Machine learning classification of binary semiconductor heterostructures

Samir Rom,[1,*] Aishwaryo Ghosh,[1,*] Anita Halder,[1,2,*] and Tanusri Saha Dasgupta[1,†]
[1] *S.N. Bose National Centre for Basic Sciences JD Block, Sector III, Salt Lake, Kolkata 700106, India*
[2] *School of Physics, Trinity College Dublin, Dublin 2, Ireland.*

(Received 25 January 2021; revised 28 February 2021; accepted 24 March 2021; published 5 April 2021)

Heterostructures of two semiconductors are at the heart of semiconductor devices with tremendous technological importance. The prediction and designing of semiconductor heterostructures of a specific type is a difficult materials science problem, posing a challenge to experimental and computational investigations. In this study, we first establish that the prediction of heterostructure type can be made with good accuracy from the knowledge of the band structure of constituent semiconductors. Following this, we apply machine learning, built on features characterizing constituent semiconductors, on a known dataset of binary semiconductor heterostructures extended by a synthetic minority oversampling technique. A significant feature of engineering made it possible to train a classifier model predicting the heterostructure type with an accuracy of 89%. Using the trained model, a large number (872 number) of unknown heterostructure semiconductor types involving elemental and binary semiconductors is theoretically predicted. Interestingly, the developed scheme is found to be extendable to heterojunctions of semiconductor quantum dots.

DOI: 10.1103/PhysRevMaterials.5.043801

I. INTRODUCTION

Semiconductor heterojunction [1], which is an interface between two dissimilar semiconductors, is one of the central topic in semiconductor research due to its potential applications as light-emitting diodes [2], solar cells [3], photovoltaic devices [4], and so on. At the interface of two dissimilar semiconductors, bulk band structures of two semiconductors merge into each other, and an electronic transition region forms, involving band bending and band-edge discontinuities.

Semiconductor heterostructures in a general sense may be classified as type I or type II, depending on the signs of the band-edge discontinuities. The band-edge discontinuities, also known as valence band and conduction band offsets, naturally occur for pairs of semiconductors with different band gaps [5]. In a type-I heterostructure, the alignment of the bands makes the valence and conduction band offset of opposite signs, so that both conduction- and valence-band edges of semiconductor A (smaller band gap) are located within the energy gap of semiconductor B (larger band gap) [6]. The electron and hole pairs excited near the interface thus tend to be in semiconductor A. For a type II heterostructure, which includes both type-II-staggered and type-II-misaligned, on the other hand, the relative alignment of the conduction and valence bands, make the conduction and valence offsets to be of same sign. This results in a staggered alignment of bands, with optical transition energy smaller than the band gap of either of the constituent semiconductors and the lowest-energy states for the electrons and the holes lying in different semiconductors, a highly attractive situation for applications [7]. Successful designing of semiconductor heterostructures of specific type involve careful material selection, which is hard given the numerous possibilities. On the other hand, the advent and advancement in data mining and machine learning technology, have made them a natural choice for materials search with targeted properties in modern day materials science projects [8–10]. The empirical trial and error method and the first-principles method for materials designing, are limited by high cost/effort with low efficiency. Machine learning, which relies on pattern recognition, can substantially reduce the effort and shorten the development cycle.

In this study, we adopt a step by step procedure for prediction of semiconductor heterostructure type by combining database driven materials search and the machine learning technique. Note that the phase space of potential candidate materials is vast, which includes combinations between elemental, binary, ternary semiconductors, and also their alloys. Within the limited scope of present study, we exclusively focus on heterostructures formed by combining elemental and binary semiconductors.

For an ideal interface, the valence and conduction band offsets, responsible for deciding the heterostructure type, are expected to be determined by intrinsic material properties of the semiconductors in contact. The band-gap difference of the two semiconductors forming the junction fixes the sum of valence- and conduction-band offsets. What is harder to determine is the relative energy positions of valence-band maxima (VBM) and conduction-band minima (CBM) at the interface, known as "band alignment" [11]. While highly accurate method of band-gap calculations and related band alignment are available in calculations employing techniques such as many-body perturbation theory [12], hybrid exchange-correlation functional [13], time-dependent density function theory [14], or quantum Monte Carlo calculations [15],

[*] These authors contributed equally to this work.
[†] t.sahadasgupta@gmail.com

none of them are computationally cheap as in conventional density functional theory (DFT) calculations within local density or generalized gradient approximation [16] (GGA) of exchange-correlation, and thus not suitable for high throughput calculations. In the present study, we use existing, digitally accessible electronic-structure database [17], apply scissor shift [18] that rigidly shifts the conduction bands to produce accurate band gaps, use the band-structure information to calculate the branch point energy [19], to be used as a measure of common absolute energy level for band alignment. Such an approach has been demonstrated to be highly successful by validating against experiment and first-principles data [19], allowing for a fast screening of very large number of materials exclusively using electronic-structure data available in online databases. We further confirm the accuracy of this method in predicting heterostructure type by comparing with about 31 number of available experimental data, and first-principles calculation of selected heterostructures within the framework of hybrid calculations [20], which allow for complete structural and electronic reconstruction upon formation of the interface.

While the above analysis confirms prediction of heterostructure type from knowledge of constituents to be a rational approach, the next step would be to built a machine learning model for heterostructure-type prediction based on features characterizing band structure of constituent semiconductors. However, we find the size of the available dataset of heterostructures with known heterostructure type to be too small for application of machine learning (ML). We thus apply synthetic minority oversampling technique [21], which corrects for the class imbalance in the dataset as well as expands the dataset to a reasonable number of 78, to which machine learning may be applied. Using the expanded dataset as the training set, we finally construct our ML model using least absolute shrinkage and selection operator (LASSO) [22] with third-order polynomial fit for feature engineering for prediction of the type of heterostructure. Extension of the approach to nanoscale heterostructure, is found to successfully describe some of the available data. This makes us hopeful about the applicability of our developed algorithm in the prediction of heterostructure type both in bulk and nanoscale, which is a technologically important problem.

II. PREDICTION OF HETEROSTRUCTURE TYPE FROM KNOWLEDGE OF CONSTITUENT BAND STRUCTURES

Complete knowledge of the electronic structure of semiconductor heterostructure involves complicated structural details of the interface that are often unknown experimentally or expensive to compute in a high throughput scheme. It will be far simpler if predictions on heterostructure type can be made solely based on the information of the individual semiconductors, i.e., the band gaps of the semiconductors in contact and the relative energy alignment. If so, the machine learning would aim on accurately capturing the individual semiconductor band-structure properties.

In the experiment, band offsets can be measured using x-ray photoemission spectroscopy [23]. As a measure of zero of energy for aligning the band energies of semiconductor A and semiconductor B, the computed energy positions of atomic core-electron levels [24], electronic transition levels of hydrogen impurities [25], or vacuum levels for the materials in contact [26] can be used. This, however, requires large simulation cells with tens or hundreds of atoms. Understandably, while such expensive calculations can be carried out for specific individual interfaces, this cannot be followed in a high-throughput fashion, for large numbers of materials.

Another quantity which can be used as universal energy level for band alignment, assuming negligible interface dipoles, is the branch-point energy (E_{BP}), also referred to as charge neutrality level or effective midgap energy [12,27,28]. The E_{BP} can be entirely traced back to the bulk band structure of a given semiconductor, defined as [28]

$$E_{BP} = \frac{1}{2N_k}\sum_{\mathbf{k}}\left(\frac{1}{N_c}\sum_{c_i}^{N_c}\epsilon_{c_i}(\mathbf{k}) + \frac{1}{N_v}\sum_{v_i}^{N_v}\epsilon_{v_i}(\mathbf{k})\right), \qquad (1)$$

where N_k is the number of k-points and ϵ_v and ϵ_c are the valence and conduction electron energy eigenvalues, N_c and N_v being the number of conduction and valence bands, respectively. Here we specifically rely on the band-structure information from the Materials Project [29] which contains band structures for at least 66 676 materials, with 45 148 band structures of semiconductors and insulators. We, however, need to remember that band structures obtained from the Materials Project suffer from the band-gap problem of DFT. To correct for this, we resort to scissor shift [19] that rigidly shifts conduction bands to produce the experimental band gap, as reported in the literature. In cases where the experimental reports are not available, we employ linear band-gap correction [18]. E_{BP} is shifted by half the band-gap correction. The comparison of VBO and CBO predicted following the above scheme with the experimentally reported offset values show a rather good correspondence (cf. Fig. 1) with correlation value of 0.98 and 0.83 for VBO and CBO, respectively. We notice the superior performance in case of VBO, as expected taking into account the difficulties with the prediction of the band gap.

Gaining the confidence on predicted VBO and CBO values based on the band-structure input of constituent semiconductors, we next construct a heterostructure heatmap, as shown in Fig. 2, containing combinations of all known elemental and binary semiconductors, classified as type I and type II. Out of the vast possibilities of 903 different heterostructures, only few (31) heterostructures have been made with their types known. We find a perfect agreement between the predicted heterostructure type and that obtained from measurement or detailed calculation in all these 31 cases. Out of the remaining 872 combinations, based on band-structure information of components, 348 (~40%) combinations are expected to be type I and 524 (~60%) to be type II (highlighted with orange and red colours in Fig. 2), opening up different possibilities of achieving both the types in heterostructures yet-to-be explored.

This exercise establishes the fact that for prediction of heterostructure type, the knowledge of the constituent's band structure is sufficient. In the machine learning study, to be taken up in the following, we thus build the machine learning model based on the features of the individual semiconductors.

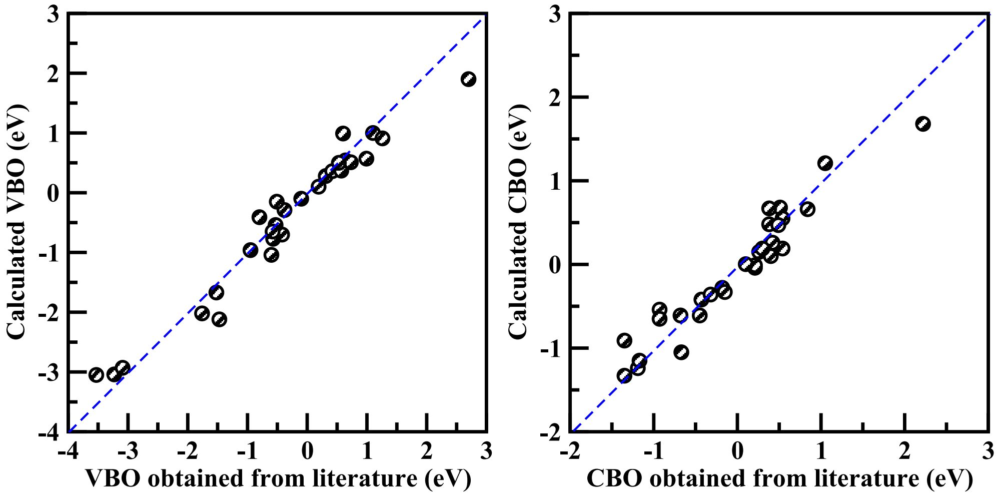

FIG. 1. Comparison of valence band offset (VBO) (left) and conduction band offset (CBO) values reported in the literature and that obtained from knowledge of corrected DFT band gap and branch point energy.

III. MACHINE LEARNING

In recent time, machine learning has become one of the popular approach in condensed matter physics and materials science for the prediction of target property in a cheap, yet accurate manner. In the context of the band-gap prediction, Lee *et al.* [30] used machine learning for the prediction of band gaps of 156 number of binary compounds. Pilania *et al.* [31] used kernel-ridge regression to predict the band gaps of 1306 double perovskites. A 16-dimensional set of element-specific descriptors was used for this purpose. Ward *et al.* [32] used a large set of 140 universal descriptors to predict band gaps that was used for identification of new solar absorbers. Weston *et al.* [33] used machine learning to study the band-gap properties of quaternary semiconductors. However, to the best of our knowledge, the classification of semiconductor heterostructures as either type I or type II has not been attempted from a machine learning perspective.

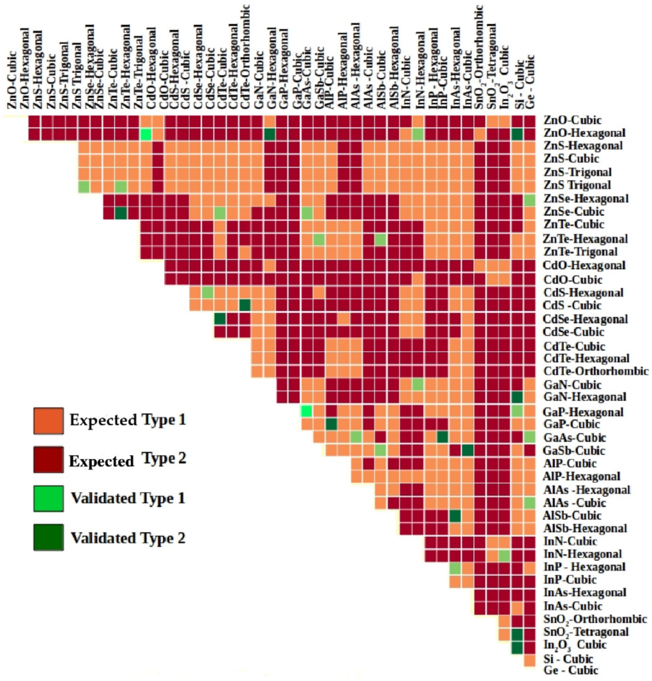

FIG. 2. Elemental and binary semiconductor heterostructures, characterized as type I (orange) or type II (red), based on band-structure knowledge of constituent semiconductors. Heterostructures that have been synthesized or calculated with types known and validated against the prediction are colored as light green (type I) and dark green (type II).

The steps followed in machine learning of the present study is schematically shown in Fig. 3, which starts with the construction of the dataset of known bulk semiconductor heterostructures along their type classification, the expansion of the dataset by creating synthetic data based on a minority oversampling technique, feature space engineering, selection of best regression model based on cross-validated error, conversion of regression model to a binary classification model (type I/type II), and finally to type prediction.

A. Experimental literature on elemental and binary semiconductor heterostructure

Exhaustive literature search for heterostructure of elemental and binary semiconductors with classified heterostructure type results in 31 heterostructures, AlP-GaP [34], AlSb-ZnTe [35], GaAs-AlAs [34], GaSb-AlSb [34], Ge-AlAs [36], Ge-GaAs [36], Ge-ZnSe [36], InN-GaN [5], InN-ZnO [37], InP-GaAs [38], InP-InAs [38], Si-GaP [35], ZnSe-GaAs [36], CdS-CdSe [39], CdS-CdTe [39], CdSe-CdTe [39], GaN-Si [40], GaN-ZnO [41], Si-In$_2$O$_3$ [42], Si-ZnO [43], Si-SnO$_2$ [43], ZnSe-CdTe [44], ZnSe-ZnTe [45], InN-In$_2$O$_3$ [46], AlN-Si [40], GaN-AlN [28], ZnS-ZnSe [45], ZnS-ZnTe [45], InN-AlN [47], ZnTe-GaSb [35], and AlSb-InAs [48]. AlSb-ZnTe, GaAs-AlAs, GaSb-AlSb, Ge-AlAs, Ge-GaAs, Ge-ZnSe, InN-GaN, InN-ZnO, InP-InAs, Si-GaP, ZnSe-GaAs, CdS-CdSe, ZnSe-CdTe, InN-In$_2$O$_3$, AlN-Si, GaN-AlN, ZnS-ZnSe, ZnS-ZnTe, InN-AlN, ZnTe-GaSb are characterized as type-I heterostructures, the rest being type II.

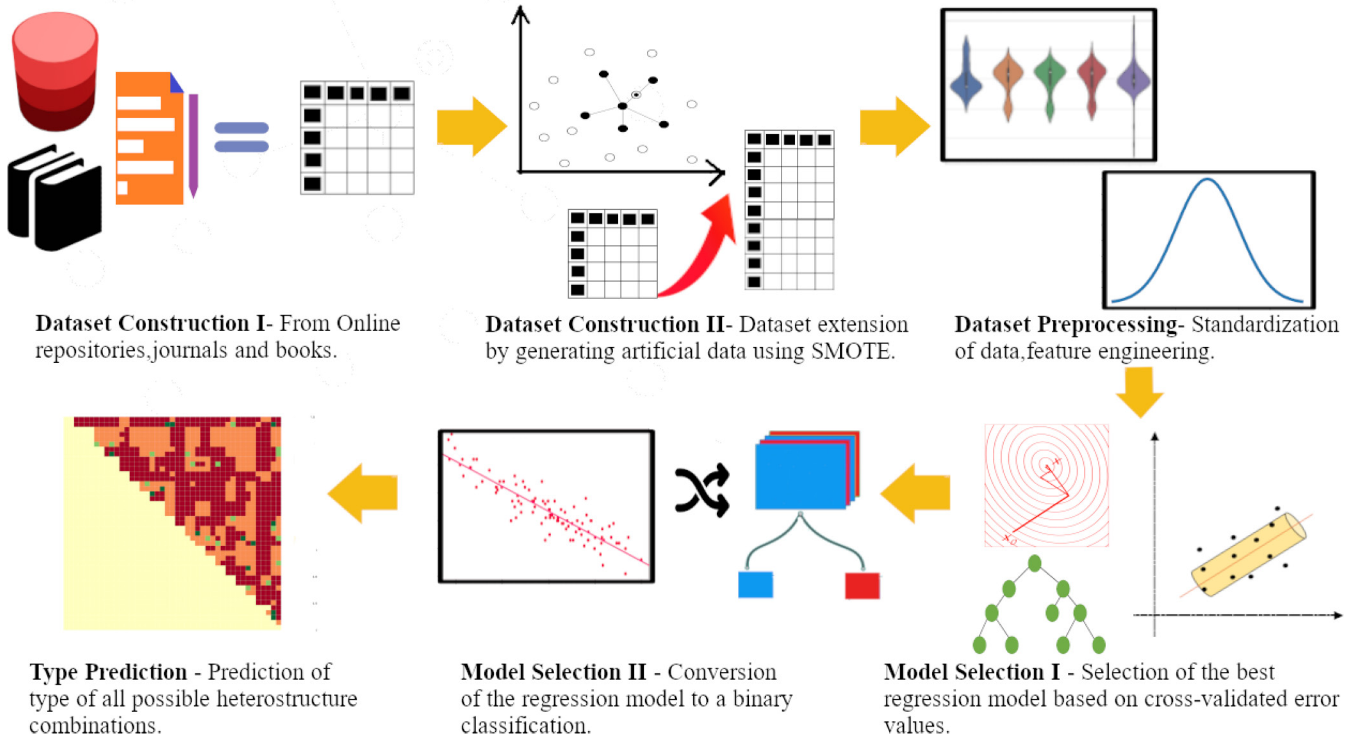
Dataset Construction I- From Online repositories,journals and books.
Dataset Construction II- Dataset extension by generating artificial data using SMOTE.
Dataset Preprocessing- Standardization of data,feature engineering.
Type Prediction - Prediction of type of all possible heterostructure combinations.
Model Selection II - Conversion of the regression model to a binary classification.
Model Selection I - Selection of the best regression model based on cross-validated error values.

FIG. 3. The machine learning flow chart adapted in the present study for prediction of semiconductor heterostructure type.

B. Creation of synthetic data

As described above, our original training dataset, constructed from information found in the literature, consists of only 31 bulk heterostructure information, which is too small a dataset for machine learning to make meaningful predictions. As also mentioned, this dataset of 31 semiconductor heterostructures consists of 20 ($\approx$64.5%) type-I heterostructures and 11 ($\approx$35.5%) type-II heterostructures. Thus, the dataset in addition to being small, has large class imbalance. Since the predictive accuracy of any ML algorithm highly depends on the nature of the training dataset, use of this dataset would lead to misclassification of minority cases. In the literature there are several ways to handle the misclassification issue. For example, one can do the undersampling of the majority class and/or oversampling of the minority class, or one can assign a distinct cost to training examples to increase the sensitivity of the classifier to minority class. In the present study, we adapt the idea of synthetic minority oversampling technique, known as "SMOTE." [21] The choice of this algorithm is guided by the fact that our number of training data is very small, in addition to the class imbalance problem.

Within the machine learning algorithm any object, belonging to the training set (in this case the binary heterostructure), is represented by a n-dimensional feature vector whose components are the n numerical features (in this case 50 numerical features, to be discussed in the next subsection). The vector space associated with these feature vectors is thus called a feature space, with the binary heterostructures as the points in the feature space. SMOTE works by taking the difference between the feature vector of a given minority class sample and each of its k minority class nearest neighbor in the feature space, and multiplying the difference by any random number between 0 to 1 and adding it with the feature vector under consideration. The nearest neighbors are those points in feature space, having closest distances, determined by the length of the n-dimensional feature vector. These k nearest neighbors are randomly chosen and it depends upon the amount of oversampling one requires. In this way the synthetic samples are generated randomly between any two specific points in the feature space. This process requires as an input a number of minority class samples, amount of SMOTE percentage, and a number of nearest neighbors. The number of nearest neighbors can be chosen as desired and it depends on the amount of oversampling one requires. In the present study, we use four nearest neighbors, which is the optimized number we found for our dataset. The goodness of the original dataset with 31 data points and the SMOTE corrected and expanded dataset with 78 data points is tested using random forest [49,50]. The compilation of the confusion matrix shows significant improvement in the expanded dataset compared to the original dataset. The percentage of false positive and false negative prediction of the model for expanded dataset turned out to be 16% compared to 42% for the original dataset, as shown in Table I. The improved performance of the extended dataset can be also measured in terms of an F1 score. In the case of

TABLE I. Confusion matrix for the original (expanded) dataset.

	Positive prediction	Negative prediction
Positive class	15 (33)	5 (7)
Negative class	8 (6)	3 (32)

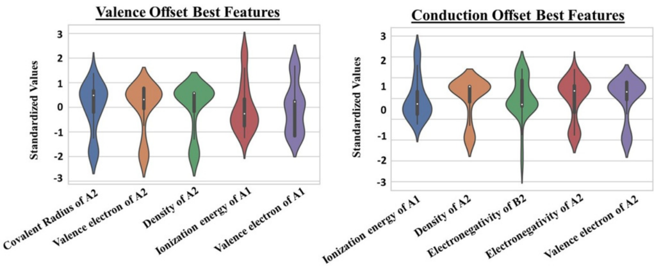

FIG. 4. A violin plot for the distributions of the predominant features in VBO (left) and CBO (right) prediction. The width of each distribution at a given value indicates the number of materials with feature value around that. See text for details.

the original dataset F1 score is found to be 0.56, whereas it shows a significantly improved value of 0.83 for the expanded dataset.

C. Feature space

A number of different features has been proposed as predictors for materials properties. In the context of band-gap prediction by using machine learning methods, different feature spaces have been used: Zhuo *et al.* [51] used 136 engineered elemental features and support vector regression (SVR) model trained and tested on 3896 various forms of semiconductors for experimental band gap prediction, achieving a root mean square error (RMSE) of 0.45 eV. By using 18 features including both elemental properties and low-level DFT computational results of compounds, Lee *et al.* [30] used SVR model on 270 binary and ternary semiconductors and achieved a RMSE of 0.24 eV. Weston *et al.* [33] trained and tested SVR model on 284 I2-II-IV-VI4 kesterite compounds with Heyd-Scuseria-Ernzerhof (HSE) functional calculated band gaps by using 12 elemental features, achieving a RMSE of 0.28 eV. Huang *et al.* [52] used 18-dimensional feature space, which was expanded to a 58-dimensional feature space. In the present study, we start with 12 elemental properties for each of the four components constituting the heterostructure, namely pauling electronegativity, covalent radius, atomic no, atomic weight, melting temperature, ionization energy, period number, number of valence electrons, density, number of the *s*-electron, number of the *p*-electron, and atomic radius giving rise to 48 features to which GGA band gaps of semiconductor A and B are added, making a total of 50-dimensional feature space. The distributions of best features in VBO and CBO are presented as violin plots in Fig. 4. The x-axis of the violin plot marks different features, while the y-axis shows the standardized value and distribution of such features. The standardized values are scaled from the original data and help in viewing miscellaneous data in the same footing. If a feature is represented by x, the corresponding standardized value is given by $x' = (x - \mu)/\sigma$, where μ is the mean of the feature and σ is the standard deviation. Standardized values are thus unit less. In the present case, we have a heterostructure built from two binary semiconductors, named as A and B. A1 and A2 are the two constituent elements in binary semiconductor A and B1 and B2 are the constituent elements in binary semiconductor B. To choose "best features," we take each of the 50 features and perform linear regression for valence and conduction offsets. The features are then sorted in increasing order of their errors, and the best features are the ones with least errors. The best five features in both the cases of valence and conduction offsets are plotted in the figure.

While this 50-dimensional feature space works reasonably well for predicting the magnitude of the band offsets using regression techniques, this set of features is found to perform poorly in the classification problem with heterostructure type. The accurate prediction of the heterostructure type demanded substantial feature engineering. As a first attempt, we constructed differences, means [51] and operations like exponentiation from the features [53], as was implemented previously. This, however, did not improve the classifier performance. Following the work by Weston *et al.* [33], we thus expand the feature space by constructing polynomial combinations of the features in the 50-dimensional feature space. Using third-order polynomial combinations it leads to a total 23 425 number of features with 86% accuracy of the classifier, comparable to that obtained in case of indirect-direct band-gap classifier problem [33].

D. Regression

The magnitude of band offsets of a heterostructure can be predicted using a number of regression models, which aim to determine a relationship between the features of each heterostructure, often called descriptors, and the band offsets of the system. In the present study, we use the LASSO algorithm [22] for the prediction of band offsets in heterostructures. The algorithm of LASSO includes a shrinkage to ordinary linear regression. It utilizes L1 regularization to penalize absolute values of coefficients. As a result, some coefficients can be made zero and the corresponding features can be removed from the model. The goal of the algorithm is to minimize the function [54]

$$\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2}$$

The tuning parameter λ determines the L1 penalty. A large value of λ amounts to discarding many descriptors in the linear regression. The value of λ is chosen by a grid-search method, and the optimum value is the one that minimizes error. We use 10-fold cross validation to determine the accuracy of our model. The model is found to have mean absolute error of 0.25 eV in valence offset prediction. The corresponding values for conduction band offset is 0.40 eV, confirming that prediction of conduction band offset from knowledge of individual semiconductors to be harder than that of valence band, as already seen in Fig. 1. However, this turned out to be still sufficient to distinguish between type I and type II. In this context, it is interesting to compare MAE reported in the literature for the ML model of the band gap of single semiconductors. Xie and Grossman [55] reported the band-gap MAE of 0.388 eV. Pilania *et al.* [56] reported band-gap MAE of 0.45 eV based on multifidelity machine learning models for the accurate band-gap prediction of solids. Gladkikh *et al.* [57] reported band gap MAE of 0.5 eV for predicting the band gaps of ABX_3 perovskites from elemental properties. Judging by

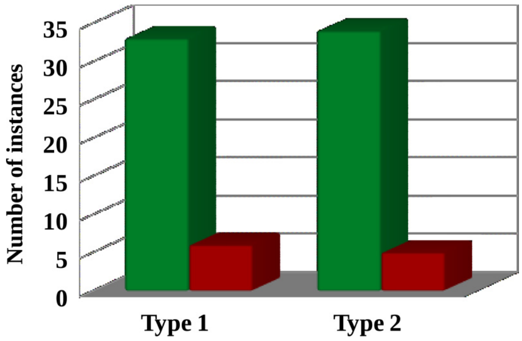

FIG. 5. Validation of machine learning classification model of semiconductor heterostructure with an extended training set of 78 instances. The true positive (type I) and true negative (type II) are shown as green bars, while the false positive (misclassified as type I) and false negative (misclassified as type II) are shown as red bars.

these reported MAEs, the above-obtained MAE appear to be reasonable.

E. Classification

Finally, we make a classification of the type of heterostructures considering only the sign of the predicted offsets. If the predicted offset of both valence and conduction band of a heterostructure have opposite signs the classifier predicts type I, whereas if they have the same signs it predicts type II. Thus, only the signs of cross-validated predictions become important for the classification problem. In the literature, this kind of problem has been dealt with using the logistic model [33]. In our case, a logistic model with 10-fold cross validation correctly predicted the type of 58 instances and incorrectly predicted 20 instances. However, using the LASSO model with 10-fold cross validation is found to provide better agreement. Out of 78 instances, 67 were found to be correctly predicted with 11 incorrectly predicted, as shown in Fig. 5. Analyzing 11 incorrectly classified cases, five are found to be those in original dataset and six are found to belong to synthetic dataset, with only two misclassified cases due to sign mismatch in valence offset and nine due to sign mismatch in conduction offset. Out of the 11 wrongly classified cases, in 9 cases the error was found due to difficulty in prediction of small offset values (∼0.2 eV or less).

Based on the machine learning model, we make predictions on the type of possible heterostructure combinations, which is similar to Fig. 2, but obtained from the machine learning algorithm. Out of 903 number of heterostructures, in 89% of cases a perfect matching between the machine learning and band alignment predictions is observed, as shown in Fig. 6, justifying the capability of machine learning as a cheap means for prediction of heterostructure type.

IV. CRYSTAL SYMMETRY-DEPENDENT HETEROSTRUCTURE TYPE

On examining Figs. 2 and 6, we find for some interesting situations, even for a specific combination of binary semiconductors, the type of heterostructure can be either type I or type II depending on the crystal structure of the semiconductors.

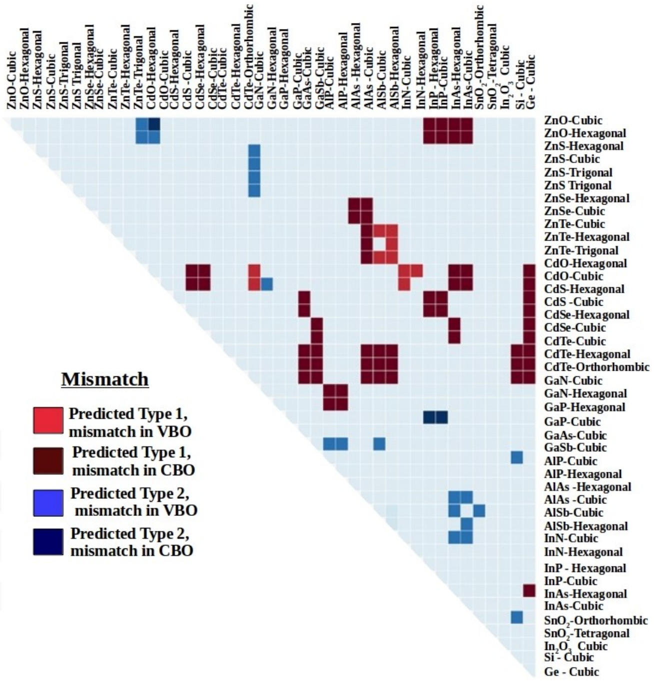

FIG. 6. The ML-predicted semiconductor heterostructures. The cases with matching between ML predictions and band alignment predictions are colored as cyan. The mismatched cases are colored as dark blue/blue (predicted type II by ML and type I by band alignment) and red/brown (predicted type I by ML and type II by band alignment).

As shown in Fig. 7, three such combinations are found, AlP-GaP, ZnO-GaN, and ZnO-InN.

To check the validity of this conjecture, first-principles calculations in the plane-wave basis within the framework of Vienna *ab initio* simulation package (VASP) [58,59] are carried out on the explicit heterostructure models. The lattice mismatch between ZnO and InN either in cubic or hexagonal phase is found to be very large (≈8%). We thus consider heterostructures between cubic-AlP/ZnO and cubic-GaP/GaN, between hexa-AlP/ZnO and hexa-GaP/GaN, and between cubic-AlP/ZnO and hexa-GaP/GaN. To build the heterostructure models, a five-bilayer (001) surface slab of one semiconductor is stacked on top of a five-bilayer (001) slab of another semiconductor, with a two-interface model within the cell. The thickness of the slab model is chosen such that to well preserve the bulk atom properties, with the interfaces built to maximize the fraction of heteropolar bonds. Care is taken so that the bonding at interfacial anionic sites resemble the bulk coordination. The z-component of the atomic coordinates as well the c-axis lattice constant of the

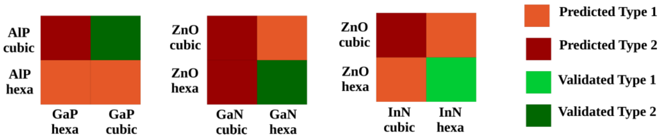

FIG. 7. Crystal structure dependency of heterostructure types of predicted and known binary semiconductors.

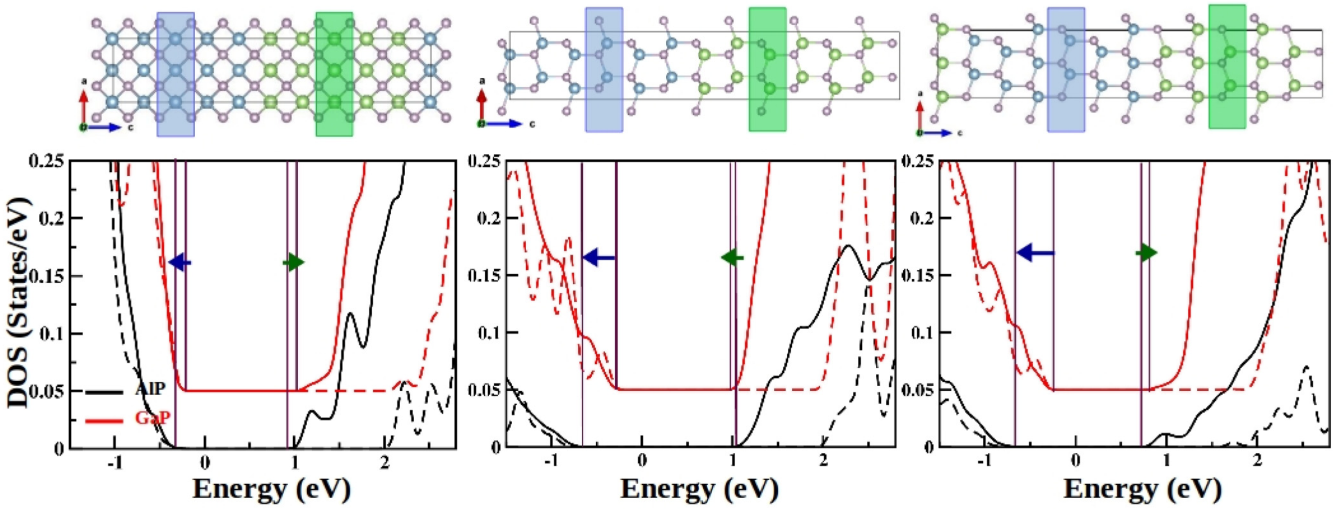

FIG. 8. The relaxed heterostructure geometry (top panels) and density of states (bottom panels) of AlP-GaP in cubic-cubic (left), hexa-hexa (middle), and cubic-hexa (right) crystal structures. The Al and Ga atoms are represented by big blue and green balls, respectively, while the P atoms are shown as small balls. In the density of states plots, shown for two scheme of calculations, GGA (solid lines) and HSE06 (dashed lines), the zero of the energy is set at respective Fermi level. The density of states projected AlP bilayer and GaP bilayer (shaded in color in top panels) are shown as black and red lines, respectively, which are shifted with respect to each other for better visualization. The VB and CB edges in GGA calculation are marked by vertical solid lines. The positive and negative values of offsets are marked by oppositely directed arrows. The offset values in hybrid calculation are similar to that of GGA and are not shown for clarity.

supercells are allowed to relax to release the internal stress. The convergence with respect to cell size has been checked by repeating calculations with seven-bilayer (001) surface slab of one semiconductor stacked on top of a seven-bilayer (001) slab of another semiconductor. The reported calculations are obtained using Monkhorst-Pack k-mesh of 6×6×2 to 10×10×2 depending on the symmetry (cubic/hexa) and heterostructure (AlP-GaP/ZnO-GaN). The convergence with respect to k-points has been checked in terms of k-mesh of 12×12×2 to 14×14×2.

The top panels of Fig. 8, show the relaxed AlP/GaP heterostructures, assuming the cubic and hexagonal symmetry of the constituent semiconductors [60]. Only cation-phosphorous bonds exist at the interface, with the interfacial four-coordinated P atoms as its bonding in bulk AlP or GaP. The GGA [16] exchange-correlation functional is employed to relax the geometry. The electronic calculation for interface supercells is performed with the GGA [16] as well as the hybrid functional, HSE06 [20] scheme. Very interestingly, while the heavily underestimated band gap in GGA is found to be nearly corrected in HSE06, the heterostructure type is found to remain same irrespective of GGA or HSE06 scheme of calculation.

The bottom panels of Fig. 8 show the density of states projected to AlP and GaP bilayers, in cubic-cubic, hexa-hexa, cubic-hexa relaxed geometry of the supercell within the GGA-PBE and HSE06 scheme of calculations. The GGA and HSE06 density of states reveal that although the magnitude of VBO and CBO differ between the two scheme of calculations, the signs of CBO and VBO are same for cubic-cubic combination confirming the experimentally observed type II nature of heterojunction, and establishing the goodness of our first-principles calculations. On the contrary the signs of VBO and CBO are found to be different for hexa-hexa combination both in GGA-PBE and HSE06 calculations, thus confirming the type I nature, as predicted from band alignment and ML consideration in Figs. 2 and 6. Similarly the interface between cubic AlP and hexa GaP is found to be of type II, again in conformity with the prediction.

The results for ZnO-GaN are presented in Fig. 9, with top panels showing the relaxed ZnO/GaN heterostructures, assuming the cubic and hexagonal symmetry of the constituent semiconductors, and the bottom panels showing the density of states. Unlike AlP-GaP which is common anion heterostructure, ZnO/GaN is neither the common anion nor common cation. In principle, two different interfaces can be formed, the Zn atoms bonded to N on one side, and bonded to O on other side, or Ga atoms bonded to O on one side, and bonded to N on other side. For the results presented in the following, we consider Zn bonded to nitrogen interfaces in all cases, which is found to be energetically favorable. The density of states plots confirm type II heterostructure between cubic-cubic ZnO-GaN, as well as between hexa-hexa ZnO-GaN, while that between cubic-hexa is found to be type I, in perfect conformity with the prediction.

Our study thus highlights the heterostructure type may depend crucially on the underlying crystal structure of the constituents, which should be investigated experimentally.

V. APPLICATION TO SEMICONDUCTOR HETEROSTRUCTURE AT NANOSCALE

Finally, it is curious to ask whether the method can be extended to heterostructures at nanoscale. Heterostructures consisting of two different semiconductor quantum dots (QDs) in a coupled quantum dot geometry, possess band

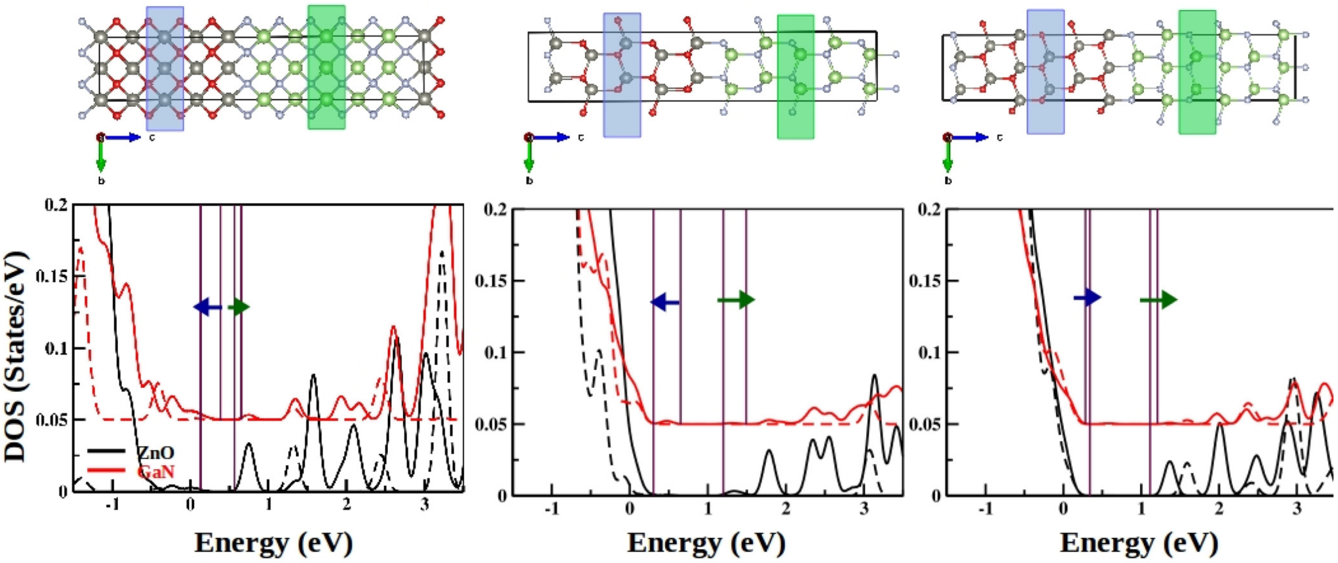

FIG. 9. The relaxed heterostructure geometry (top panels) and density of states (bottom panels) of ZnO-GaN in cubic-cubic (left), hexa-hexa (middle), and cubic-hexa (right) crystal structures. The Zn and Ga atoms are represented by big gray and green balls, respectively, while the O and N atoms are shown as small balls. In the density of states plots, shown for two scheme of calculations, GGA (solid lines) and HSE06 (dashed lines), the zero of the energy is set at respective Fermi level. The density of states projected ZnO bilayer and GaN bilayer (shaded in color in top panels) are shown as black and red lines, respectively, which are shifted with respect to each other for better visualization. The VB and CB edges in GGA calculation are marked by vertical solid lines. The positive and negative values of offsets are marked by oppositely directed arrows. The offset values in hybrid calculation are similar to that of GGA and are not shown for clarity.

offsets at conduction band and valence band depending upon the relative alignments of the energy levels, the CBO and VBO being determined by the band gaps of the constituent QDs, which largely depends on the sizes of the QD's.

The available data for nanoscale heterostructures are extremely limited. Our exhaustive literature search resulted in only two past reports. In one of the experimental studies [61], by using the size-tuned CdS QDs in contact with fixed-sized ZnSe QD, it was shown that band offset at the interface can be tuned selectively. In another computational study from the literature [62] interfaces between small $A_{12}B_{12}$ nanoclusters containing 12 cations and 12 anions with A = Cd/Zn and B = S/Se/Te were investigated. These small magic-sized clusters consist of six member, four member, and two member rings, and 2-bond as well as 6-bond configurations were considered in the computational study [62]. Based on the energy-resolved first-principles charge density plots, the offset in highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) which correspond to VBO and CBO in bulk was calculated in this study [62].

To describe such a situation, we apply the band alignment algorithm starting from the bulk band structure information available in the Materials Project [29], corrected by scissor shift and the quantum confinement correction, the latter being given by the Brus equation [63]

$$\Delta E_g = \frac{\hbar^2\pi^2}{2R^2}\left(\frac{1}{m_e^*} + \frac{1}{m_h^*}\right), \quad (3)$$

where R is the radius of the nanocrystal and m_e^* and m_h^* are the electron and hole effective masses. The electron and hole effective masses can be readily calculated from the Materials Project [29] band curvatures at the CBM and VBM, respectively, using parabolic fits to the band structures.

As shown in the top panel of Fig. 10, even such an approximate approach is able to reproduce the experimental trend of size-dependent CBO and VBO between 2.25-nm-sized ZnSe QDOT and CdS QDOT of sizes 1.3 nm, 2.1 nm, and 2.8 nm, provided the choice of cubic symmetry is made for CdS. Experimentally both ZnSe and CdS QDOTs were found to be in cubic symmetry, in conformity with our predictions.

It is far more challenging to make predictions on small nanoclusters of only 24 atoms from the information of bulk properties. A comparison of energy needed to create low-energy exciton calculated using Brus's model [63] with experimental results show a marked deviation already at radius of 1 nm or so [64]. Nevertheless, we carried out the band alignment of AB coupled QDOTs for common anion and common cation heterostructures, CdS-ZnS, CdSe-ZnSe, CdTe-ZnTe, ZnS-ZnSe, CdS-CdSe, ZnSe-ZnTe, CdSe-CdTe, ZnS-ZnTe, CdS-CdTe assuming a radius of 0.4 nm and cubic (hexagonal) symmetry for CdTe, ZnS, ZnSe, ZnTe (CdS, CdSe). While the quantitative values of the offsets are found to be off by a factor of 10, compared to that found in quantum-chemical calculations on $A_{12}B_{12}$ nanoclusters, the general trend of the valence offsets (HOMO offset) were reproduced. In particular, a decrease of HOMO offset upon moving from 3p (S) to 4p (Se) to 5p (Te) in common anion systems and a increase of VBO between 4d (Cd) and 3d (Zn) in common cation systems are reproduced, which in turn confirms the applicability of such an approximate scheme to capture the essential chemical trend.

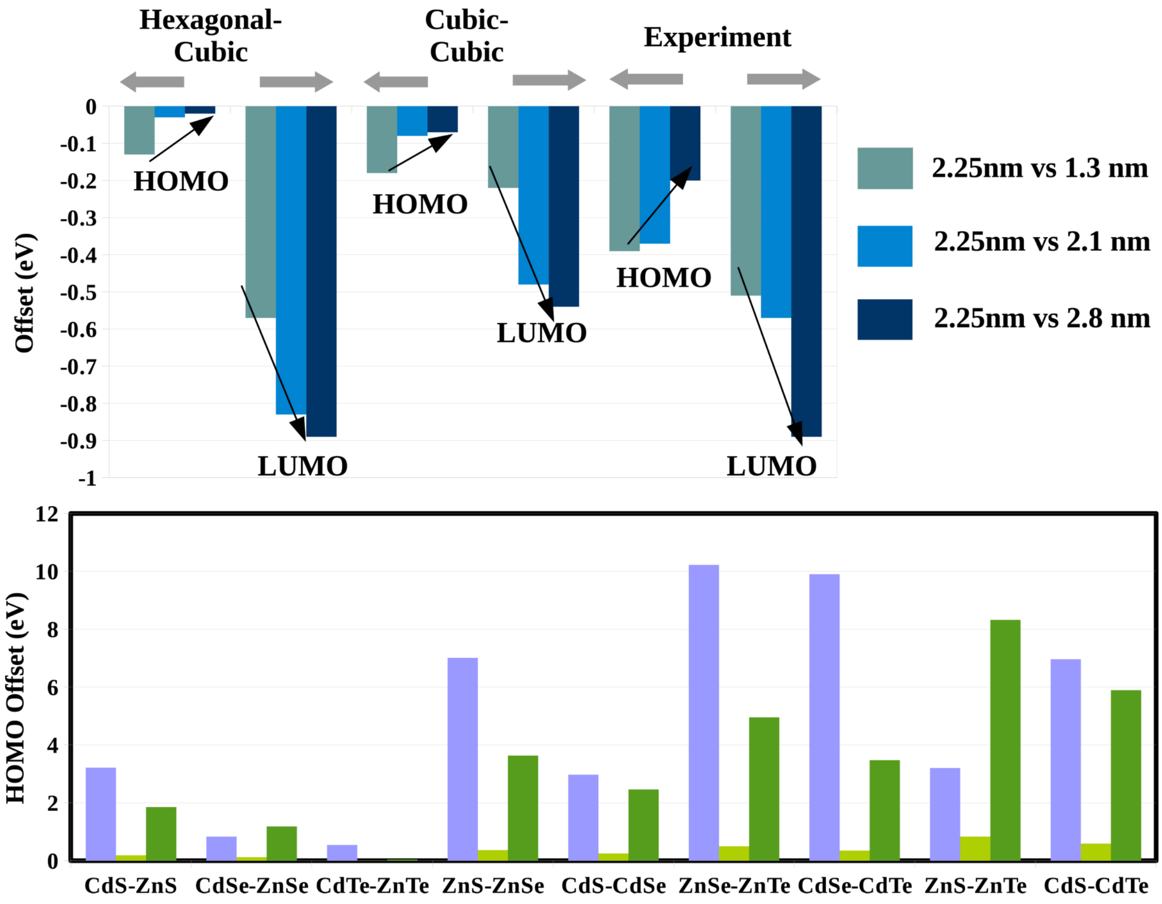

FIG. 10. Top panel: Comparison between predicted and experimentally measured [61] valence or highest occupied molecular orbital (HOMO) and conduction or lowest unoccupied molecular orbital (LUMO) offset in coupled QDOTs of 2.5 nm ZnSe and CdS of varying sizes, 1.3 nm, 2.1 nm, and 2.8 nm, marked in three different colors. For predictions, both cubic and hexagonal crystal symmetries are considered. Bottom panel: Predicted HOMO offset (blue) for common anion and common cation of II-V $A_{12}B_{12}$ semiconductors with computed values [62], scaled by a factor of 10 (green).

To test the applicability of ML expanding the scope to nanostructured heterostructures, we repeat the SMOTE to a dataset consisting of 31 numbers of bulk heterostructures and five nanoscale heterostructures. Out of five nano heterostructures, two are type I and three are type II. To generate synthetic data starting from this 36 dataset of bulk and nanoheterostructures, we include two additional features for the quantum confinement correction terms for semiconductor A and B. The application of SMOTE generated 78 data with 57 ($\approx$73.08%) of them bulk heterostructures and 21 ($\approx$26.92%) of them nanoheterostructures. Accounting for the class imbalance by SMOTE, out of 78 heterostructures 38 ($\approx$48.71%) are type I and 40 ($\approx$51.28%) are type II heterostructures. Starting with an F1 score of 0.58 for the original dataset of 36, the F1 score is improved to 0.83 for the SMOTE-corrected dataset of 78. Thus, we conclude that "SMOTE" works satisfactorily for the expanded dataset including the nanostructures. Upon inclusion of nanoheterostructures in the dataset of the ML classification, the MAE values became 0.29 eV and 0.49 eV for CBO and VBO, respectively. The corresponding classification model rightly classified 82% of the 78 instances. The slight decrease in accuracy in the dataset including nanostructures compared to the bulk-only dataset is pertaining to the fact that very less data about nanoheterostructures could be extracted from the literature and added to our dataset.

VI. SUMMARY AND DISCUSSION

The designing of heterostructures of a specific type, formed by bringing together two semiconductors of comparable lattice constants but differing band gaps, is one of the important topics in semiconductor industry. Understandably this calls for materials engineering, which requires either a cost and/or effort intensive path of experimental synthesis and characterization, or first-principles calculations involving large simulation cells with several tens or hundreds of atoms and accurate numerical schemes going beyond the conventional DFT [65]. In the present study, we propose and demonstrate that machine learning can be a viable alternative for materials selection for semiconductor heterostructure design with targeted heterostructure type. While there exists machine learning studies for the band-gap prediction of semiconductors, to the best of our knowledge, no such study exists for the prediction of heterostructure type, which is admittedly more challenging. Validating the scheme on few synthesized bulk semiconductor heterostructures with known type, we make predictions on a large number (872) of bulk semiconductor heterostructures which are either not been synthesized, or not characterized for their type. The predictions furthermore bring out an interesting factor that, depending on the crystal structure of constituent semiconductors, either a type I or type II situation may be realized.

As is well known, the biggest restriction on the materials selection process of semiconductor heterostructure is lattice constant. For coherent growth of the heterostructure, lattice match, or small mismatch between two adjacent semiconductor materials that allows epitaxial growth of one on top of the other is an essential requirement [66]. However, there have been attempts to form lattice-mismatched heterostructures via heteroepitaxy [67], mechanical-thermal direct bonding [68], or grafting [69]. Following the later method, successful formation of lattice-mismatched semiconductor heterostructures has been achieved between Ge/Si (diamond lattice), Si/GaAs (zinc blende lattice), GaAs/GaN (hexagonal lattice), and Si/GaN heterostructures with a lattice mismatch of 4.2%, 4.9%, 77.1%, and 70.2%, respectively. Among the yet-to-synthesized 872 heterostructures, for 139 cases the lattice mismatch is found to be less than 2%, while for 466 cases the lattice mismatch is found to be within 6%, making synthesis of the predicted heterostructures probable.

The situation becomes further hopeful by the fact that same scheme seems to also show satisfactory performance for the semiconductor heterostructure in nanoscale. This establishes the proposed scheme to be a powerful computational tool for fast materials selection for heterostructure design.

Finally, it is to be noted that within the scope of the present machine learning project, the aim was to make a prediction on heterostructure classification as type I and type II. The present exercise though does not provide a prediction on whether the band gap of the heterostructure will be direct or indirect, the knowledge of which is important for applications like optoelectronics. It might be difficult to apply machine learning for this purpose, as the data size of heterostructures with classification as direct/indirect band gap is expected to be even smaller. To the best of our knowledge, machine learning prediction on direct/indirect classification for semiconductors has been made in the context of a single semiconductor, and not for semiconductor pairs, as is needed for heterostructure, that, in additiion, is for only a specific semiconductor family I2-II-IV-V4 [33]. This calls for future investigation.

ACKNOWLEDGMENTS

The authors acknowledge the support of DST Nanomission for the computational facility used in this study. The authors thank Indra Dasgupta for suggesting the problem and Arup Chakraborty for sharing the thesis data. T.S.D. acknowledges J.C. Bose National Fellowship (Grant No. JCB/2020/000004) for research support.

[1] H. Kroemer, Proc. IEEE **51**, 1782 (1963).

[2] R. W. Chuang, R.-X. Wu, L.-W. Lai, and C.-T. Lee, Appl. Phys. Lett. **91**, 231113 (2007).

[3] Y. Lin, X. Li, D. Xie, T. Feng, Y. Chen, R. Song, H. Tian, T. Ren, M. Zhong, K. Wangb, and H. Zhu, Energy and Environ Sci. **6**, 108 (2013).

[4] S. K. Behura, C. Wang, Y. Wen, and V. Berry, Nat. Photon. **13**, 312 (2019).

[5] B. Roul, M. Kumar, M. K. Rajpalke, T. N. Bhat, and S. B. Krupanidhi, J. Phys. D **48**, 423001 (2015).

[6] J. H. Davies, *The Physics of Low-Dimensional Semiconductors* (Cambridge University Press, Cambridge, England, 2012).

[7] Zh. L. Alferov, Rev. Mod. Phys. **73**, 767 (2001).

[8] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, npj Comput. Mater. **5**, 83 (2019).

[9] Y. Liu, T. Zhao, W. Ju, and S. Shi, J. Materiomics **3**, 159 (2017).

[10] A. Halder, S. Rom, A. Ghosh, and T. Saha-Dasgupta, Phys. Rev. Appl. **14**, 034024 (2020).

[11] P. D. C. King, T. D. Veal, P. H. Jefferson, S. A. Hatfield, L. F. J. Piper, C. F. McConville, F. Fuchs, J. Furthmüller, F. Bechstedt, H. Lu, and W. J. Schaff, Phys. Rev. B **77**, 045316 (2008).

[12] Y. Hinuma, A. Gruneis, G. Kresse, and F. Oba, Phys. Rev. B **90**, 155405 (2014).

[13] J. Heyd, J. E. Peralta, G. E. Scuseria, and R. L. Martin, J. Chem. Phys. **123**, 174101 (2005).

[14] D. Wing, J. B. Haber, R. Noff, B. Barker, D. A. Egger, A. Ramasubramaniam, S. G. Louie, J. B. Neaton, and L. Kronik, Phys. Rev. Mater. **3**, 064603 (2019).

[15] Y. Yang, V. Gorelov, C. Pierleoni, D. M. Ceperley, and M. Holzmann, Phys. Rev. B **101**, 085115 (2020).

[16] R. Elmér, M. Berg, L. Carlén, B. Jakobsson, B. Norén, A. Oskarsson, G. Ericsson, J. Julien, T.-F. Thorsteinsen, M. Guttormsen, G. Løvhøiden, V. Bellini, E. Grosse, C. Müntz, P. Senger, and L. Westerberg, Phys. Rev. Lett. **78**, 1396 (1997).

[17] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, Comput. Mater. Sci. **50**, 2295 (2011).

[18] W. Setyawan, R. M. Gaume, S. Lam, R. S. Feigelson, and S. Curtarolo, ACS Comb. Sci. **13**, 382 (2011).

[19] E. P. Shapera and A. Schleife, Adv. Theory Simul. **1**, 1800075 (2018).

[20] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003); **124**, 219906 (2006).

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, J. Artif. Intell. Res. **16**, 321 (2002).

[22] R. Tibshirani, J. R. Statist. Soc. B **58**, 267 (1996).

[23] S. Dias and S. B. Krupanidhi, AIP Adv. **5**, 047137 (2015).

[24] S.-H. Wei and A. Zunger, Appl. Phys. Lett. **72**, 2011 (1998).

[25] C. G. Van de Walle and J. Neugebauer, Nature **423**, 626 (2003).

[26] R. Anderson, Solid State Electron. **5**, 341 (1962).

[27] J. Tersoff, Phys. Rev. B **30**, 4874 (1984).

[28] A. Schleife, F. Fuchs, C. Rodl, J. Furthmuller, and F. Bechstedt, Appl. Phys. Lett. **94**, 012104 (2009).

[29] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, APL Mater. **1**, 011002 (2013).

[30] J. Lee, A. Seko, K. Shitara, K. Nakayama, and I. Tanaka, Phys. Rev. B **93**, 115104 (2016).

[31] G. Pilania, A. Mannodi-Kanakkithodi, B. Uberuaga, R. Ramprasad, J. Gubernatis, and T. Lookman, Sci. Rep. **6**, 19375 (2016).

[32] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, npj Comput. Mater. **2**, 16028 (2016).
[33] L. Weston and C. Stampfl, Phys. Rev. Mater. **2**, 085407 (2018).
[34] A. Wadehra, J. W. Nicklas, and J. W. Wilkins, Appl. Phys. Lett. **97**, 092119 (2010).
[35] E. T. Yu, M. C. Phillips, D. H. Chow, D. A. Collins, M. W. Wang, J. O. McCaldin, and T. C. McGill, Phys. Rev. B **46**, 13379 (1992).
[36] K. Steiner, W. Chen, and A. Pasquarello, Phys. Rev. B **89**, 205309 (2014).
[37] A. L. Yang, H. P. Song, H. Y. Wei, X. L. Liu, J. Wang, X. Q. Lv, P. Jin, S. Y. Yang, Q. S. Zhu, and Z. G. Wang, Appl. Phys. Lett. **94**, 163301 (2009).
[38] J. R. Waldrop, R. W. Grant, and E. A. Kraut, J. Vacuum Sci. Tech. B **7**, 815 (1989).
[39] S.-H. Wei, S. B. Zhang, and A. Zunger, J. Appl. Phys. **87**, 1304 (2000).
[40] S. W. King, R. J. Nemanich, and R. F. Davis, J. Appl. Phys. **118**, 045304 (2015).
[41] C. Chen, M. Dutta, and M. A. Stroscio, J. Appl. Phys. **95**, 2540 (2004).
[42] B. Höffling, A. Schleife, C. Rödl, and F. Bechstedt, Phys. Rev. B **85**, 035305 (2012).
[43] B. Höffling, A. Schleife, C. Rödl, F. Fuchs, and F. Bechstedt, Appl. Phys. Lett. **97**, 032116 (2010).
[44] A. Smith, A. Mohs, and S. Nie, Nat. Nanotechnol. **4**, 56 (2009).
[45] S.-H. Wei and A. Zunger, J. Appl. Phys. **78**, 3846 (1995).
[46] H. P. Song, A. L. Yang, H. Y. Wei, Y. Guo, B. Zhang, G. L. Zheng, S. Y. Yang, X. L. Liu, Q. S. Zhu, Z. G. Wang, T. Y. Yang, and H. H. Wang, Appl. Phys. Lett. **94**, 222114 (2009).
[47] G. Martin, S. Strite, A. Botchkarev, A. Agarwal, A. Rockett, H. Morkoç, W. R. L. Lambrecht, and B. Segall, Appl. Phys. Lett. **65**, 610 (1994).
[48] H. Kroemer, Physica E **20**, 196 (2003).
[49] Leo Breiman, Mach. Learn. **45**, 5 (2001).
[50] A. Liaw and M. Wiener, R News **2**, 18 (2002).
[51] Y. Zhuo, A. M. Tehrani, and J. Brgoch, J. Phys. Chem. Lett. **9**, 1668 (2018).
[52] Y. Huang, C. Yu, W. Chen, Y. Liu, C. Li, C. Niu, F. Wang, and Y. Jia, J. Mater. Chem. C **55**, 10856 (2019).
[53] N. Kumar, P. Rajagopalan, P. Pankajakshan, A. Bhattacharyya, S. Sanyal, J. Balachandran, and U. V. Waghmare, Chem. Mater. **31**, 314 (2019).
[54] S. Lau, J. Gonzalez, and D. Nolan, The Principles and Techniques of Data Science [online] (2020), https://www.textbook.ds100.org/intro.html [January, 2021].
[55] T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).
[56] G. Pilania, J. E. Gubernatis, and T. Lookman, Comput. Mater. Sci. **129**, 156 (2017).
[57] V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, and K. S. Kim, J. Phys. Chem. C **124**, 8905 (2020).
[58] G. Kresse and J. Hafner, Phys. Rev. B **47**, R558 (1993).
[59] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[60] As the shapes of cubic and hexagonal cells are different, to create the interface between cubic and hexagonal systems, we cut the cubic cell by a plane (111) which exposes the surface with an atomic arrangement in hexagonal symmetry, to which the hexagonal cell is attached. Though the heterostructures made from cubic-hexa and hexa-hexa look almost similar from the top panel of Figs. 8 and 9, they are quite different in the details. For example, the hexa-hexa lattice parameters are $a = b$ (7.771/6.578 Å for AlP-GaP/ZnO-GaN) $\neq c$ (38.500/26.400 Å for AlP-GaP/ZnO-GaN) but for cubic-hexa a (6.745/3.274 Å for AlP-GaP/ZnO-GaN) $\neq b$ (7.788/5.671 Å for AlP-GaP/ZnO-GaN) $\neq c$ (38.300/31.800 Å for AlP-GaP/ZnO-GaN).
[61] A. Dalui, A. Chakraborty, U. Thupakula, A. H. Khan, S. Sengupta, B. Satpati, D. D. Sarma, I. Dasgupta, and S. Acharya, J. Phys. Chem. C **120**, 10118 (2016).
[62] A. Chakraborty, Ph.D. thesis, Calcutta University, 2019.
[63] L. Brus, J. Phys. Chem. **90**, 2555 (1986).
[64] T. Kippeny, L. A. Swafford, and S. J. Rosenthal, J. Chem. Educ. **79**, 1094 (2002).
[65] G. Onida, L. Reining, and A. Rubio, Rev. Mod. Phys. **74**, 601 (2002); A. Seidl, A. Görling, P. Vogl, J. A. Majewski, and M. Levy, Phys. Rev. B **53**, 3764 (1996); L. K. Wagner and D. M. Ceperley, Rep. Prog. Phys. **79**, 094501 (2016).
[66] Z. I. Alferov, V. M. Andreyev, V. I. Korol'Kov, and E. L. Portnoi, Kristall Technik **4**, 495 (1969).
[67] Y. Chen and J. Washburn, Phys. Rev. Lett. **77**, 4046 (1996).
[68] A. Black, IEEE J. Sel. Top. Quantum Electron. **3**, 943 (1997).
[69] D. Liu *et al.*, arXiv:1812.10225.