# روش تلفیقی مدل‌های آماری انباشته در بهبود نتایج یادگیری ماشین در پیش بینی گاف انرژی مواد

انوشیروان غفاری پور[1*]، بهروز واثقی[2]

[2] گروه آمار، دانشکده علوم پایه، دانشگاه یاسوج، یاسوج، ایران

[1] گروه فیزیک، دانشکده علوم پایه، دانشگاه یاسوج، یاسوج، ایران

اطلاعات مقاله

تاریخ دریافت: ۱۴۰۴/۱۰/۲۹
تاریخ پذیرش: ۱۴۰۴/۱۱/۱۹
تاریخ چاپ: ۱۴۰۴/۱۱/۲۰

**شاپای چاپی:** 2588-493x
**شاپای الکترونیکی:** 2588-4921

* نویسنده مسئول
aghaffaripour@yu.ac.ir

## چکیده

در این پژوهش با استفاده از رویکرد یادگیری ماشین به پیش بینی گاف انرژی دسته ای از مواد نیمه رسانا پرداخته شده است. هدف اصلی پژوهش بر این مبنا بوده که با تلفیق روش های مختلف در یادگیری ماشین، سعی بر ارائه روشی کارآمد در پیش بینی گاف انرژی مواد داشته باشیم. روش انباشته یک روش جدید در میان آلگوریتم های یادگیری ماشین می باشد که با بهره گیری از روش های مرسوم، تلفیق آنها با یکدیگر و استفاده از مزیت هریک از روش های آماری روشی کارآمد و با دقت بالاتر نسبت به هریک از روش های مورد استفاده، می باشد. با استفاده از روش هایی چون از درخت تصمیم تقویت گرادیان، تقویت گرادیان سبک، جنگل تصادفی و تقویت گرادیان حداکثری یک مدل ترکیبی انباشته ارائه داده ایم که نتایج شبیه سازی بهتری نسبت به هریک از روش های نام برده ارائه کرده است.

***واژگان کلیدی:*** یادگیری ماشین، گاف انرژی، XGBoost، Stacking، ضریب رگرسیون.

## مقدمه

در دهه اخیر، پیشرفت‌های چشمگیر در حوزه آمار محاسباتی و یادگیری ماشین، رویکرد تحلیل داده در علوم مهندسی و فیزیک را به‌طور اساسی متحول کرده است. در این میان، روش‌های داده‌محور به‌عنوان ابزارهایی قدرتمند برای مدل‌سازی روابط پیچیده، غیرخطی و چندبعدی مطرح شده‌اند؛ روابطی که غالباً فراتر از توان روش‌های تحلیلی کلاسیک یا مدل‌های فیزیکی ساده قرار دارند. یادگیری ماشین را می‌توان در این چارچوب به‌عنوان بسطی از مدل‌سازی آماری دانست که با تکیه بر داده، وظیفه تقریب نگاشت‌های پیچیده میان متغیرهای ورودی و خروجی را بر عهده دارد[۱و۲].

یکی از کاربردهای مهم این رویکرد، پیش‌بینی خواص مواد است؛ حوزه‌ای که در آن خواص فیزیکی و الکترونیکی معمولاً حاصل برهم‌کنش‌های پیچیده میان پارامترهای شیمیایی، ساختاری و الکترونی هستند. از منظر آماری، بسیاری از این مسائل را می‌توان به‌صورت مسائل رگرسیونی چندمتغیره با ساختار غیرخطی و نویز ذاتی در نظر گرفت. در این میان، گاف نواری انرژی مواد نیمه‌رسانا به‌عنوان یکی از کلیدی‌ترین کمیت‌ها، نقش تعیین‌کننده‌ای در عملکرد ادوات اپتوالکترونیکی، فوتوالکتروشیمیایی و الکترونیکی ایفا می‌کند و پیش‌بینی دقیق آن همواره مورد توجه پژوهشگران بوده است. روش‌های محاسباتی مبتنی بر اصول اولیه، به‌ویژه نظریه

تابعی چگالی[1]، چارچوب فیزیکی معتبری برای محاسبه گاف نواری فراهم می‌کنند. با این حال، این روش‌ها از دیدگاه آماری و محاسباتی با محدودیت‌هایی مواجه‌اند؛ از جمله هزینه محاسباتی بالا، حساسیت به انتخاب تقریب‌های تابعی، و دشواری تعمیم به فضای طراحی بزرگ. در مقابل، یادگیری ماشین با تکیه بر داده‌های موجود، امکان ساخت مدل‌های آماری جانشین[2] را فراهم می‌کند که قادرند با هزینه محاسباتی بسیار کمتر، پیش‌بینی‌هایی سریع و نسبتاً دقیق ارائه دهند[ ۳و۴و۵].

## ۱- نتایج حاصل از مدل سازی

در ساده‌ترین سطح، مدل‌های رگرسیونی کلاسیک نظیر رگرسیون خطی و رگرسیون چندجمله‌ای می‌توانند به‌عنوان نقطه شروع در تحلیل داده‌های مواد مورد استفاده قرار گیرند. این مدل‌ها اگرچه از شفافیت و تفسیرپذیری بالایی برخوردارند، اما به دلیل فرض خطی‌بودن یا محدودیت در انعطاف‌پذیری تابعی، معمولاً قادر به توصیف کامل روابط پیچیده میان ویژگی‌های عنصری و گاف نواری نیستند. از این رو، استفاده از مدل‌های آماری پیشرفته‌تر که توانایی مدل‌سازی غیرخطی را دارند، ضرورت می‌یابد.

الگوریتم‌هایی نظیر رگرسیون بردار پشتیبان[3] ، روش‌های مبتنی بر درخت تصمیم مانند جنگل تصادفی[4]، و مدل‌های تقویتی نظیر تقویت گرادیان[5]، به‌طور گسترده در مسائل رگرسیونی پیچیده به‌کار گرفته شده‌اند. هر یک از این مدل‌ها از منظر آماری دارای ویژگی‌های خاص خود هستند؛ برای مثال، SVR با استفاده از توابع کرنل، فضای ویژگی‌ها را به فضایی با بعد بالاتر نگاشت کرده و امکان مدل‌سازی روابط غیرخطی را فراهم می‌کند، در حالی که مدل‌های مبتنی بر درخت با تقسیم‌بندی فضای داده، قادر به شناسایی تعامل‌های محلی میان متغیرها هستند. با این وجود، عملکرد این مدل‌ها به‌شدت به ساختار داده، انتخاب ابرپارامترها و نحوه نمونه‌برداری وابسته است و هیچ‌یک به‌تنهایی تضمین‌کننده بهترین عملکرد در تمامی نواحی فضای داده نیستند. این چالش از دیدگاه آماری، به مسئله بایاس–واریانس و حساسیت مدل‌ها به نویز و پراکندگی داده بازمی‌گردد. در چنین شرایطی، چارچوب‌های **یادگیری تجمعی**[6] به‌عنوان راهکاری مؤثر برای بهبود عملکرد مدل‌ها مطرح می‌شوند. ایده اصلی در این روش‌ها، ترکیب چندین مدل پایه با ساختارهای متفاوت است، به‌گونه‌ای که نقاط ضعف هر مدل توسط نقاط قوت مدل‌های دیگر جبران شود.

در میان روش‌های ensemble، الگوریتم‌هایی مانند Bagging و Boosting به‌طور گسترده مورد استفاده قرار گرفته‌اند .Bagging عمدتاً با کاهش واریانس مدل از طریق آموزش مدل‌ها بر روی نمونه‌های بوت‌استرپ‌شده عمل می‌کند، در حالی که Boosting با تمرکز تدریجی بر نمونه‌های دشوار، تلاش می‌کند خطای سیستماتیک مدل را کاهش دهد. با این حال، این روش‌ها نیز معمولاً بر یک نوع مدل پایه متکی هستند و از تنوع ساختاری محدودی برخوردارند[ ۶ و۷].

در این میان، **مدل** انباشه به‌عنوان یکی از پیشرفته‌ترین چارچوب‌های یادگیری تجمعی، جایگاه ویژه‌ای از منظر آماری دارد. انباشته با ترکیب مدل‌های پایه ناهمگن و استفاده از یک مدل سطح دوم فرا-یادگیرنده[7]، امکان یادگیری ساختار خطاهای مدل‌های پایه و وزن‌دهی بهینه به پیش‌بینی‌های آن‌ها را فراهم می‌کند. از دیدگاه آماری، این رویکرد معادل با یادگیری یک نگاشت ثانویه بر روی فضای پیش‌بینی‌ها است که می‌تواند منجر به کاهش هم‌زمان بایاس و واریانس شود.

اهمیت استفاده از انباشته در مسائل پیش‌بینی خواص مواد، به‌ویژه زمانی برجسته می‌شود که دیتاست‌ها دارای اندازه محدود، ناهمگونی بالا و روابط غیرخطی پیچیده باشند؛ شرایطی که در بسیاری از مسائل علم مواد، از جمله پیش‌بینی گاف نواری، به‌طور طبیعی وجود دارد. در چنین مسائلی، مدل‌های منفرد ممکن است تنها بخشی از ساختار آماری داده را یاد بگیرند، در حالی که مدل انباشته با ترکیب دیدگاه‌های آماری متفاوت، تصویر کامل‌تری از رابطه بین ویژگی‌های ورودی و خروجی ارائه می‌دهد.

در این پژوهش، پیش‌بینی گاف نواری گروهی از مواد نیمه رسانا به‌عنوان یک مسئله رگرسیونی داده‌محور مورد مطالعه قرار گرفته است. با استفاده از یک دیتاست برگرفته از سیستم مورد بررسی و تمرکز بر ویژگی‌های عنصری مواد، مجموعه‌ای از مدل‌های آماری شامل رگرسیون‌های کلاسیک، الگوریتم‌های یادگیری ماشین منفرد و چارچوب‌های ensemble مورد ارزیابی قرار گرفته‌اند. در نهایت، مدل انباشته به‌عنوان رویکرد اصلی انتخاب شده است تا با بهره‌گیری از توان مکمل مدل‌های پایه، دقت پیش‌بینی و قابلیت تعمیم مدل بهبود یابد. این مطالعه بر نقش کلیدی مدل‌سازی آماری و یادگیری ماشین در تحلیل مسائل پیچیده

---

[1] Density Functional Theory
[2] Surrogate Models
[3] Support Vector Regression
[4] Random Forest
[5] Gradient Boosting
[6] Ensemble Learning
[7] Meta-Learner

فیزیکی تأکید داشته و نشان می‌دهد که رویکردهای داده‌محور می‌توانند به‌عنوان ابزاری کارآمد در پیش‌بینی خواص مواد مورد استفاده قرار گیرند.

سیستم های مورد بررسی در این مدل سازی ۳۰۰ نمونه ماده نیمرسانا بوده که هر یک دارای ۸ اتم در ساختار خود می باشند. متغیرهای ورودی در شبیه سازی یادگیری ماشین برای این ۳۰۰ ترکیب در فایل کمکی مقاله ارائه شده است. شعاع یونی، الکترونگاتیوی و انرژی یونیزاسیون متغیرهای ورودی برای این ۳۰۰ ترکیب می باشند. از آنجا که هدف این مقاله ارائه روشی کارآمد برای پیش بینی گاف انرژی مواد نیمرسانا با تقریب HSE06 می باشد، مقدار گاف انرژی ترکیبات با تقریب GGA-PBE نیز به عنوان متغیر ورودی نیز در مدل سازی یادگیری ماشین نیز وارد شده است. در ادامه به تفسیر ارتباط فیزیکی هریک از متغیرهای در نظر گرفته شده با گاف انرژی مواد پرداخته شده است.

شعاع یونی اتم ها نقش مهمی در تعیین فاصله بین اتم‌ها و میزان هم‌پوشانی اوربیتال‌های الکترونی ایفا می‌کند. افزایش شعاع یونی معمولاً منجر به افزایش فاصله بین اتمی و کاهش هم‌پوشانی اوربیتال‌ها می‌شود که این امر پهنای نوارهای الکترونی را کاهش داده و می‌تواند به افزایش موضعی‌بودن حالات الکترونی منجر شود. از منظر فیزیک نوارها، کاهش پهنای نوار ظرفیت و رسانش معمولاً باعث تغییر در مقدار گاف نواری می‌شود. در بسیاری از نیمه‌رساناها، افزایش شعاع یونی عناصر سازنده با کاهش قدرت پیوند و تضعیف برهم‌کنش‌های الکترونی همراه است که می‌تواند موجب تغییر محسوس در مقدار گاف انرژی گردد.

الکترونگاتیوی نیز به‌عنوان معیاری از تمایل اتم برای جذب الکترون، تأثیر مستقیمی بر ماهیت پیوند شیمیایی و توزیع بار الکترونی دارد. اختلاف الکترونگاتیوی میان عناصر تشکیل‌دهنده یک ماده، میزان یونی یا کووالانسی بودن پیوندها را تعیین می‌کند. افزایش الکترونگاتیوی معمولاً منجر به تمرکز بیشتر چگالی الکترونی در اطراف اتم‌های با الکترونگاتیوی بالاتر شده و جدایش انرژی حالات اشغال‌شده و اشغال‌نشده را افزایش می‌دهد. از دیدگاه فیزیکی، این جدایش انرژی می‌تواند به افزایش گاف نواری منجر شود، زیرا اختلاف پتانسیل مؤثر تجربه‌شده توسط الکترون‌ها در نوار ظرفیت و رسانش تقویت می‌شود.

انرژی یونیزاسیون نیز بیانگر میزان انرژی لازم برای جدا کردن یک الکترون از اتم در حالت پایه است و مستقیماً با پایداری حالات الکترونی و عمق ترازهای انرژی مرتبط است. اتم‌هایی با انرژی یونیزاسیون بالاتر، الکترون‌های خود را قوی تر نگه می‌دارند که این موضوع منجر به پایین‌تر قرار گرفتن ترازهای انرژی نوار ظرفیت می‌شود. در نتیجه، افزایش انرژی یونیزاسیون معمولاً با افزایش گاف نواری همراه است، زیرا فاصله انرژی بین بالاترین حالت‌های اشغال‌شده و پایین‌ترین حالت‌های غیر اشغال‌شده افزایش می‌یابد. این پارامتر در کنار الکترونگاتیوی، تصویری فیزیکی از شدت برهم‌کنش الکترون–هسته و میزان محبوس‌سازی الکترون‌ها در شبکه بلوری ارائه می‌دهد.

در مجموع، گاف انرژی مواد نیمه‌رسانا حاصل برهم‌کنش پیچیده این ویژگی‌های اتمی است و تغییر هر یک از آن‌ها می‌تواند به‌صورت غیرخطی و وابسته به ساختار بلوری، مقدار گاف نواری را تحت تأثیر قرار دهد. این پیچیدگی فیزیکی، توجیه‌کننده استفاده از روش‌های یادگیری ماشین برای مدل‌سازی دقیق این روابط چندمتغیره است.

ضریب همبستگی پیرسون برای بررسی رفتار دوتایه بین متغیرها در مدل سازی حاضر مورد بررسی قرار گرفت. این کمیت بصورت زیر تعریف می شود[۶]:

$$r = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \left(y_i - \bar{y}\right)^2}} \qquad (۱)$$

این ضریب دارای مقداری بین ۱- تا ۱+ را دارا می باشد. مقدار ۱+ آن به این معنا می باشد که با افزایش متغیر x متغیر y نیز افزایش می یابد و برعکس. اما مقدار ۱- آن بیان می کند که هرگاه متغیر x افزایش یابد به چه میزان متغیر y کاهش می یابد و یک همبستگی منفی بین x و y برقرار است. مقدار ضریب همبستگی صفر هم به این معناست که هیچ گونه ارتباطی بین این دو متغیر نمی توان پیش بینی کرد.

## ۲- تفسیر فیزیکی و آماری همبستگی بین متغیرها

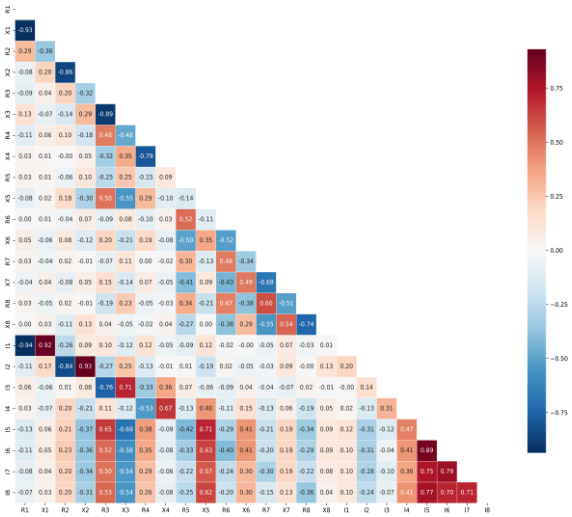
نتایج ضریب همبستگی را می توان در شکل ۱ برای متغیرهای مسئله مشاهده کرد. نتایج همبستگی پیرسون نشان می‌دهد که بین شعاع یونی و الکترونگاتیوی یک همبستگی منفی قوی برقرار است که از دیدگاه فیزیکی کاملاً قابل انتظار است. اتم‌هایی با شعاع یونی بزرگ‌تر معمولاً دارای الکترون‌های لایه والانس دورتر از هسته بوده و نیروی جاذبه مؤثر هسته بر الکترون‌ها در آن‌ها ضعیف‌تر است. این امر باعث کاهش تمایل اتم به جذب الکترون و در نتیجه کاهش الکترونگاتیوی می‌شود. به‌طور مشابه، همبستگی منفی بین شعاع یونی و انرژی یونیزاسیون نیز بازتاب همین اصل فیزیکی است؛

به‌طوری که افزایش شعاع یونی، جداسازی الکترون از اتم را آسان‌تر کرده و انرژی یونیزاسیون را کاهش می‌دهد.

در مقابل، مشاهده همبستگی مثبت قوی بین الکترونگاتیوی و انرژی یونیزاسیون بیانگر آن است که هر دو کمیت ریشه در شدت برهم‌کنش الکترون–هسته دارند. اتم‌هایی که الکترونگاتیوی بالاتری دارند، معمولاً انرژی یونیزاسیون بیشتری نیز از خود نشان می‌دهند، زیرا الکترون‌ها در این اتم‌ها به‌طور مؤثرتری به هسته مقید شده‌اند. این هم‌راستایی فیزیکی میان این دو پارامتر سبب می‌شود که اثرات آن‌ها بر گاف انرژی نیز هم‌جهت بوده و در بسیاری از مواد، افزایش همزمان این دو کمیت به افزایش گاف نواری منجر شود.

از دیدگاه آماری، ضرایب همبستگی پیرسون محاسبه‌شده نشان می‌دهند که متغیرهای ورودی مدل از استقلال آماری برخوردار نبوده و بین آن‌ها وابستگی‌های دوتایی معناداری وجود دارد. وجود همبستگی‌های قوی مثبت یا منفی میان متغیرها بیانگر آن است که اطلاعات ارائه‌شده توسط برخی ویژگی‌ها تا حدی همپوشانی دارد. این مسئله در مدل‌سازی رگرسیونی کلاسیک می‌تواند منجر به چندهمخطی و کاهش پایداری ضرایب مدل شود، اما در چارچوب الگوریتم‌های یادگیری ماشین غیرخطی و به‌ویژه روش‌های ensemble، این همبستگی‌ها لزوماً مضر نبوده و حتی می‌توانند به بهبود قدرت پیش‌بینی کمک کنند.

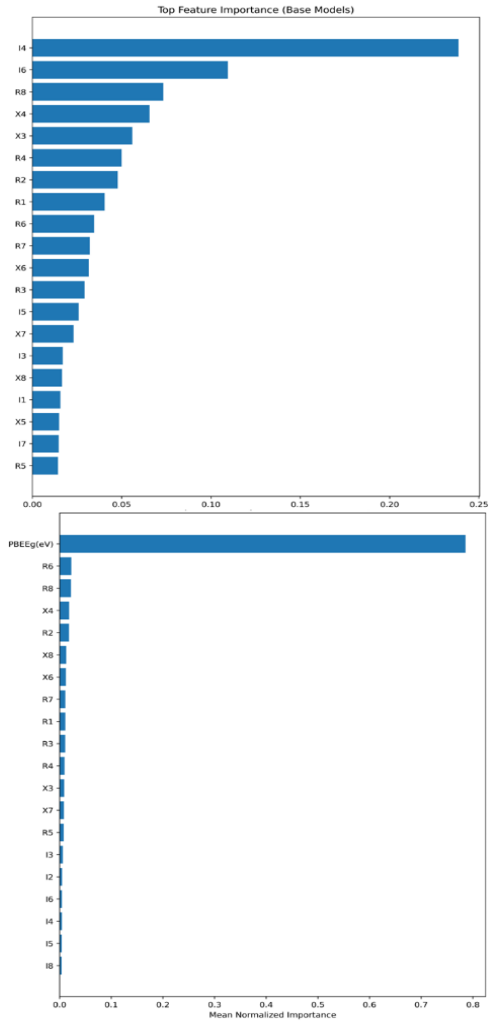
در روش‌های مبتنی بر درخت تصمیم و مدل‌های تقویتی، ساختار غیرخطی و سلسله‌مراتبی مدل‌ها امکان بهره‌برداری مؤثر از این همبستگی‌ها را فراهم می‌کند. الگوریتم‌های ensemble مانند انباشته با ترکیب چندین مدل پایه، قادرند الگوهای مشترک و مکمل موجود در متغیرهای همبسته را شناسایی کرده و وزن‌دهی بهینه‌ای به آن‌ها اختصاص دهند. از این منظر، همبستگی مشاهده‌شده میان شعاع یونی، الکترونگاتیوی و انرژی یونیزاسیون نه‌تنها بازتابی از واقعیت فیزیکی سیستم است، بلکه از دیدگاه آماری نیز نقش مهمی در افزایش دقت و پایداری پیش‌بینی گاف انرژی در مدل نهایی ایفا می‌کند.



**شکل ۱: نمودار همبستگی پیرسون برای متغیرهای مسئله**

در گام بعد از محاسبات مدل سازی یادگیری ماشین، به بررسی اهمیت متغیرهای ورودی در شبیه سازی پرداخته شده است. بصورت کلی با داشتن ۸ اتم پایه در ساختار و ۳ ویژگی برای هر اتم، ۲۷ متغیر اتمی در محاسبات مدل سازی وارد شده اند. برای بررسی اهمیت هر یک از متغیرها نتایج حاصل برای ۲۰ متغیر اول در زیر یکبار با درنظر گرفتن متغیر گاف PBE و یکبار بدون آن ارائه شده است.

همان طور که می بینیم نقش متغیر PBE بسیار چشمگیر تر از هر یک از متغیرهای دیگر می باشد. نتایج به ۱ بهنجار شده اند. برای نشان دادن نقش دیگر متغیرها، بدون در نظر گرفتن متغیر PBE ، نیز نتایج را تکرار کرده ایم. همان طور که پیداست به غیر از انرژی یونیزاسیون اتم ۴، الکترونگاتیوی و شعاع یونی دارای بیشترین نقش در مدل سازی سیستم های مورد بررسی را دارد.




شکل ۲- اهمیت متغیرها (شکل بالا) بدون در نظر گرفتن گاف نواری PBE و (شکل پایین) با در نظر گرفتن PBE.

همان طور که در قسمت قبل گفته شد، هدف اصلی این مقاله ارائه روش تلفیقی انباشته برای بهبود نتایج حاصل از مدل سازی یادگیری ماشین با استفاده از تک مدل های مرسوم می باشد. در روش انباشته ابتدا با تقسیم دیتاست به قسمت های مختلف و ارسال هر یک به تک روش های زیر مجموعه تشکیل دهنده روش انباشته ، هریک از مدل ها را بهینه کرده و در نهایت با تلفیق همه آنها باهمدیگر، بهینه ترین نتایج را ارائه می کند. به منظور مقایسه نتایج و بررسی کیفیت مدل سازی آنها پارامترهای زیر را به عنوان متغیرهای مقایسه در مدل سازی در نظر گرفته ایم[۸]:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \overline{y_i}\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y_i}\right)^2} \qquad (۲)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{N}\left(y_i - y_i\right)^2 \qquad (۳)$$

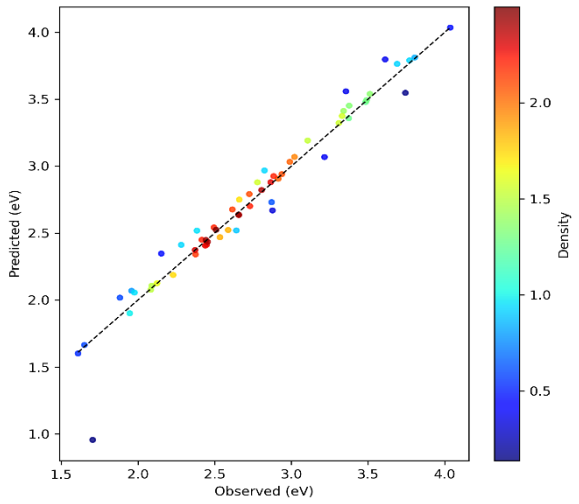
$$MAE = \frac{1}{n}\sum_{i=1}^{N}\left|y_i - y_i\right| \qquad (۴)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}\left(y_i - y_i\right)^2} \qquad (۵)$$

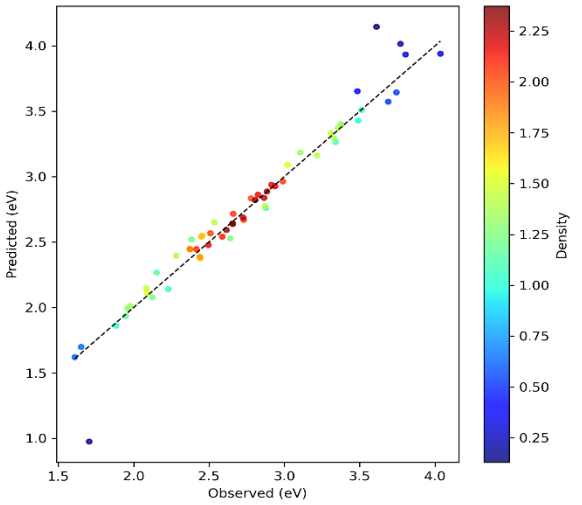
که در آن $y_i$ مقادیر گاف انرژی HSE حاصل از محاسبات شبیه سازی DFT ، $y_i$ مقادیر پیش بینی شده حاصل از مدل سازی یادگیری ماشین و $\overline{y_i}$ مقدار میانگین $y_i$ ها می باشد.

مدل های در نظر گرفته شده در این مدل سازی عبارتند از درخت تصمیم تقویت گرادیان(GBDT) ، تقویت گرادیان سبک(LGB) ، جنگل تصادفی(RF) و تقویت گرادیان حداکثری(XGB) . همچنین از روش XGB به عنوان متارگرسور استفاده شده است. در طی مراحل اجرای کد یادگیری ماشین، ۸۰ درصد از داده ها در فرآیند یادگیری مدل ها مورد استفاده قرار گرفته اند و ۲۰ در صد از داده ها برای تست مدل ها استفاده شده اند.

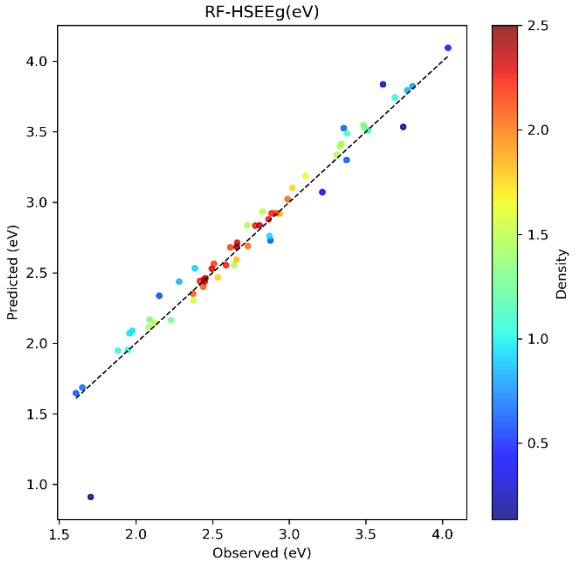
نتایج حاصل از اجرای تک مدل ها و در نهایت روش انباشته در شکل های زیر نشان داده شده است. همچنین پارامترهای آماری مورد استفاده به منظور سنجش مدل نیز در جدول ۱ لیست شده اند. به خوبی از نتایج پیداست که روش انباشته دارای نتایج پیش بینی بهتری نسبت به هر یک از تک مدل های استفاده شده در فرآیند یادگیری را دارا می باشد. مقدار $R^2$ به دست آمده برای روش انباشته برابر با مقدار $R^2$=0.9685 می باشد که بیانگرپیش بینی مناسب این روش از نتایج گاف انرژی ترکیبات می باشد. همچنین مقدار خطای مربعی میانگین ریشه این روش نیز دارای کمترین مقدار در بین روش های دیگر می باشد که برابر با مقدار 0.1032 به دست آمد.



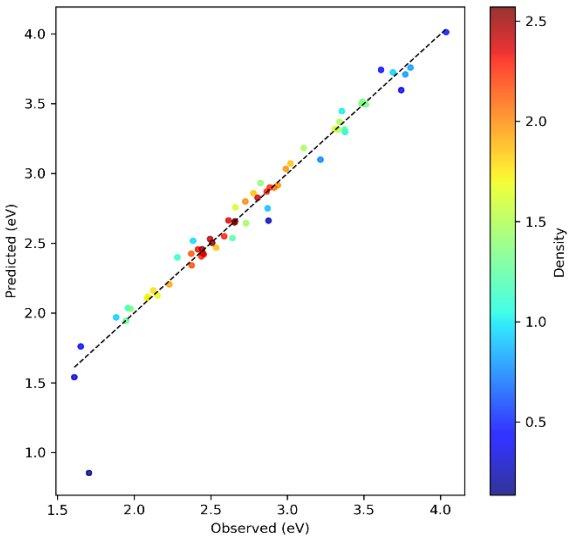

شکل ۳: نتایج مدل سازی روش GBDT



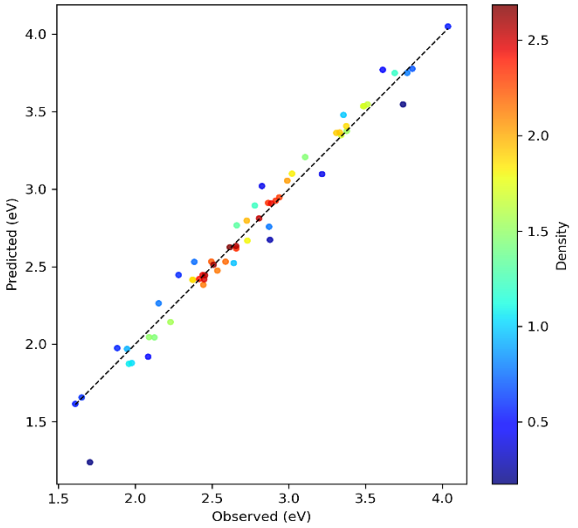

شکل ۴: نتایج مدل سازی روش LGB




شکل ۵: نتایج مدل سازی روش RF




شکل ۶: نتایج مدل سازی روش XGB




شکل ۷: نتایج مدل سازی روش Stacking

جدول ۱- نتایج آماری حاصل از مدل سازی

| LGB | RF | XGB | GBDT | Stacking | Metrics |
|---|---|---|---|---|---|
| 0.1387 | 0.1328 | 0.1301 | 0.1282 | 0.1032 | RMSE |
| 0.0798 | 0.078 | 0.0683 | 0.0727 | 0.0717 | MAE |
| 0.0309 | 0.0319 | 0.0287 | 0.029 | 0.0285 | MAPE |
| 0.9431 | 0.9479 | 0.9499 | 0.9514 | 0.9685 | $R^2$ |

## ۳- نتیجه گیری

انتخاب متغیرهای شعاع یونی، الکترونگاتیوی و انرژی یونیزاسیون به‌عنوان ویژگی‌های ورودی مدل‌سازی، از منظر فیزیک حالت جامد مبتنی بر نقش بنیادی این

کمیت‌ها در تعیین ساختار الکترونی و ماهیت پیوندهای شیمیایی مواد نیمه‌رسانا است. همبستگی‌های مشاهده‌شده میان این متغیرها بازتاب مستقیمی از اصول فیزیکی حاکم بر برهم‌کنش الکترون–هسته و میزان محبوس‌سازی الکترون‌ها در شبکه بلوری است. همبستگی منفی قوی بین شعاع یونی و دو کمیت الکترونگاتیوی و انرژی یونیزاسیون نشان می‌دهد که افزایش اندازه یونی با کاهش شدت پیوند و تضعیف جذب الکترونی همراه است، در حالی که همبستگی مثبت میان الکترونگاتیوی و انرژی یونیزاسیون بیانگر تقویت همزمان این دو اثر در افزایش جدایش انرژی ترازهای الکترونی است. این روابط فیزیکی به‌طور مستقیم بر مقدار گاف نواری اثر گذاشته و تأیید می‌کنند که متغیرهای انتخاب‌شده قادرند اطلاعات معنادار و مکملی از مکانیزم‌های فیزیکی مؤثر بر گاف انرژی را در اختیار مدل قرار دهند.

از دیدگاه آماری، وجود همبستگی‌های قابل توجه میان متغیرهای ورودی نشان‌دهنده ساختار داده‌ای پیچیده و غیرمستقل است که مدل‌های رگرسیونی ساده قادر به توصیف کامل آن نیستند. این وابستگی‌ها اگرچه می‌توانند در مدل‌های کلاسیک منجر به چندهمخطی شوند، اما در چارچوب الگوریتم‌های یادگیری ماشین غیرخطی و به‌ویژه روش‌های تجمعی، به‌عنوان منبعی غنی از اطلاعات مورد بهره‌برداری قرار می‌گیرند. استفاده از مدل انباشته با ترکیب مدل‌های پایه ناهمگن، امکان استخراج همزمان الگوهای محلی، غیرخطی و مکمل موجود در متغیرهای همبسته را فراهم کرده و با یادگیری بهینه وزن پیش‌بینی هر مدل، منجر به کاهش همزمان بایاس و واریانس شده است. بهبود قابل توجه شاخص‌های آماری، به‌ویژه افزایش ضریب تعیین و کاهش خطا، نشان می‌دهد که رویکرد انباشته نه‌تنها از نظر آماری کارآمد است، بلکه به‌طور مؤثری توانسته پیچیدگی فیزیکی روابط میان متغیرهای اتمی و گاف انرژی را در مدل‌سازی داده‌محور منعکس کند.

## منابع


[1] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, "Predicting the Band Gaps of Inorganic Solids by Machine Learning", *Journal of Physical Chemistry Letters* Vol 9, 1668-1673, 2018.

[2] A. Sabagh Moeini, F. Shariatmadar Tehrani, A. Naeimi-Sadigh "Machine learning-enhanced band gaps prediction for low-symmetry double and layered perovskites", *Scientific Reports* Vol. 14, 26736, 2024.

[3] S.G. Jung, G. Jung, J.M. Cole, "Automatic Prediction of Band Gaps of Inorganic Materials Using a Gradient Boosted and Statistical Feature Selection Workflow", *Journal of Chemical Information and Modeling* Vol. *64, 1187-1200,* 2024.

[4] E. Ogoshi et al., "Learning from machine learning: the case of band-gap directness in semiconductors", *Discover Materials* Vol. 4, 6, 2024.

[5] A.Ch. Rajan et al., "Machine-Learning Assisted Accurate Band Gap Predictions of Functionalized MXene" *Chemistry of Materials* Vol. 4, 112, 2018.

[6] T. Wang, K. Zhang, J. The, H. Yu, "Accurate prediction of band gap of materials using stacking machine learning model" *Computational Materials Science* Vol. 201, 110899, 2022.

[7] S. Priyanga et al., "Prediction of nature of band gap of perovskite oxides ($ABO_3$) using a machine learning approach", *Journal of Materiomics* Vol. 8, 937e948, 2022.

[8] J. Xu et al., "Machine learning predictions of band gap and band edge for $(GaN)_{1-x}(ZnO)_x$ solid solution using crystal structure information" Journal of Material Sciences Vol. 58, 7986–7994, 2023.

JMRPh. (10)(2): 18-25, Autumn & Winter 2025

# Efficient machine learning method, Stacking, to improvement of material band gaps prediction

[1*] Anoshirvan Ghaffaripour, [2] Behrooz Vaseghi

[1*] Department of Statistics, College of Sciences, Yasouj University, Yasouj, Iran

[1*] Department of Physics, College of Sciences, Yasouj University, Yasouj, Iran

**Article details**

Received: 2026/01/19
Accepted: 2026/02/8
Published: 2026/02/9

ISSN: 2588-493x
eSSN: 2588-4821

Correspondence email: aghaffaripour@yu.ac.ir

## Abstract

In this study, the energy band gaps of 300 selected materials were predicted using machine learning techniques. Several widely used machine learning algorithms were employed and systematically combined to develop an efficient predictive framework for material band gap estimation. Specifically, Gradient Boosting Decision Trees, Light Gradient Boosting Machine, Random Forest, and Extreme Gradient Boosting models were integrated within a stacking ensemble strategy. By leveraging the complementary strengths of these individual learners, the proposed stacking approach demonstrates improved accuracy and robustness in band gap prediction compared to single-model methods.

**Keywords:** Machine Learning, Energy Band Gap, XGBoost ،Stacking, Regression Coefficient.