Predicting Critical Temperature Of Superconductors Using Machine Learning Techniques

Arun Pavithran Rajasekaran, Jemilsan Jeyakumar, Shabbir Kutbuddin, Yuxiao Pu

May 2, 2025

Abstract

This project looks at the Superconductivity Data dataset in the UCI Machine Learning Repository. This dataset consists of 21,263 superconducting materials where each of them has a total of 81 extracted features (like atomic mass, valence, electronegativity, etc). The primary goal will be to make prediction models to try and estimate the critical temperature (T_c) of superconductors based on the features. We will use regression methods and aims to find the most important predictors to T_c and examine different machine learning algorithms on predicting T_c. The results will build an understanding of superconducting materials as well as help discover new superconductors with convenient properties.

1 Introduction

Superconductivity is the state of a material that demonstrates zero electrical resistance while rejecting magnetic fields at all temperatures below a critical temperature (T_c). Superconductivity has received considerable attention from researchers studying the potential implications of the application of superconducting properties in technology and industry. Application areas for superconductors include MRI, maglev trains, and eventually lossless power lines and quantum computing. However, the discovery of new superconducting materials has occurred primarily through trial and error methods.

Recently, the growth of large datasets and advancements in machine learning provide researchers several new paths of looking for new superconductors. The 'Superconductivity Data' dataset, obtained from the UCI Machine Learning Repository, has 21,263 material samples described and 81 features based on composition. The 81 features for superconductivity include atomic mass, electronegativities, thermal conducts, and valences of the different elements within the material sample that may affect the material's superconducting behaviour.

The aim of this study is:

- To determine whether machine learning models can accurately predict the critical temperature (T_c) of superconducting materials based on their features.

To address these questions, we applied a variety of regression models, ranging from linear and regularized methods to advanced ensemble techniques like Random Forest, XGBoost, and LightGBM. We also performed feature selection and dimensionality reduction to enhance model interpretability and efficiency.

Prior to model development, an exploratory data analysis (EDA) was conducted to better understand the structure of the dataset and the distribution of the target variable. The critical temperature values span a wide range, with a right-skewed distribution centered around lower T_c values. Correlation analysis revealed that features related to thermal conductivity, atomic radius, and valence exhibited strong associations with T_c, suggesting that these properties could be influential predictors.

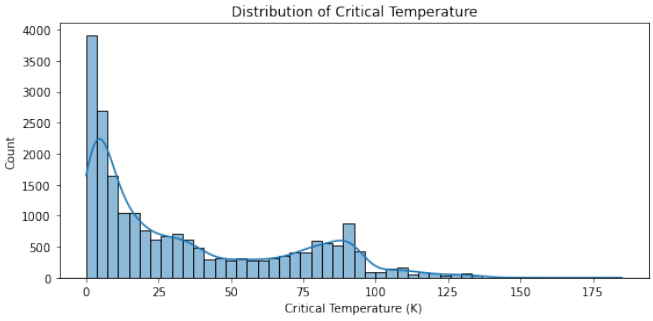

Figure 1: Distribution of critical temperatures (T_c) in the superconductivity dataset

This initial analysis highlighted the complexity of the relationships between material composition and superconducting behavior, reinforcing the importance of employing diverse machine learning techniques and rigorous evaluation to model such phenomena effectively.

2 Literature Review

The discovery and understanding of superconducting materials have historically been rooted in experimental science. Since the initial discovery of superconductivity

in mercury [1], researchers have sought materials that exhibit higher critical temperatures (T_c) to enable more practical applications. Traditional experimental methods, while successful in identifying classes of superconductors such as cuprates and iron-based compounds, have proven to be time-consuming and costly.

In recent years, machine learning techniques have emerged as powerful tools for materials science, offering the potential to accelerate the discovery of new superconductors. Various studies have shown that supervised learning algorithms (like Random Forests) can predict the critical temperature of superconductors with promising accuracy, using features derived from elemental properties [2]. Gradient boosting machines (XGBoost), for instance, have demonstrated strong predictive capabilities, outperforming traditional empirical approaches [3].

One of the most widely used resources in this domain is the SuperCon database, maintained by the National Institute for Materials Science (NIMS). It contains over 12,000 superconducting materials with reported critical temperatures and chemical compositions. This dataset has been extensively used in prior studies to train and validate machine learning models [3]. While comprehensive, SuperCon predominantly includes known high-T_c materials such as cuprates and iron-based superconductors, introducing potential biases in the training data.

Researchers have also explored the application of deep learning techniques to superconductivity datasets. Deep neural networks are capable of modeling complex, non-linear relationships between features and critical temperature, although challenges related to data sparsity, noise, and model interpretability persist [5]. While deep models offer strong predictive power, their "black-box" nature limits the physical insights that can be directly obtained from them.

Despite these advances, challenges remain in the accurate and interpretable prediction of superconductivity. Many existing studies have relied on relatively small datasets, and there is still ongoing debate about which material properties are most influential in determining T_c [6]. Furthermore, many machine learning models focus heavily on predictive performance without providing actionable scientific understanding.

The Superconductivity Data dataset from the UCI Machine Learning Repository presents an opportunity to address some of these challenges. With its large number of samples and detailed feature set, it allows for a comprehensive evaluation of machine learning models and feature selection strategies. Building upon prior work, this study aims to develop predictive models that not only achieve high accuracy but also offer insights into the fundamental factors driving superconductivity.

3 Methodology

3.1 Data Collection

The dataset used in this study was obtained from the UCI Machine Learning Repository, titled Superconductivity Data. It contains 21,263 observations of superconducting materials, each represented by 81 numerical features extracted from the chemical formulae of the compounds. These features are derived from basic physical and chemical properties of the constituent elements, such as atomic mass, electron affinity, thermal conductivity, valence electron count, and electronegativity.

No additional data collection was performed by the group members; all data used in this analysis were publicly available and preprocessed to a degree suitable for machine learning applications. As there are no missing values in the dataset, it was well-suited for direct application of regression algorithms after appropriate scaling and transformation.

3.2 Data Preprocessing

To prepare the superconductivity dataset for machine learning, several data preprocessing steps were performed to ensure model robustness and optimal performance. The dataset originally contained no missing values, so attention was primarily focused on outlier removal, feature scaling, and dimensionality reduction.

Outliers were detected and removed using the Interquartile Range (IQR) method for each numerical feature. This step was critical in reducing the influence of extreme values, particularly in target-relevant features that could bias model learning. After this process, the dataset was reduced from its original size, but the remaining data retained sufficient diversity for training.

3.3 Feature Engineering and Selection

A superconductivity dataset with 81 features based on some physical baby property of materials poses a problem in high-dimensionality so we took a three-pronged complementary approach to feature selection.

Correlation Analysis: To analyse the relationship between each feature and the critical temperature (T_c) by means of a correlation matrix. Each feature was classified into weak (0.0–0.2), moderate (0.2–0.5), and strongly correlated (0.5–1.0) bands. The correlation analysis showed that the features `wtd_std_ThermalConductivity` (0.72), `range_ThermalConductivity` (0.69), and `range_atomic_radius` (0.65) had the greatest correlation with T_c with positive correlations and `wtd_mean_Valence`

(−0.63), `wtd_gmean_Valence` (−0.62), and `mean_Valence` (−0.60) had the three features to have the greatest negative correlation with T_c.

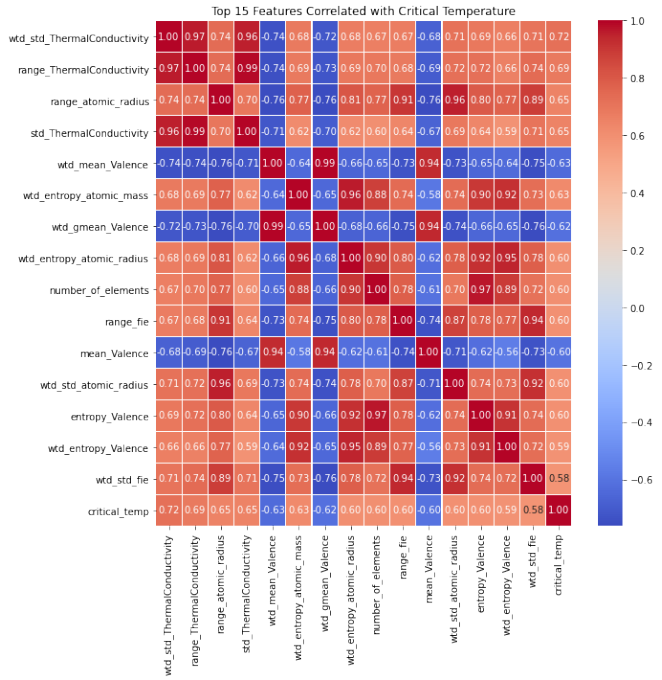

Figure 2: Correlation heatmap showing the top 15 features most correlated with critical temperature (T_c)

Multicollinearity Reduction: To address issues of multicollinearity, highly correlated features (threshold = 0.8) were removed and the dimensionality features were reduced from 82 features to 27 features. Features were kept interpretable while also eliminating redundant information that might influence model performance stability.

Random Forest Feature Importance: A RandomForestRegressor was used to evaluate the impact of features used in the model, yielding `wtd_mean_Valence` (0.20), `wtd_std_ElectronAffinity` (0.17), and `wtd_mean_ThermalConductivity` (0.12) as the most significant predictors of critical temperature. This data was useful for gaining physical knowledge and helped inform future models.

Principal Component Analysis: We conducted Principal Component Analysis (PCA) on the scaled features retaining 95% of the variance with only 15 principal components (dimensionality reduction of 81.5%). The cumulative explained variance plot indicated that the vast majority of the variance information was still captured in these components allowing significantly more efficient data computation while still reflecting the variance structure of the data.

3.4 Machine Learning Models

Different regression models have been explored to predict superconducting critical temperatures, starting with baseline models, and progressively using more advanced model and regression techniques.

Model 1: Baseline Linear Models:

Regression models predict a target variable based on a linear combination of input variables or feature. In the context of predicting superconductivity, the model assumes that the critical temperature (T_c) is expressed as a linear combination (or weighted sum) of theoretical material properties. The basic linear regression model is formulated as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \qquad (1)$$

Where y represents the critical temperature, x_i are the material properties (features), β_i are the coefficients that the model learns, and ϵ is the error term.

Linear models offer strong interpretability, as each coefficient directly communicates the extent to which its corresponding feature influences the critical temperature. However, linear models cannot capture complex nonlinear relationships or interactions between features. This limitation contributes to their relatively poor performance in this context, with a root mean squared error (RMSE) of approximately 19.24 K, reflecting the underlying complexity of superconductivity, which involves intricate quantum mechanical behavior.

To improve generalization and address overfitting, regularized analogues such as Ridge, Lasso, and ElasticNet regression introduce penalty terms into the optimization objective:

- **Ridge regression** adds an L2 penalty (the sum of squared coefficients)
- **Lasso regression** adds an L1 penalty (the sum of absolute coefficients)
- **ElasticNet** combines both L1 and L2 penalties

These regularization strategies help mitigate multicollinearity and facilitate variable selection. However, they do not address the fundamental limitation of linearity in the model structure.

Model 2: Tree-Based Ensemble Methods

Tree-based models divide the feature space into regions, and predictions come from expected values within those regions. Tree-based models can recognize nonlinear relationships and complicated interactions without

specifying a functional form.

A decision tree essentially partitions the data by recursively splitting on feature values that minimize a loss function (normal mean squared error for regression):

1. At each node, the algorithm chooses the threshold that leads to the minimum impurity (variance) in the child nodes.
2. The algorithm minimizes the following objective for a split on feature j at threshold t:
$$\min_{j,t} \left[\frac{n_{\text{left}}}{n} \text{Var}_{\text{left}} + \frac{n_{\text{right}}}{n} \text{Var}_{\text{right}} \right]$$
3. This process repeats until a stopping criterion is met (e.g., maximum depth, minimum samples per leaf, etc.).
4. The final prediction is the average of the target values within the leaf node in which a given sample falls.

While single decision trees tend to overfit, ensemble methods such as Random Forest and Gradient Boosting mitigate this issue:

- **Random Forest** builds many decorrelated trees by bootstrapping the training data, using a random subset of features at each split, and averaging the predictions from all trees.
- **Gradient Boosting** builds trees sequentially, where each new tree corrects the errors of the previous model by fitting to the residuals and updating the predictions with a learning rate.

LightGBM and **XGBoost** are efficient implementations of gradient boosting that include additional innovations to improve speed and accuracy. Their superior performance in this study (RMSE $\approx$ 11.75–12.27 K) indicates their ability to capture the highly non-linear patterns inherent in the superconductivity data.

Model 3: Support Vector Regression

Support Vector Regression (SVR) follows a similar approach as Support Vector Machines, and instead of focusing on minimizing squared error versus observations as in previous regression methods, SVR uses a function documented in a regression space that does not differ from observed target variables by more than a margin of ϵ , while remaining as flat as possible.

All models were assessed at length using 5-fold cross-validation and were evaluated using negative root mean squared error (RMSE) as the primary performance measure. Cross-validation involved stratifying the data into five folds. Each fold served as the validation set while being trained on the other four folds while estimating model performance accounts for the variability in the data and reduces the likelihood of overfitting.

4 Results and Evaluation

The initial model comparison through 5-fold cross-validation indicated that there was substantial variation in performance for each algorithm employed, and where tree based ensemble methods performed significantly better than both linear models and SVR:

1. LightGBM had a highly comparable performance baseline (RMSE = 12.51 K) that suggests its leaf-wise growth strategy and gradient-based sampling can adequately capture the underlying relationships with comparable accuracy.
2. Random Forest had the best baseline performance (RMSE = 12.52 K), leveraging ensemble averaging and bootstrapped sampling to minimize overfitting and capture complex feature interaction.
3. XGBoost had a similar (if slightly lower) baseline performance (RMSE = 12.73 K), demonstrating that its regularized objective function and second-order approximation can capture the non-linear patterns in the superconductivity data well.
4. Gradient Boosting performed moderately well (RMSE = 13.90 K), with a sequential error-correction process yielding reasonable predictions when compared to the specialized implementations of XGBoost and LightGBM.
5. Decision Tree performed relatively poorly, demonstrating very high error (RMSE = 15.86 K, std = 0.68 K), which was expected as the consistency and generality offered by ensemble methods over single trees were shown to be significant for a complex problem still predictive at the ratio stage (even without full leaf splitting depth).
6. SVR with an RBF kernel provided moderate accuracy (RMSE = 16.92 K), indicating that while flexible, it suffers from scalability and sensitivity to hyperparameters in larger datasets.
7. Lasso and Elastic net performed the worst (RMSE 22.54 K) suggesting that Lasso, and possibly even ElasticNet, with its sparsity inducing properties, may have discarded some non-linear contributing features altogether, creating further error.
8. Linear models struggled with these data, with both Linear Regression and Ridge performing similarly (RMSE $\approx$ 22.54 K), suggesting that linear models, which function on the premise of linear decision boundaries, would be incapable of consistently capturing the complexity of the non-linear relationships between the material properties and measured critical temperature.

These results demonstrate quantitatively that relationships between material properties and critical temperature must be predominantly non-linear in nature, sufficient enough such that tree-based ensemble methods were able to demonstrate a 39% reduction in prediction error compared to linear based approaches to superconductivity.

Table 1: Model performance comparison based on test RMSE and R^2

Index	Model	BestParams	Test RMSE	Test R^2
0	LightGBM	{num_leaves: 50, n_estimators: 200, learning_rate: 0.1}	12.51	0.856
1	RandomForest	{n_estimators: 200, max_depth: None}	12.52	0.855
2	XGBoost	{n_estimators: 200, max_depth: 6, learning_rate: 0.1}	12.73	0.850
3	GradientBoosting	{n_estimators: 200, max_depth: 5, learning_rate: 0.1}	13.90	0.822
4	DecisionTree	{min_samples_leaf: 5, max_depth: None}	15.86	0.768
5	SVR	{kernel: 'rbf', gamma: 'auto', C: 10}	16.92	0.736
6	ElasticNet	{l1_ratio: 0.8, alpha: 0.01}	22.54	0.531
7	Ridge	{alpha: 10}	22.54	0.531
8	Lasso	{alpha: 0.001}	22.54	0.531
9	LinearRegression	{}	22.54	0.531

4.1 Hyperparameter optimization

Upon considerations of baseline performance and the balance of computational efficiency, the LightGBM model was chosen for hyper-parameter tuning via RandomizedSearchCV. RandomizedSearchCV is a useful tool that employs Monte Carlo sampling over the parameter space of the candidate method and is the most efficient way to explore an intractable dimension space for hyperparameters. The tuning process was done in a structured way by tuning multiple parameters in the LightGBM model:

(a) Boosting type: gbdt (traditional gradient boosting), dart (dropouts meet multiple additive regression trees), and goss (gradient-based one-side sampling) - evaluating multiple approaches to constructing the model ensemble.

(b) Number of leaves: 20, 31, 40 - the number of leaves on each tree making each of the trees less complex.

(c) Learning rate: 0.01, 0.05, 0.1 - the contribution of the trees to the final prediction.

(d) Number of estimators: 100, 200, 300 - how many trees total are in the sequential ensemble.

(e) Max depth: -1 (infinite), 5, 10 - the maximum number of nodes from root to leaf.

(f) Regularization parameters (alpha,lambda) - 0, 0.001, 0.01- the strength of L1 and L2 regularization respectively.

(g) Sampling of both features and data - 0.8, 1.0 - the fraction of features and observations that will be used in the construction of the tree.

The optimized LightGBM model produced an RMSE of 12.22 K, MAE of 8.30 K and, a R^2 of 0.86 on the held-out test set. The optimal configuration employed gradient based one side sampling (GOSS), 31 leaves, a learning rate of 0.1, 300 estimators, max depth of 10 and decent amount of regularization (alpha=0.01, lambda=0.01). This model had a good balance of complexity or capacity against generalization ability indicatively capturing the salient interaction effects in the complex data of superconductivity while avoiding shared pitfalls of overfitting.

4.2 Error Characteristics

- **Residual Spread:** At low true Critical Temperature (< 20 K), predictions occasionally overshoot (predicted $>$ actual), whereas at very high Critical Temperature (> 120 K), there is a tendency to undershoot. This "regression-toward-the-mean" effect is common when models are trained to minimize squared error.

- **Heteroscedasticity:** The vertical scatter around the y = x line increases slightly with Critical Temperature, suggesting that prediction uncertainty grows for more extreme superconductors. A future refinement could involve quantile regression or conformal methods to better quantify prediction intervals.

4.3 Final Model Performance

The tuned LightGBM model with PCA-reduced features produced positive results across test metrics. RMSE(12.22 K), which is the root-mean-square error between predicted and actual critical temperature. R2(0.86) meaning the model explains 86% of the variance in critical temperature.

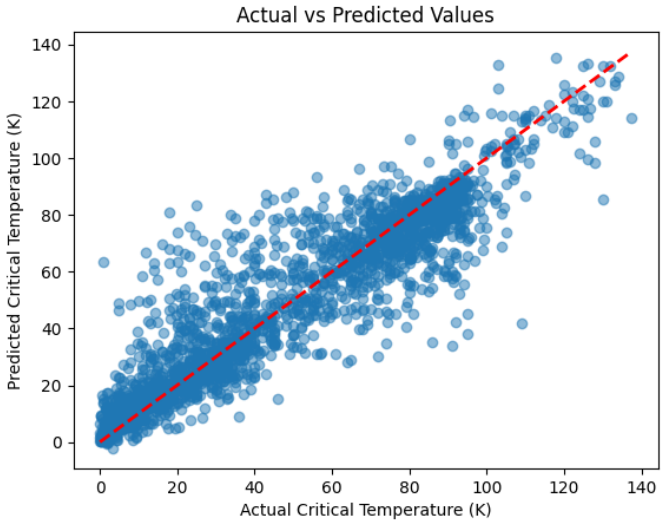

Figure 3: Actual vs predicted critical temperatures for the LightGBM model. The red dashed line indicates ideal prediction ($y = x$).

The scatter plot of true versus predicted critical temperatures showed mostly solid agreement with the identity line, yet heteroscedasticity was present with increasing error variance at increased values. Overall, the model does demonstrate the most certainty on materials with lower critical temperatures (< 50 K), which represented the majority of samples, and higher uncertainty regarding high-temperature superconductors. This asymmetry in prediction accuracy is likely associated to the limited number of high-T_c samples within training data and potential more complex physicochemical relationships in these compounds.

5 Discussion

The experimental results support machine learning, and specifically tree-based ensembles, have the potential for quantifiable predictive power to determine the critical temperature of superconductors with high accuracy. This research comes with several important observations: First, when we analyzed feature importance, we saw that valence electrons, electron affinity, and thermal conductivity played a central role in characterizing superconductivity.

The very strong negative correlations with features related to valence electrons and correlations with thermal conductivity features are consistent with BCS (Bardeen-Cooper-Schrieffer) theory, which proposes electron pair creation and phonon coupling as fundamental ways to reinforce the superconducting state. BCS theory states that an electron on its own does not cause superconductivity, but will when another electron share similar (or distinct) characteristics and the atoms phonon are coordinated enough to induce coupling.

Based on the data available, our empirical evidence supports there exist particular parameters which may indicate a material is more prone to this method of forming Cooper pairs. The other observations of significant positive correlation between conductivity features suggest phonon transport characteristics that could drive higher critical temperatures, likely due to the strength of the electron-phonon coupling.

Second, our dimensionality reduction analysis indicated that the feature space was high dimensional (81 features), however, we were able to account for about 95% of variance with only 15 principal components - quite efficient. This suggests that there is significant multicollinearity in the original features and suggests there are some latent features contributing to superconductivity, and thus this is an interesting insight related to modeling. The improved performance of PCA indicates superconductivity is likely governed by simpler fundamental material properties despite the quantum mechanics governing them, and this may simplify the search for new superconductors.

Third, while non-linear models outperformed linear models, they also quantitatively demonstrated complex non-linear interactions among the material properties (predictor variables) contributing to critical temperature. For example, the ensemble methods reduced the prediction error by on average 39% from the linear models. This evidence supports superconducting phenomena originates from interactions among multiple material properties, rather than additive effects. This newfound complexity

facilitates the development of physical theories explaining superconductivity across the entire material class, while potentially also demonstrating the value of machine learning models as complementary to physics-based models.

Nevertheless, there are more limitations to acknowledge. Despite the promising predictive capability, the model was still able to make accuracy predictions on average of about 12K, which is not insignificant percentage error for materials with low critical temperatures, and it can potentially become catastrophic if used for more practical use cases needing a highly controlled temperature.

Furthermore, while we felt we had a large ecosystem (21,263 samples), our data may still have had some degree of sampling bias as it could be indicative of a certain class of superconductors, likely defined by their usability and capacity to be replicated and had better historical documentation for classification andr report. For instance, it would be premature to extend orders views directly to new compositions or novel structures, and with specific mechanisms with and unconventional superconductors (a class notoriously difficult to explain), they may prove out of scope.

Finally, the black-box nature of ensemble learning models presents challenges for direct physical interpretations of the models, although indicators like feature importance offers potential better insights indirectly.

6 Conclusions and Future Work

This research illustrates the potential of machine learning techniques to predict the critical temperature of superconducting materials based on their physical and chemical descriptors (features). The optimal LightGBM model allowed for a 0.86 coefficient of determination (R^2) for predicting critical temperature, which means that 86% of the variance in critical temperature can be explained by the grouping of features together. While this result is far from perfect, it is significant for the field of materials science where historically, complex quantum phenomena such as superconductivity have been notoriously difficult to predict from classical descriptors (e.g. physical and chemical properties). Key findings include:

1. Tree-based ensemble techniques greatly outperform linear models for the outlined prediction trail, with Random Forest displaying a 39% lower RMSE against the best performing linear model in this research. This empirical gain substantiates the complexity of the non-linear relationships with regard to supervision material properties and superconductivity since linear models could not account for them.

2. The most important predictors of the critical temperature were valence-related features, electron affinity, and thermal conductivity, with weighted mean valence contributing to 20.4% of the total feature importance. This evidence also supports theoretical models on superconductivity that highlight electron configuration and phonon interactions.

3. Feature reduction methodologies were able to decrease model complexity while retaining high predictive capabilities, with PCA achieving 81.5% dimensionality reduction while keeping 95% of the total variance. This level of compressibility suggests superconductivity may be a function of much fewer specific material properties, than originally configured for. Future directions of work could explore the following directions:

    - Examining transfer learning approaches, by taking information from applicable materials science fields like thermoelectric or topologically insulators, where materials may behave similarly and be affected by the similarity of other physical properties.
    - Expanding the method to be predictive of and address alternate key properties, not just critical temperature.

References

[1] K. Onnes, "Further experiments with liquid helium. C. On the change of electric resistance of pure metals at very low temperatures, etc.," *Communications from the Physical Laboratory of the University of Leiden*, 1911.

[2] V. Stanev, C. Oses, A. G. Kusne, et al., "Machine learning modeling of superconducting critical temperature," *npj Computational Materials*, vol. 4, no. 1, pp. 1–14, 2018.

[3] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor," *Computational Materials Science*, vol. 154, pp. 346–354, 2018.

[4] L. Ward, R. Liu, A. Krishna, et al., "Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations," *Physical Review B*, vol. 96, no. 2, 2016.

[5] Y. Zhuo, A. Mansouri Tehrani, A. Ouyang, J. Brgoch, "Predicting the Band Gaps of Inorganic Solids by Machine Learning," *The Journal of Physical Chemistry Letters*, vol. 9, no. 7, pp. 1668–1673, 2018.

[6] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547–555, 2018.

7 CRediT statement

Arun Pavithran Rajasekaran: Conceptualization, Data curation & preprocessing, Writing – original draft preparation (Literature Review).

Jemilsan Jeyakumar: Conceptualization, Formal analysis & feature engineering, Methodology, Writing – original draft preparation (Abstract, Introduction & Methods).

Shabbir Kutbuddin: Software & model implementation, Hyperparameter optimization, Investigation & metrics calculation, Writing – original draft preparation (Hyperparameter Optimization).

Yuxiao Pu: Visualization, Software, Results interpretation, Writing – original draft preparation (Results, Discussion & Conclusions).

All authors: Writing – review & editing, Visualization review, Project administration.