistry 24(6): 1142-1146.

[9] Yin, S., Lin, X., Huang, Y., Zhang, Z. and Li, X. (2023). Application of improved support vector machine in geochemical lithology identification. Earth. Sci. Inform. 16: 205–220.

[10] Mahdiyanfar, H., Mohammadpoor, M. and Mahdavi, M. (2022). Determination of alteration genesis and quantitative relationship between alteration and geochemical anomaly using support vector machines. International Journal of Mining and Geo-Engineering 56(1): 33-391.

[11] Trott, M., Leybourne, M., Hall, L. and Layton-Matthews, D. (2022). Random forest rock type classification with integration of geochemical and photographic data. Applied Computing and Geosciences 15: 100090.

[12] Zhang, Y., Ye, X., Xie, S., Dong, J., Yaisamut, O., Zhou, X. and Zhou, X. (2023). Prediction of Au-Polymetallic Deposits Based on Spatial Multi-Layer Information Fusion by Random Forest Model in the Central Kunlun Area of Xinjiang, China. Minerals 13(10): 1302.

[13] Chen, Y. and Zhao, Q., (2021). Mineral exploration targeting by combination of recursive indicator elimination with the ℓ2-regularization logistic regression based on geochemical data. Ore Geology Reviews 135: 104213.

[14] Hanson, D.R. and Lawson, H.E. (2023). Using Machine Learning to Evaluate Coal Geochemical Data with Respect to Dynamic Failures. Minerals 13(6): 808.

[15] Puzyrev, V., Zelic, M. and Duuring, P. (2023). Applying neural networks-based modelling to the prediction of mineralization: A case-study using the Western Australian Geochemistry (WACHEM) database. Ore Geology Reviews 152: 105242.

[16] Tahmooresi, M., Babaei, B. and Dehghan, S. (2022). Geochemical exploration numerical modeling using convolutional neural network (Case study: Gonabad region). Analytical and Numerical Methods in Mining Engineering 12(31): 47-58.

imbalanced data classification in healthcare. BioData Mining 16: 15.

[41] Hengyu, Z. (2020). Improved SMOTE algorithm for imbalanced dataset. Chinese Automation Congress (CAC), Shanghai, China, 693-697.

[42] Lee, H., Kim, J. and Kim, S. (2017). Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions. Int. J. Fuzzy Log. Intell. Syst. 17(4): 229-234.

[43] He, H., Bai, Y., Garcia, E.A. and Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning, IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 1322-1328.

[44] Brandt, J. and Lanzén, E. (2021). A Comparative Review of SMOTE and ADASYN in Imbalanced Data Classification. Department of Statistics, Uppsala University, 42 P.

[45] Kurniawati, Y.E., Permanasari, A.E. and Fauziati, S. (2018). Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data, 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 1-6.

[46] Qing, Z., Zeng, Q., Wang, H., Liu, Y., Xiong, T. and Zhang, S. (2022). ADASYN-LOF Algorithm for Imbalanced Tornado Samples. Atmosphere, 13(4): 544.

[47] Devi, D., Biswas, S.K. and Purkayastha, B. (2020). A Review on Solution to Class Imbalance Problem: Undersampling Approaches, International Conference on Computational Performance Evaluation (ComPE), Shillong, India, 626-631.

[48] Mazhari, S.A. and Safari, M. (2013). High-K Calc-alkaline Plutonism in Zouzan, NE of Lut Block, Eastern Iran: An Evidence for Arc Related Magmatism in Cenozoic. Journal Geological Society of India 81: 698-708.

[49] Geranian, H. and Carranza, E.J.M. (2022). Mapping of Regional-scale Multi-Element Geochemical Anomalies Using Hierarchical Clustering Algorithms. Natural Resources Research 31(4): 1841-1865.

[50] Seyedrahimi-Niaraq, M., Mahdiyanfar, H. and Mokhtari, A. R. (2023). Application of geochemical structural methods to determine lead-contaminated areas related to mining activities. *Journal of Analytical and Numerical Methods in Mining Engineering* 13(34): 41-55.

[51] Kubat, M. and Matwin, S. (1997). Addressing the course of imbalanced training sets:

[29] Wang, W., and Sun, D. (2021). The improved AdaBoost algorithms for imbalanced data classification. Information Sciences 563: 358-374.

[30] Salehi, A. R., and Khedmati, M. (2024). A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data. Scientific Reports 14(1): 5152.

[31] Araf, I., Idri, A., and Chairi, I. (2024). Cost-sensitive learning for imbalanced medical data: a review. Artificial Intelligence Review 57(4): 80.

[32] Xiao, J., Li, S., Tian, Y., Huang, J., Jiang, X., and Wang, S. (2025). Example dependent cost sensitive learning based selective deep ensemble model for customer credit scoring. Scientific Reports 15(1): 6000.

[33] Liu, Y., Li, Z., Chen, J., Zhang, T., Pan, T., and He, S. (2025). A batch-adapted cost-sensitive contrastive feature learning network for industrial diagnosis with extremely imbalanced data. Measurement 244: 116478.

[34] Yuan, Y., Wei, J., Huang, H., Jiao, W., Wang, J. and Chen, H. (2023). Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring. Engineering Applications of Artificial Intelligence 126: 106911.

[35] Abhishek, K. and Abdelaziz, M. (2023). Machine Learning for Imbalanced Data: Tackle imbalanced datasets using machine learning and deep learning techniques, Packt Publishing, 344 p.

[36] Yang, Y., Akbarzadeh Khorshidi, H. and Aickelin, U. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. Front. Digit. Health 26: 1430245.

[37] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16: 321–357.

[38] Hu, S., Liang, Y., Ma, L. and He, Y. (2009). MSMOTE: Improving Classification Performance when Training Data is imbalanced. 2009 Second International Workshop on Computer Science and Engineering, 13-17.

[39] Tahmooresi, M., Babaei, B. and Dehghan, S. (2022). Geochemical exploration numerical modeling using convolutional neural network (Case study: Gonabad region). *Journal of Analytical and Numerical Methods in Mining Engineering* 12(31): 47-58.

[40] Kosolwattana, T., Liu, C., Hu, R. Han, S., Chen, H. and Lin, Y. (2023). A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly

One-sided selection. Proceedings of the 14th international conference on machine learning, Morgan Kaufmann, pp. 179-186.

[52] Jia, C. and Zuo, Y. (2017). S-SulfPred: A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique. Journal of Theoretical Biology 422: 84-89.

[53] Batista, G., Bazzan, A. and Monard, MC. (2003). Balancing Training Data for Automated Annotation of Keywords: A Case Study. II Brazilian Workshop on Bioinformatics, 10-18.

[54] Hart, P.E. (1968). The Condensed Nearest Neighbour Rule. IEEE Transactions on Information Theory 14(5): 515-516.

[55] Hassani Pak, A.A. (2016). Principles of Geochemical Exploration. Tehran University Press, Tehran [In Persian].

[56] Fakhari, S., Jafarirad, A., Afzal, P., and Lotfi, M. (2019). Delineation of hydrothermal alteration zones for porphyry systems utilizing ASTER data in Jebal-Barez area, SE Iran. Iranian Journal of Earth Sciences, 11: 80-92.

[57] Mokhtari, Z., and Seifi, A. (2021). Detection of Hydrothermal Alteration Zones Using ASTER Remote Sensing Data in Turquoise mine of Neyshabur. Journal of Analytical and Numerical Methods in Mining Engineering, 11(28): 1-22 [In Persaian].

[1] Extreme Gradient boosting (XGboost)
[2] Imbalanced Dataset
[3] Majority Class
[4] Minority Class
[5] Imbalance Ratio
[6] Oversampling
[7] Undersampling
[8] Random Oversampling (ROS)
[9] Synthetic Minority Oversampling Technique (SMOTE)
[10] Borderline Oversampling (BOS)
[11] Adaptive Synthetic Sampling (ADASYN)
[12] Random Undersampling (RUS)
[13] Condensed Nearest Neighbor Rule (CNN)
[14] Near Miss Undersampling (NMUS)
[15] Tomek Links Undersampling (TLUS)
[16] Edited Nearest Neighbors Rule (ENN)
[17] One-Sided Selection (OSS)
[18] Neighborhood Cleaning Rule (NCR)
[19] Support Vector Machine (SVM)
[20] Artificial Neural Network (ANN)
[21] Receiver Operating Characteristic (ROC)