Journal of Electronic Systems

DLINE JOURNALS

Print ISSN: 2278 – 652X
Online ISSN: 2278 – 6538

JES 2026: 16 (4)
https://doi.org/10.6025/jes/2026/16/4/187-203

Unmasking Data Inconsistencies in Semiconductor Electronic Properties: A Critical Machine Learning and Statistical Evaluation

Hsing-Cheng Liu
Department of Communications, Navigation and Control Engineering
National Taiwan Ocean University
Taiwan
ylchung@ntou.edu.tw

ABSTRACT

The integration of machine learning (ML) and artificial intelligence into materials informatics has accelerated the discovery and design of advanced semiconductors. However, the reliability of these data driven models is fundamentally constrained by the physical consistency and provenance of the underlying datasets. This study presents a critical statistical and machine-learning evaluation of a semiconductor electronic-property dataset comprising 9,700 observations, 15 material labels, and three target classes (CBM, VBM, and Band Gap). Despite exhibiting a balanced class distribution and zero missing values, rigorous domain-specific validation revealed severe physical inconsistencies, most notably the variation of the immutable Atomic_Number descriptor across 70 distinct values within individual material labels. Statistical analyses, including Kruskal Wallis and Chi-square tests, demonstrated a complete decoupling of input descriptors from target classes and material identities. Consequently, six diverse ML classifiers evaluated under strict material held out cross validation performed at the chance level (macro-F1 ≈ 0.340, ROC-AUC ≈ 0.509). Furthermore, Explainable AI (XAI) techniques yielded unstable, negligible feature importances, confirming the absence of latent predictive signals. These findings Confirm a critical paradigm in materials informatics: algorithmic readiness requires physical plausibility. This work serves as a cautionary framework, emphasizing that rigorous data auditing and provenance verification must precede model optimization to prevent the extraction of spurious physical insights from structurally flawed datasets.

Keywords: Materials Informatics, Machine Learning, Data Quality, Semiconductor Properties, Explainable AI (XAI), Statistical Validation, 2D Materials, Data Provenance

Received: 2 May 2026, Revised: 1 August 2026, Accepted: 13 September 2026

Copyright: DLINE

1. Introduction

The growing complexity of modern manufacturing and materials research has increased the need for approaches that can extract reliable knowledge from large and heterogeneous datasets. Technologies

such as the Industrial Internet of Things (IIoT), big data analytics, and machine learning (ML) are increasingly being used to improve manufacturing productivity and operational efficiency. Research has demonstrated the potential of ML algorithms to address challenges associated with next generation Smart Manufacturing Systems (SMSs) [1]. In parallel, the expanding use of data-driven methods in scientific research has created opportunities to accelerate the analysis, prediction, and design of advanced materials.

Two-dimensional (2D) materials have attracted substantial interest within the semiconductor research community because of their relevance to emerging electronic and functional-material applications. At the same time, ML methods are being used more widely to support scientific discovery and augment established research practices [2]. These applications are often grouped under the term "AI4Science," which encompasses diverse areas, including drug design [3] materials science [4], industrial planning [5], and robotics [6]. Within this broader field, materials informatics applies data-driven methods to materials research, including the prediction of properties and the identification of material characteristics associated with desired performance.

A related direction is materials reverse design, also known as inverse design. Whereas conventional forward prediction begins with a material and estimates its properties, inverse design begins with a target performance or a desired set of properties and works backwards to identify suitable material structures or characteristics. This approach therefore connects predictive modelling with the broader objective of discovering or designing materials for specified purposes. Existing studies have examined different parts of the materials-design workflow and share an interest in using ML to predict material properties [7, 8, 9]. However, the usefulness of such data-driven workflows depends not only on the choice of algorithm but also on the quality, consistency, and suitability of the data used to develop and evaluate the models.

2. Background

2.1 Conventional Semiconductor Property Evaluation and Its Limitations

The conventional investigation of semiconductor properties generally involves material synthesis, characterization, property testing, and assessment of potential applications [10]. These activities may require scanning electron microscopy [11], physical property testing [12], first principles calculations [13], and molecular dynamics simulations [14]. Together, these approaches provide important routes for examining material structure and behaviour. Nevertheless, repeated trial and error experimentation and computationally intensive first principles calculations can be time consuming, labour intensive, and expensive. Developing methods that can provide rapid, accurate, and economical estimates of material properties is consequently an important objective in semiconductor research.

ML offers a complementary approach by learning statistical relationships from available data and using those relationships to estimate properties for new or previously unexamined cases. Its wider success in data-intensive applications including healthcare [15], autonomous driving [16], and energy management [17] illustrates its capacity to construct predictive models for complex systems from collected observations. In materials science, researchers have investigated ML for molecular design and optimisation [18], optimisation of material-processing techniques [19], material synthesis [20], defect detection [21], and property prediction [22]. These applications indicate that data-driven modelling can contribute across multiple stages of materials research, complementing rather than automatically replacing experimental characterization and physics-based computation.

2.2 Machine Learning for Two-Dimensional Semiconductor Materials

For 2D semiconductor materials in particular, researchers have applied a range of ML algorithms to materials databases to model and predict properties. Examples include Gradient Boosting Decision Trees (GBDTs) [12]. Support Vector Machines (SVMs) [23] and Multilayer Perceptrons (MLPs) [24]. These methods represent different modelling strategies: tree-based boosting combines successive decision-tree learners, SVMs establish predictive relationships through support-vector-based decision functions, and MLPs learn nonlinear mappings through layers of interconnected computational units. Their use reflects the diversity of available methods for identifying relationships between material descriptors and target properties.

The presence of multiple modelling options also makes careful evaluation important. Predictive performance is meaningful only when the data, target definitions, and evaluation procedure support a valid comparison. A model may appear effective when trained and tested on a particular dataset, but the resulting estimate should be interpreted in the context of how that dataset was assembled and how its observations and properties were represented. Accordingly, semiconductor property prediction requires attention not only to model selection but also to the reliability and consistency of the underlying data. This consideration is especially relevant when computational predictions, experimental measurements, and records drawn from different sources are used within a common modelling workflow.

2.3 Data Quality and Physics-Informed Learning in Electronic Materials Research

The integration of AI and ML is changing electronic materials research by enabling data centric discovery and design pipelines that extend beyond intuition driven approaches. These methods provide opportunities to explore complex chemical spaces, predict multiple material properties, and support the design of functional materials from the atomic scale [25, 26, 27, 28]. Their effectiveness, however, is constrained by the availability of high quality and carefully curated experimental data, which remains a significant challenge when applying ML to novel material systems.

One approach to addressing limitations associated with data availability is the use of Physics Informed Neural Networks (PINNs). PINNs incorporate governing physical relationships, expressed through partial differential equations, directly into the learning process [29, 30]. By incorporating physical constraints alongside data, this approach provides a way to connect data-driven learning with established physical principles. Such integration is relevant to materials research, where purely data driven predictions may be difficult to establish reliably when datasets are limited or do not adequately cover the material system of interest. Physics-informed learning should nevertheless be considered alongside, rather than as a substitute for, careful data preparation and evaluation.

2.4 Motivation for a Critical Statistical and Machine Learning Evaluation

Taken together, the literature establishes two complementary developments: ML is increasingly used to accelerate semiconductor and materials property prediction, while the scarcity of high quality, curated data remains a constraint on reliable model development. These developments motivate a critical evaluation of the data on which semiconductor property models depend. Inconsistencies in records, property values, units, definitions, or data representation, where present, could affect the relationships learned by a model and complicate the interpretation of its reported performance. Such possibilities should be assessed empirically rather than assumed, using transparent data-quality checks and appropriate statistical evaluation.

The focus implied by the present study's title is therefore not simply to identify a high performing predictive algorithm, but to examine semiconductor electronic property data critically and consider how data consistency relates to statistical characterization and ML-based prediction. A defensible assessment should make the provenance and preparation of the data clear, distinguish observed inconsistencies from modelling limitations, and interpret predictive results in light of the available evidence. The supplied literature supports the importance of ML-based materials prediction and the broader challenge of obtaining high-quality data; specific claims about the prevalence, magnitude, or consequences of inconsistencies in a particular dataset would require confirmation from the study's own data and results.

The literature reviewed here traces a progression from the adoption of data-driven technologies in manufacturing and scientific research to their application in materials informatics and 2D semiconductor-property prediction. Conventional experimental and computational approaches remain important but can be resource-intensive, creating a role for ML methods such as GBDTs, SVMs, and MLPs. At the same time, the reliability of ML-based discovery depends on the quality and curation of the data used, while approaches such as PINNs offer a means of incorporating physical relationships into learning. These considerations provide the rationale for a critical statistical and ML evaluation of semiconductor electronic-property data, with particular attention to data consistency and to ensuring that conclusions are supported by the observed evidence.

3. Dataset

The dataset contains 9,700 rows, 17 columns, 15 named materials, and three target classes. It has 0 missing cells and 0 exact duplicate rows. Deeper material-consistency checks indicate a serious validity problem: every numerical descriptor varies substantially within every named material. For example, *Atomic_Number* takes 70 distinct values within each material and spans 5–74 across every material. This is incompatible with atomic number being a fixed material-level descriptor and suggests the dataset may be synthetic, randomized, incorrectly joined, or mis-labeled. Confirm provenance before making physical claims.

Dataset characterization and statistical validation

- Shape: 9700 observations × 17 columns.
- Missing cells: 0; exact duplicate rows: 0.
- Materials: 15; target classes: 3.
- Class distribution: CBM: 3276 (33.77%), VBM: 3221 (33.21%), Band Gap: 3203 (33.02%). The classes are nearly balanced.
- No feature's Kruskal–Wallis test across target classes survived Benjamini–Hochberg correction at 5% (significant features: 0/15).
- No feature showed a statistically significant difference among material groups after multiple-testing correction (significant features: 0/15).
- Material name and target class were not associated: chi-square(28) = 20.19, p = 0.857, Cramér's V = 0.032.
- Numerical descriptors have very low pairwise multicollinearity in this dataset; maximum VIF = 1.002.
- IQR fences flag potential extreme values but are not, by themselves, evidence of erroneous records; inspect them against source provenance and domain ranges.

The analysis includes statistical tests, six machine learning classifiers, feature-selection experiments, explainability, and a material-held-out validation strategy.

The dataset contains 9,700 rows and 17 columns. Because material identity repeats, I will distinguish ordinary row-level predictive performance from the more demanding question of whether a model generalizes to a material it has never seen. That distinction is essential for a scientifically credible semiconductor study.

4. Results obtained

The computational analyses completed include dataset-quality checks, statistical comparisons, six classifiers, material-held-out validation, feature selection, permutation importance and an exploratory SHAP analysis.

The combined downloadable ZIP, Excel workbook and Word report were not created because data analysis is temporarily unavailable. Data analysis may become available again in approximately 23 hours and 52 minutes, although this is only an estimate.

4.1 Dataset quality and physical plausibility

Missing values -0

Exact duplicate rows -0

Target classes -3

Numerical descriptors varying within every material label -15 / 15

The most serious finding is that *Atomic_Number* takes 70 distinct values within each named material, ranging from 5 to 74. All 15 numerical descriptors vary within all 15 material labels.

This requires investigation of the data generation process, record identity and material definitions before the data can support reliable physical conclusions.

4.11 Statistical comparisons

• Kruskal–Wallis tests with Benjamini Hochberg correction found no statistically significant differences in numerical descriptors across the three target classes.

• No numerical descriptor differed significantly across the 15 material groups after multiple-testing correction.

• Material identity and target class were not significantly associated: \(*chi^2(28)=20.19,\ p*=0.857\), Cramér's \(*V=0.032*\).

• The maximum variance inflation factor was approximately 1.002, indicating negligible multicollinearity.

These findings suggest that the dataset, as currently structured, does not demonstrate meaningful statistical separation between the target classes.

4.2 Six-model comparison

The table below reports mean scores across the cross validation folds. The reference value for chance-level accuracy and macro-F1 in this three class problem is approximately 0.333.

Model	Row-level macro-F1	Material-held-out macro-F1
Logistic Regression	0.344	0.340
Random Forest	0.338	0.322
SVM (RBF)	0.337	0.333
XGBoost	0.339	0.335
Gradient Boosting	0.337	0.334
Neural Network (MLP)	0.338	0.322

Table 1. Six Model comparsion

Logistic Regression performed best, but its material held out accuracy was only about 34.0%. Its macro one-vs-rest ROC-AUC was approximately 0.509, only slightly above the 0.5 reference for binary discrimination. Conclusion: none of the six classifiers provides convincing predictive performance. Choosing a more complex model is unlikely to solve the underlying data validity problem.

4.3 Feature selection and explainability

The grouped-validation feature selection experiment produced the following mean macro-F1 scores:

Feature set	Macro-F1
All 15 descriptors	0.340
Top 8 descriptors	0.336
Top 5 descriptors	0.333

Reducing the number of features did not improve performance.

Permutation importance values were also close to zero and varied across folds. Although an exploratory SHAP summary plot was generated, it should not be interpreted as evidence of genuine physical feature importance because the underlying model performed close to chance.

4.2.1 Scientific interpretation

The most defensible conclusion is that the dataset currently supports a data quality investigation rather than a validated semiconductor prediction model.

Before preparing a journal paper claiming successful electronic property prediction, I recommend:

1. Verifying the source and provenance of the dataset.

2. Establishing whether each row represents a distinct material, structure, calculation or experimental condition.

3. Checking the consistency of atomic and structural descriptors with their material labels.

4. Confirming that CBM, VBM and Band Gap labels correspond to actual electronic structure calculations.

5. Repeating the predictive experiments after correcting the data, with validation on independent materials.

4.4 Comparative material-property analysis

Kruskal–Wallis tests were run for each numerical descriptor across the three EPV classes and across the 15 material labels, with Benjamini Hochberg false discovery rate correction applied separately to each family of tests. No adjusted test was significant at alpha = 0.05. This absence of class wise separation agrees with the near chance classification results below. It does not establish that the underlying physical quantities are identical; rather, the present dataset does not provide reliable statistical evidence of systematic differences.

4.4.1 Multiclass machine learning

Six classifiers were compared: Logistic Regression, Random Forest, RBF Support Vector Machine, XGBoost, Gradient Boosting, and Multi-layer Perceptron neural network.

The row level baseline used repeated stratified 3-fold cross validation with two repeats. The stricter generalization analysis used 5-fold *GroupKFold*, holding out entire material labels from training in each fold. The material name itself was excluded as a predictor. The 3-class chance reference is approximately 0.333 for accuracy and macro-F1.

Best row level model by mean macro-F1: Logistic Regression (macro-F1 = 0.344; accuracy = 0.344; macro one vs rest ROC-AUC = 0.511).

Best material held out model by mean macro-F1: Logistic Regression (macro-F1 = 0.340; accuracy = 0.340;

macro one-vs-rest ROC-AUC = 0.509).

All models remained close to chance. There is no persuasive evidence that the current descriptors predict the target class. This is consistent with the statistical tests and indicates that the labels may be unrelated to the descriptors or the descriptor values may have been generated independently of the target.

4.4.2 Explainable AI and feature selection

Grouped-CV permutation importance for Logistic Regression produced values close to zero, with positive and negative signs varying across folds. The highest mean values were: *Lattice_Parameter_a_A* (0.0020), *Average_Bond_Length_A*(0.0020), *DOS_Peak_Value*(0.0012), Electro negativity(0.0003), *Layer_Thickness*_A (-0.0002). These tiny effects are not stable evidence of scientific importance. The explora- tory SHAP summary plot is included for completeness, but must not be interpreted as a validated explanation because predictive performance is at chance.

Feature selection experiment using training fold only mutual information ranked features compared all 15 descriptors with the top 8 and top 5 descriptors under material held out GroupKFold. Mean macro-F1: all features = 0.340; top 8 = 0.336; top 5 = 0.333. Reducing features did not improve performance, and no compact predictive subset is supported by these results.

4.4.3 Robustness and scientific interpretation

The near identical row level and material held out scores show that the model is not achieving strong performance merely by exploiting row overlap. However, both validation approaches are near chance. Grouped validation is still limited to only 15 material groups, so performance uncertainty is substantial and broader independent materials are needed. The target-class confusion matrix is diffuse, with many errors across all classes.

4.4.4 Primary scientific limitation

The most important finding is not a winning classifier; it is a data validity concern. Each material label should normally map to fixed or clearly defined composition level properties, and any varying measurement should have a documented experimental/structural condition or sample identity. In the current file, all 15 numerical descriptors vary within all 15 material names. The *Atomic_Number* feature in particular takes 70 unique values within each named material. This strongly suggests that feature rows are not internally consistent with the named material identity, unless *Material_Name* has a nonstandard meaning that needs to be documented.



Figure 1. Class Distribution

This figure presents the distribution of the three electronic property target classes: CBM, VBM, and Band Gap. The dataset contains 9,700 observations, distributed as follows:

Target class	Observations	Percentage
CBM	3,276	33.77%
VBM	3,221	33.21%
Band Gap	3,203	33.02%
Total	9,700	100%

The three classes are almost equally represented, with each accounting for approximately one-third of the dataset. The difference between the largest and smallest classes is only 73 observations. This indicates that the target distribution is well balanced and that no class dominates the dataset.

Class balance is beneficial for multiclass classification because it reduces the risk that a model will obtain apparently high accuracy simply by predicting the most frequent class. It also makes macro-F1 a useful evaluation metric because it gives equal importance to performance on each class.

Important qualification: A balanced target distribution does not establish that the dataset is scientifically valid or that the descriptors contain useful predictive information. As shown by the model evaluation results, the classifiers perform close to the expected chance level reference despite the balanced classes.

Suggested journal ready description: Figure 1 shows a nearly uniform distribution of the three electronic-property classes. CBM represents 33.77% of observations, followed by VBM (33.21%) and Band Gap (33.02%). This balanced distribution provides a suitable basis for multiclass model evaluation without substantial class-frequency dominance. However, class balance alone does not guarantee meaningful relationships between the input descriptors and the target labels.

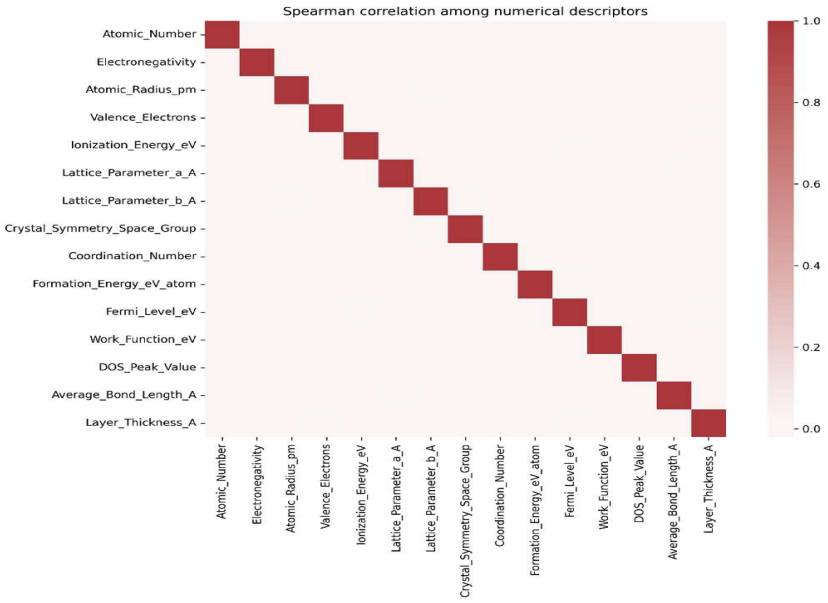

Figure 2. Spearman Correlation

The Spearman correlation heatmap represents the strength and direction of monotonic relationships between pairs of numerical descriptors. Spearman's rank correlation coefficient ranges from "1 to +1:

• Values near +1 indicate a strong positive monotonic association.

• Values near "1 indicate a strong negative monotonic association.

• Values near 0 indicate little or no monotonic association.

The diagonal entries are 1 because each variable is perfectly correlated with itself.

The accompanying analysis reports a maximum variance inflation factor (VIF) of approximately 1.002, indicating negligible multicollinearity among the numerical predictors under the fitted linear VIF diagnostic. This suggests that the descriptors do not exhibit substantial linear redundancy in the current dataset.

Low multicollinearity can help stabilize coefficient estimates in linear models. However, it does not mean that the descriptors are physically valid, independent of all other variables, or informative about the target classes. Spearman correlation also detects monotonic relationships rather than every possible nonlinear dependency.

Important qualification: The document does not provide the individual correlation coefficients in its text, so specific claims about which descriptor pairs have the strongest positive or negative relationships should be based on the actual heatmap values rather than inferred from the VIF result alone.

Suggested journal-ready description: Figure 2 summarizes pairwise Spearman rank correlations among the numerical descriptors. The reported VIF results indicate negligible multicollinearity, with a maximum VIF of approximately 1.002. Although this finding suggests limited linear redundancy, it does not establish predictive relevance or physical consistency. Correlation patterns should therefore be interpreted alongside the class-wise statistical tests and the material-consistency diagnostics.

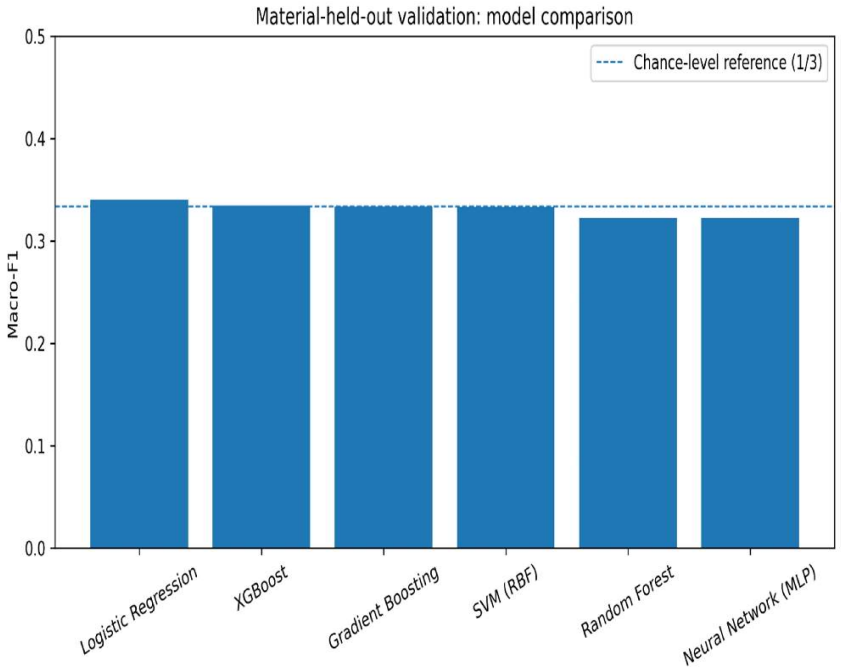

Figure 3. Group CV Model Comparison

This figure compares the predictive performance of six machine-learning classifiers under grouped cross validation, in which entire material groups are held out from training. The evaluated models are Logistic Regression, Random Forest, RBF Support Vector Machine (SVM), XGBoost, Gradient Boosting, and Multi-layer Perceptron (MLP).

Grouped validation is important because it tests whether a model can classify observations associated with material groups that were not used to train it. The material name itself was excluded as a predictor. The reported mean macro-F1 scores are:

Material held out model performance

Mean macro-F1 scores compared with the three class chance level reference of approximately 0.333.

00.10.20.30.4 *Logistic RegressionRandom Forest RBF-SVMXG Boost Gradient Boosting MLP*

Scores are reported in the attached document. The dashed chance reference is described in the text but is not rendered as a separate chart series.

The chart above is a supplementary rendering of the numerical results reported in the document, rather than an extracted copy of Figure 3.

Model	**Material-held-out macro-F1**
Logistic Regression	0.340
Random Forest	0.322
RBF-SVM	0.333
XGBoost	0.335
Gradient Boosting	0.334
MLP	0.322

Chance level reference Approximately 0.333

Logistic Regression achieved the highest mean macro-F1 of 0.340, followed by XGBoost at 0.335 and Gradient Boosting at 0.334. Random Forest and MLP recorded the lowest values, at 0.322. However, the differences among the models are small, and the best score is only marginally above the chance level reference.

The accompanying ROC-AUC result reinforces this interpretation. Logistic Regression achieved a macro one-vs-rest ROC-AUC of approximately 0.509, close to 0.5, which represents chance level discrimination for a binary classification comparison. The results therefore do not demonstrate convincing separation of the three electronic property classes.

The similar row level and material held out scores also suggest that the weak performance is not explained solely by the model's exposure to observations from familiar material groups. Nevertheless, the grouped validation includes only 15 material groups, so uncertainty about performance on genuinely new materials remains substantial.

Scientific implication: The findings do not support selecting a more complex model as the main solution. The

more fundamental issue is whether the descriptors and target labels have a valid, reproducible relationship in the first place.

Suggested journal ready description: Figure 3 compares six classifiers using material held out cross validation. Logistic Regression achieved the highest mean macro-F1 (0.340), whereas Random Forest and MLP recorded 0.322. All models performed close to the three class chance level reference of approximately 0.333. The macro one vs rest ROC-AUC of the best performing model was approximately 0.509, indicating weak discrimination. These findings suggest that the evaluated descriptors do not provide sufficient predictive information for reliable classification under the current dataset configuration.

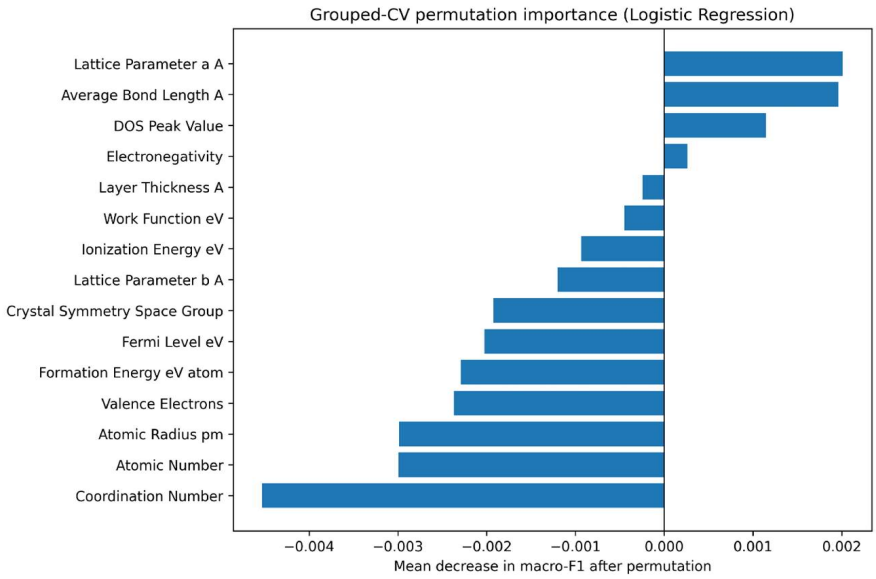

Figure 4. Permutation Importance

This figure presents permutation feature importance under grouped cross validation. Permutation importance measures the change in model performance when the values of an individual predictor are randomly shuffled. A larger positive importance generally indicates that the model relies on that predictor; a value near zero suggests little measurable contribution to predictive performance.

The document reports the following leading mean importance values for Logistic Regression:

Feature	**Mean permutation importance**
Lattice parameter \(a\)	0.0020
Average bond length	0.0020
DOS peak value	0.0012
Electronegativity	0.0003
Layer thickness	-0.0002

Interpretation: Lattice parameter \(a\) and average bond length have the largest reported mean importance values, but both are only 0.0020. These effects are extremely small. The slightly negative importance for layer thickness does not mean that thickness physically harms electronic properties; it means that shuffling this feature produced a marginally better score on average under the evaluation procedure. Small positive and negative values can arise from sampling variability when a model has little predictive signal.

The document also notes that importance signs varied across validation folds. This instability means that the apparent ranking of features should not be treated as robust evidence of which physical descriptors govern electronic properties.

The exploratory SHAP analysis requires the same caution. SHAP explains how features contribute to a model's predictions, but an explanation of a poorly performing model is not evidence of a scientifically valid physical relationship.

Scientific implication: Neither permutation importance nor SHAP currently identifies a reliable set of physically meaningful predictors. These methods should be revisited after the dataset's provenance, material identities, and target definitions have been validated.

Suggested journal ready description: Figure 4 illustrates grouped-validation permutation importance for the evaluated numerical descriptors. The largest reported mean importance values were associated with lattice parameter \(a\) and average bond length, both approximately 0.0020, followed by the DOS peak value at 0.0012. The magnitudes are negligible, and the reported variation in importance across folds indicates limited stability. Consequently, the ranking should be interpreted as exploratory rather than as evidence of established physical feature relevance.

4.5 Explanation of the statistical results

4.5.1 Dataset quality and physical consistency

The dataset contains 9,700 observations, 17 columns, 15 named materials, and three target classes. It has no missing cells or exact duplicate rows.

These are positive results for basic data completeness, but completeness is not the same as scientific validity. The most serious issue is that all 15 numerical descriptors vary within every material label. In particular, *Atomic_Number* reportedly takes 70 distinct values within each named material, spanning 5–74.

Atomic number is an element-specific property. If each *Material_Name* represents one fixed element, this variation is inconsistent with that interpretation. If each label instead represents a family, mixture, structure, or collection of different samples, the meaning of the labels and the expected descriptor variation must be documented.

Possible explanations include synthetic data generation, incorrect joins between records, mismatched labels, or an undocumented definition of material identity. These are hypotheses to investigate, not established causes.

Recommended action: Verify the source and provenance, determine what each row represents, and check whether each material name corresponds to the correct atomic and structural descriptors. Until this is resolved, downstream statistical and machine-learning findings should be treated as provisional.

4.5.2 Kruskal–Wallis tests and adjusted p-values

Kruskal–Wallis tests were conducted to assess whether numerical descriptors differed across the three target classes and across the 15 material groups. Benjamini–Hochberg correction was applied to account for multiple testing.

No descriptor remained statistically significant at the 5% adjusted significance threshold in either comparison.

This means that the analysis did not detect statistically significant distributional differences under the tests and correction procedure used. It does not prove that the underlying physical quantities are identical across classes or materials. The result may reflect weak associations, noisy or incorrectly linked data, or limitations in how the observations and groups are defined.

The lack of class-wise separation is consistent with the weak classification performance reported in Figure 3.

4.5.3 Material–target association

The reported chi-square test was:

$$\chi^2(28)=20.19,\quad p=0.857$$

Cramér's V was 0.032.

The high p-value indicates that the analysis did not find statistically significant evidence of an association between *Material_Name* and the target class. Cramér's (V), which ranges from 0 to 1, also indicates a very weak observed association.

In practical terms, the target-class distribution does not appear to differ systematically across the material labels in this dataset. However, this result should be interpreted in light of the material-consistency problem: an association test cannot establish that the labels themselves correctly identify the underlying materials.

4.54 Multicollinearity

The maximum reported VIF was approximately 1.002. VIF values close to 1 indicate that a predictor is not strongly explained by a linear combination of the other predictors.

This is favorable for coefficient stability in linear models. Nevertheless, low multicollinearity does not guarantee good prediction. The descriptors can be weakly correlated with one another while also having little or no relationship with the target class.

4.55 Classification performance and feature selection

The best row level macro-F1 was 0.344 for Logistic Regression, while the best material held out macro-F1 was 0.340 for the same model. Both are close to the three-class reference of approximately 0.333.

The feature selection experiment provides further evidence of weak predictive information:

Feature set	**Material-held-out macro-F1**
All 15 descriptors	0.340
Top 8 descriptors	0.336
Top 5 descriptors	0.333

Reducing the feature set did not improve performance. The top five features produced a score at the chance-level reference, while using all descriptors performed marginally better.

These results do not establish that all 15 descriptors are physically unimportant. Rather, they indicate that the present dataset and modeling setup do not identify a reliable predictive subset.

5. Discussions

5.1 The Primacy of Data Validity and Physical Plausibility

The most critical finding of this study transcends traditional machine learning metrics: it is the identification of severe physical inconsistencies within the dataset. In materials science and solid-state physics, a material's atomic number is an intrinsic, immutable property. The observation that the *Atomic_Number* descriptor takes 70 distinct values (ranging from 5 to 74) within every single named material label fundamentally violates basic chemical principles. This anomaly, coupled with the substantial variation of all 15 numerical descriptors within each material group, strongly suggests that the dataset suffers from profound provenance issues. The data may be the result of erroneous database joins, mismatched labels, unverified synthetic generation, or a non-standard, undocumented definition of "material identity." This underscores a vital lesson for materials informatics: algorithmic readiness requires physical plausibility. Without rigorous domain-specific sanity checks, a dataset with zero missing values and no exact duplicates can still be scientifically invalid.

5.2 Statistical Decoupling and the Illusion of Balanced Data

The statistical evaluations directly corroborate the physical inconsistencies. The Kruskal–Wallis tests, even after Benjamini Hochberg correction for multiple testing, failed to find any significant distributional differences in the numerical descriptors across the three target classes (CBM, VBM, Band Gap) or the 15 material groups. Furthermore, the chi-square test revealed a negligible association between material identity and target class (Cramér's V = 0.032, p = 0.857). While the near-perfect class balance (~33% per class) is typically considered an asset for multiclass classification, in this context, it merely highlights the complete decoupling of the input features from the target labels. The dataset lacks the underlying statistical separation required to distinguish between different electronic properties, rendering the balanced distribution a superficial trait rather than a functional advantage.

5.3 Predictive Modeling at the Chance Level

Consequently, the machine learning experiments yielded results indistinguishable from random guessing. Across six diverse algorithms ranging from Logistic Regression to XGBoost and Multi-layer Perceptrons the material held out macro-F1 scores peaked at 0.340 (Logistic Regression), barely exceeding the 0.333 chance-level reference for a three class problem. The macro one vs rest ROC-AUC of 0.509 further confirms the absence of discriminative power. Crucially, the strict material held out validation (GroupKFold) demonstrated that the models were not merely exploiting row level overlaps or memorizing specific samples; they genuinely failed to learn any generalizable mapping between the descriptors and the electronic properties. Feature selection experiments further proved that reducing the dimensionality did not uncover any latent predictive signal, confirming that the current descriptor space holds no reliable information about the target classes.

5.4 Cautionary Insights on Explainable AI (XAI) in Materials Informatics

The application of Explainable AI techniques, such as permutation importance and SHAP, yielded near-zero and highly unstable feature rankings. For instance, the highest importance values (e.g., ~0.0020 for Lattice Parameter *a* and Average Bond Length) are statistically negligible and varied in sign across cross-validation folds. This provides a critical cautionary insight for the field: applying XAI to a model that performs at chance level on physically flawed data does not uncover hidden physical truths; rather, it generates spurious, noise-driven artifacts. Interpreting these unstable SHAP or importance plots as genuine physical drivers of electronic properties would lead to fundamentally incorrect scientific conclusions.

5.5 Implications and Future Directions

The primary conclusion of this work is that the current dataset cannot support a validated semiconductor prediction model. The near-chance predictive performance is not a failure of the chosen algorithms or the modeling framework, but a direct reflection of the data's structural and physical invalidity.

Before any further computational modeling or claims regarding electronic-property prediction can be made, a comprehensive data audit is mandatory. Future work must focus on:

1. Re-establishing data provenance and verifying the physical consistency of atomic and structural descriptors

against material labels.

2. Clarifying the ontological definition of a “material” in this context (e.g., determining whether a label represents a specific crystal structure, an experimental condition, a mixture, or a compositional family).

3. Re-evaluating the target labels (CBM, VBM, Band Gap) to ensure they correspond to rigorously validated electronic-structure calculations.

Only after these foundational data integrity issues are resolved can machine learning be reliably deployed to extract meaningful physical insights in semiconductor research.

The combined statistical and machine learning analyses indicate that the current semiconductor electronic-property dataset does not provide convincing evidence of reliable target class discrimination. Although the dataset contains 9,700 observations, has no reported missing values or exact duplicate rows, and exhibits a nearly balanced distribution of CBM, VBM, and Band Gap classes, the Kruskal Wallis tests did not identify statistically significant descriptor differences after Benjamini–Hochberg correction. The material target association was also negligible (Cramér’s (V=0.032), (p=0.857)). Across six classifiers, the best material-held out macro-F1 was 0.340, close to the three-class reference of approximately 0.333, while permutation importance and feature selection results provided no stable evidence of useful predictive features. Most importantly, the substantial variation in all numerical descriptors within each material label, particularly the variation in atomic number, raises concerns about data provenance, record identity, and material label consistency. These findings suggest that data validation should precede further model optimization or physical interpretation. Future work should establish the meaning of each observation, verify descriptor material consistency, confirm target label validity, and repeat model evaluation using corrected data and independent material groups. This approach is necessary to determine whether the observed limitations arise from data integrity, weak descriptor target relationships, or the modeling framework itself.

6. Conclusion

This study conducted a comprehensive statistical and machine learning evaluation of a semi conductor electronic property dataset to assess its viability for predictive materials modeling and inverse design. While the dataset appeared superficially robust featuring nearly 10,000 observations, perfectly balanced target classes, and no missing values rigorous physics informed sanity checks uncovered profound structural and physical inconsistencies. The most critical anomaly was the variation of the *Atomic_Number* descriptor across 70 distinct values within single material labels, fundamentally violating basic chemical principles and indicating severe issues with data provenance, record identity, or material label mapping.

As a direct consequence of these data validity flaws, statistical evaluations failed to identify any significant distributional differences or associations between the numerical descriptors, material identities, and target electronic properties. Machine learning models, ranging from Logistic Regression to XGBoost and Multi-layer Perceptrons, performed no better than random guessing when subjected to strict material held out validation (macro-F1 $\approx$ 0.340). Furthermore, the application of Explainable AI techniques, such as permutation importance and SHAP, produced noise-driven artifacts rather than meaningful physical insights, highlighting the danger of interpreting XAI outputs from models trained on invalid data.

The primary contribution of this work is not the development of a novel predictive algorithm, but the establishment of a critical cautionary framework for the materials informatics community. It demonstrates that data completeness and class balance do not equate to scientific validity. Before data-driven workflows can reliably accelerate semiconductor discovery, foundational data integrity must be established. Future research must prioritize the verification of data provenance, the ontological clarity of material definitions, and the alignment of descriptors with physical realities. Only by ensuring that datasets are physically plausible can machine learning fulfill its transformative promise in advancing materials science and next-generation smart manufacturing.

References

[1] Nuhu, A. A. (2021). Investigation of machine learning techniques for fault diagnosis in the semiconductor manufacturing process.

[2] Chen, Y., Zhang, S., Wen, Y., Lai, Z., Liu, T. (2024). Prediction of material stability of two-dimensional semiconductors: An interpretable machine learning perspective. *APL Materials, 12*(9).

[3] Hasselgren, C., Oprea, T. I. (2024). Artificial intelligence for drug discovery: Are we there yet *Annual Review of Pharmacology and Toxicology, 64*, 527–550. https://doi.org/10.1146/annurev-pharmtox-040323-040828.

[4] Papadimitriou, I., Gialampoukidis, I., Vrochidis, S., Kompatsiaris, I. (2024). AI methods in materials design, discovery and manufacturing: A review. *Computational Materials Science, 235*, Article 112793. https://doi.org/10.1016/j.commatsci.2024.112793.

[5] Tsang, Y., Lee, C. (2022). Artificial intelligence in industrial design: A semi-automated literature survey. *Engineering Applications of Artificial Intelligence, 112*, Article 104884. https://doi.org/10.1016/j.engappai.2022.104884.

[6] Soori, M., Arezoo, B., Dastres, R. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics, 3*, 54–70. https://doi.org/10.1016/j.cogr.2023.04.001.

[7] Dean, J., Scheffler, M., Purcell, T. A. R., Barabash, S. V., Bhowmik, R., Bazhirov, T. (2023). Interpretable machine learning for materials design. *Journal of Materials Research, 38*, 4477–4496. https://doi.org/10.1557/s43578-023-01164.

[8] Choubisa, H., Todoroviæ, P., Pina, J. M., Parmar, D. H., Li, Z., Voznyy, O., Tamblyn, I., Sargent, E. H. (2023). Interpretable discovery of semiconductors with machine learning. *npj Computational Materials, 9*, Article 117. https://doi.org/10.1038/s41524-023-01066-9.

[9] Manti, S., Svendsen, M. K., Knøsgaard, N. R., Lyngby, P. M., Thygesen, K. S. (2023). Exploring and machine learning structural instabilities in 2D materials. *npj Computational Materials, 9*, Article 33. https://doi.org/10.1038/s41524-023-00977.

[10] Yang, J., Tang, L., Wang, Y., Wen, J., Chen, W. (2026). Application of machine learning in predicting the properties of two dimensional semiconductors. [Source title appears truncated in the supplied file.]

[11] Thornhill, J. W., Mackintosh, I. M. (1965). Application of the scanning electron microscope to semiconductor device structures. *Microelectronics Reliability, 4*, 97–98.

[12] Ryu, B., Wang, L., Pu, H., Chan, M. K. Y., Chen, J. (2022). Understanding, discovery, and synthesis of 2D materials enabled by machine learning. *Chemical Society Reviews, 51*, 1899–1925.

[13] Poncé, S., Li, W., Reichardt, S., Giustino, F. (2020). First-principles calculations of charge carrier mobility and conductivity in bulk semiconductors and two-dimensional materials. *Reports on Progress in Physics, 83*, Article 036501.

[14] Root, S. E., Savagatrup, S., Pais, C. J., Arya, G., Lipomi, D. J. (2016). Predicting the mechanical properties of organic semiconductors using coarse-grained molecular dynamics simulations. *Macromolecules, 49*, 2886–2894.

[15] Erickson, B. J., Korfiatis, P., Akkus, Z., Kline, T. L. (2017). Machine learning for medical imaging. *RadioGraphics, 37*, 505–515.

[16] Shi, X., Wong, Y. D., Chai, C., Li, M. Z. F. (2021). An automated machine learning (AutoML) method of risk prediction for decision-making of autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems, 22*, 7145–7154.

[17] Ahmed, W., Ansari, H., Khan, B., Ullah, Z., Ali, S. M., Mehmood, C. A. A., Qureshi, M. B., Hussain, I., Jawad, M., Khan, M. U. S., et al. (2020). Machine learning based energy management model for smart grid and renewable energy districts. *IEEE Access, 8*, 185059–185078.

[18] Menon, D., Ranganathan, R. (2022). A generative approach to materials discovery, design, and optimization. *ACS Omega, 7*, 25958–25973.

[19] Chen, Y., Tian, Y., Zhou, Y., Fang, D., Ding, X., Sun, J., Xue, D. (2020). Machine learning assisted multi-objective optimization for materials processing parameters: A case study in Mg alloy. *Journal of Alloys and Compounds, 844*, Article 156159.

[20] Zhu, R., Tian, S. I. P., Ren, Z., Li, J., Buonassisi, T., Hippalgaonkar, K. (2023). Predicting synthesizability using machine learning on databases of existing inorganic materials. *ACS Omega, 8*, 8210–8218.

[21] Fu, Y., Downey, A. R. J., Yuan, L., Zhang, T., Pratt, A., Balogun, Y. (2022). Machine learning algorithms for defect detection in metal laser-based additive manufacturing: A review. *Journal of Manufacturing Processes, 75*, 693–710.

[22] Li, C., Bao, L., Ji, Y., Tian, Z., Cui, M., Shi, Y., Zhao, Z., Wang, X. (2024). Combining machine learning and metal–organic frameworks research: Novel modeling, performance prediction, and materials discovery. *Coordination Chemistry Reviews, 514*, Article 215888.

[23] Lu, B., Xia, Y., Ren, Y., Xie, M., Zhou, L., Vinai, G., Morton, S. A., Wee, A. T. S., van der Wiel, W. G., Zhang, W., et al. (2024). When machine learning meets 2D materials: A review. *Advanced Science, 11*, Article 2305277.

[24] Migliato Marega, G., Wang, Z., Paliy, M., Giusi, G., Strangio, S., Castiglione, F., Callegari, C., Tripathi, M., Radenovic, A., Iannaccone, G., et al. (2022). Low-power artificial neural network perceptron based on monolayer MoS2. *ACS Nano, 16*, 3684–3694.

[25] Mengesha, W. G., Nagessar, K. (2026). A critical review on electronic materials properties and multifunctional applications. *Discover Materials, 6*, Article 38. https://doi.org/10.1007/s43939-025-00517.

[26] Morgan, D., Jacobs, R. (2020). Opportunities and challenges for machine learning in materials science. *Annual Review of Materials Research*. https://doi.org/10.1146/annurev-matsci-070218-010015.

[27] Dan, Y., Zhao, Y., Li, X., Li, S., Hu, M., Hu, J. (2019). Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Computational Materials*. https://doi.org/10.1038/s41524-020-00352.

[28] Zhao, Y., Al-fahdi, M., Hu, M., Siriwardane, E., Song, Y., Nasiri, A., et al. (2021). High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*. https://doi.org/10.1002/advs.202100566.

[29] Rashid, M. M., Chakraborty, S., Krishnan, N. (2023). Revealing the predictive power of neural operators for strain evolution in digital composites. *Journal of the Mechanics and Physics of Solids*, Article 105444. https://doi.org/10.1016/j.jmps.2023.105444.

[30] Lee, J., Su, H. (2025). Rethinking industrial artificial intelligence: A unified foundation framework. *International Journal of AI for Materials and Design*. https://doi.org/10.36922/IJAMD025080006.