INTISARI

PREDIKSI *BAND GAP* MATERIAL 2D MENGGUNAKAN RANDOM FOREST DAN XGBOOST DENGAN ANALISIS INTERPRETABILITAS SHAP PADA DATASET C2DB

Oleh:

Iren Yolanda Sitompul

21/481208/PA/20942

Studi ini memprediksi *band gap* Kohn-Sham material 2D semikonduktor menggunakan pendekatan *machine learning* berbasis data hasil perhitungan Density Functional Theory dari Computational 2D Materials Database (C2DB). Analisis menggunakan fitur fisik global yang tersedia dalam dataset, yaitu energi total per atom, energi pembentukan, energi terhadap *convex hull*, ketebalan lapisan, nomor *space group*, dan simetri inversi. Analisis tidak melibatkan informasi struktur atom secara eksplisit. Model Random Forest dan XGBoost digunakan untuk memodelkan hubungan nonlinear antara fitur fisik global dan *band gap* pada rentang 0.5 sampai 3.0 eV. Random Forest menghasilkan nilai MAE sebesar 0.399 eV, RMSE 0.520 eV, dan R^2 sebesar 0.407. XGBoost menghasilkan MAE sebesar 0.392 eV, RMSE 0.524 eV, dan R^2 sebesar 0.397. Nilai ini menunjukkan bahwa kedua model mampu menjelaskan sebagian variasi *band gap* dalam dataset, meskipun belum sepenuhnya menangkap kompleksitas sifat elektronik material. Analisis interpretabilitas menggunakan SHAP menunjukkan bahwa energi pembentukan, energi terhadap *convex hull*, dan ketebalan lapisan memberi kontribusi terbesar terhadap prediksi *band gap*. Fitur-fitur tersebut merepresentasikan aspek kestabilan termodinamika dan karakter fisik global material 2D yang secara statistik berkorelasi dengan variasi *band gap*, tanpa mengklaim hubungan sebab-akibat langsung pada tingkat struktur pita elektronik. Hasil ini menunjukkan bahwa *machine learning* dapat digunakan sebagai alat praktis untuk analisis awal dan penyaringan material 2D semikonduktor. Pendekatan ini membantu mengidentifikasi tren global dalam data DFT dan mempersempit ruang pencarian kandidat material sebelum dilakukan perhitungan fisika komputasi lanjutan.

Kata kunci: *band gap*, *machine learning*, material 2D, Random Forest, XGBoost.

ABSTRACT

PREDICTING *BAND GAP*S OF 2D MATERIALS USING RANDOM FOREST AND XGBOOST WITH SHAP-BASED INTERPRETABILITY ON THE C2DB DATASET

Oleh:

Iren Yolanda Sitompul

21/481208/PA/20942

This study predicts the Kohn-Sham band gap of two-dimensional semiconductors using a machine learning approach based on Density Functional Theory data from the Computational 2D Materials Database. The analysis uses global physical features available in the dataset, including total energy per atom, formation energy, energy above the convex hull, material thickness, space group number, and inversion symmetry. The study does not use explicit atomic structure information. Random Forest and XGBoost models capture the nonlinear relationship between global physical features and the band gap in the range of 0.5 to 3.0 eV. The Random Forest model achieves a mean absolute error of 0.399 eV, a root mean square error of 0.520 eV, and an R^2 value of 0.407. The XGBoost model achieves a mean absolute error of 0.397 eV, a root mean square error of 0.524 eV, and an R^2 value of 0.394. These results show that both models explain part of the band gap variation in the dataset, but they do not fully capture the complexity of electronic properties. SHAP-based interpretability analysis shows that formation energy, energy above the convex hull, and material thickness contribute most to the band gap predictions. These features represent thermodynamic stability and global physical characteristics of 2D materials that statistically correlate with band gap variation, without implying a direct causal link at the level of electronic band structure. The results show that machine learning serves as a practical tool for early-stage analysis and pre-screening of 2D semiconductors. This approach helps identify global trends in DFT data and narrows the search space for candidate materials before more detailed computational studies.

Keywords: 2D materials, band gap, machine learning, Random Forest, XGBoost.