Machine Learning-Based Phase Prediction and Structural Stability Analysis of High Entropy Nitrides

Priyanshu Bhatia

Department of Materials Science and Engineering

Indian Institute of Technology Kanpur

Abstract

High Entropy Nitrides (HENs) represent a distinctive class of advanced materials characterized by the incorporation of multiple principal elements bonded with nitrogen. Their diverse compositional space and configurational complexity endow them with exceptional properties, including **high hardness**, **oxidation resistance**, and **thermal stability**. Despite their promising potential, the prediction of structural stability and phase classification of HENs poses a significant challenge due to the absence of robust computational models.

This research addresses this critical issue by employing a **machine learning-based approach** to classify the phases and predict the structural stability of HENs. By utilizing a combination of **structural** and **thermodynamic descriptors**, several machine learning models were developed, including **K-Nearest Neighbors (KNN)**, **Random Forest (RF)**, **Support Vector Machine (SVM)**, and **Gaussian Naive Bayes (GNB)**. Techniques such as **ADASYN** were implemented to balance the dataset, thereby enhancing the performance of the models.

Among the various models, the KNN demonstrated the highest prediction accuracy and robustness. This work significantly contributes to the **data-driven discovery** of stable HENs, providing a pathway to expedite experimental synthesis and characterization efforts.

1 Introduction

High Entropy Nitrides (HENs) represent a cutting-edge class of **multi-principal element ceramics**. Unlike conventional compounds that typically involve one or two primary elements, HENs integrate four or more metallic elements in nearly equiatomic proportions. The incorporation of multiple elements introduces high **configurational entropy**, which stabilizes simple solid solution phases even at elevated temperatures. This thermodynamic advantage results in unique combinations of physical and chemical properties, including **superior hardness**, **thermal conductivity**, and **resistance to wear and corrosion**.

However, this same complexity that gives HENs their unique properties also poses significant challenges for researchers. The traditional methods of predicting phases based on binary or ternary phase diagrams are inadequate in the high-dimensional compositional space of HENs. Consequently, there is a growing interest in leveraging **machine learning (ML)** techniques to understand and predict the structural stability and phase formation in these materials. By utilizing advanced algorithms and computational datasets, ML offers a promising avenue for accelerating material discovery and optimizing the properties of HENs.

2 Research Objective

The primary aim of this study is to design and validate **machine learning models** that can achieve the following objectives:

1. **Accurate Phase Classification:** Effectively classify the phase stability of High Entropy Nitrides (HENs), distinguishing between **single-phase** and **multi-phase** compositions based on engineered features.

2. **Structural Stability Prediction:** Predict the structural stability of HENs utilizing a **semi-synthetic dataset** generated from both literature and computational modeling, thereby enhancing the reliability of predictions.

3. **Addressing Class Imbalance:** Tackle the challenge of class imbalance by employing oversampling techniques such as **ADASYN** (Adaptive Synthetic Sampling Approach

for Imbalanced Learning) to improve the predictive performance of the models.

3 Approach

To tackle the challenges associated with predicting the structural stability and phase classification of High-Entropy Nitrides (HENs), the following methodology was employed:

1. **Dataset Generation:** A semi-synthetic dataset was generated using an **atomic environment mapping** based structural plot, leveraging existing datasets to enhance the quality and diversity of the data.

2. **Sorting Criteria Development:** Two distinct sorting criteria were developed for both **quinary** and **quaternary compositions** to produce candidates for the single-phase class.

3. **Synthetic Data Generation:** The **ADASYN** (Adaptive Synthetic Sampling Approach for Imbalanced Learning) technique was implemented to generate a synthetic dataset from structural modeling and literature data, effectively oversampling the minority class to improve model training.

4. **Machine Learning Implementation:** Four machine learning algorithms were implemented: **K-Nearest Neighbors (KNN)**, **Random Forest (RF)**, **Support Vector Machine (SVM)**, and **Gaussian Naive Bayes (GNB)**. These models were trained on both balanced and imbalanced datasets to evaluate their performance.

5. **Feature Pool Design:** A comprehensive feature pool was designed, incorporating both structural and thermodynamic parameters to enhance the predictive capabilities of the models.

4 Dataset Construction and Feature Engineering

4.1 Data Generation

To construct a **representative dataset**, a comprehensive approach was adopted, integrating both literature data and **semi-synthetic samples**. Structural modeling tools were utilized to generate **atomic environment plots** and **structural descriptors** for various quaternary and quinary nitride compositions.

In addition to the generated data, supplementary information was sourced from **thermodynamic simulations** and **empirical observations**. This multifaceted data collection strategy ensures a robust dataset that captures the complexity and diversity of High Entropy Nitrides (HENs), facilitating more accurate predictions and analyses in subsequent machine learning applications.

Table 1: Thermodynamic Parameters for Selected HEN Compositions

Alloy	ΔS_{conf}	ΔH_{mix}	ΔG (Formation)
TiZrHfVNbTa-N	-0.1387	0.756	Multi-phase
(Zr-Ti-Cr-Nb-Si)N	13.3809	-42.400	Multi-phase
$(Al_{0.5}CrFeNiTi_{0.25})N_x$	13.3809	-22.720	Single-phase
(FeCoNiCuAlCrV)N	13.8644	-7.600	Single-phase
Hf-Nb-Ti-V-Zr-N	13.3809	0.160	Multi-phase

4.2 Sorting Criteria

Empirical sorting rules were applied to identify potentially stable single-phase compositions. These were derived from observed tendencies in MN-type nitrides:

- Atomic Packing Efficiency ($1.34 \leq X_p \leq 1.94$)
- Atomic Radius Mismatch ($1.89 \leq P_{rad} \leq 2.54$)
- Enthalpy of Mixing ($-2.72 \leq \Delta H_{mix} \leq 0.76$)

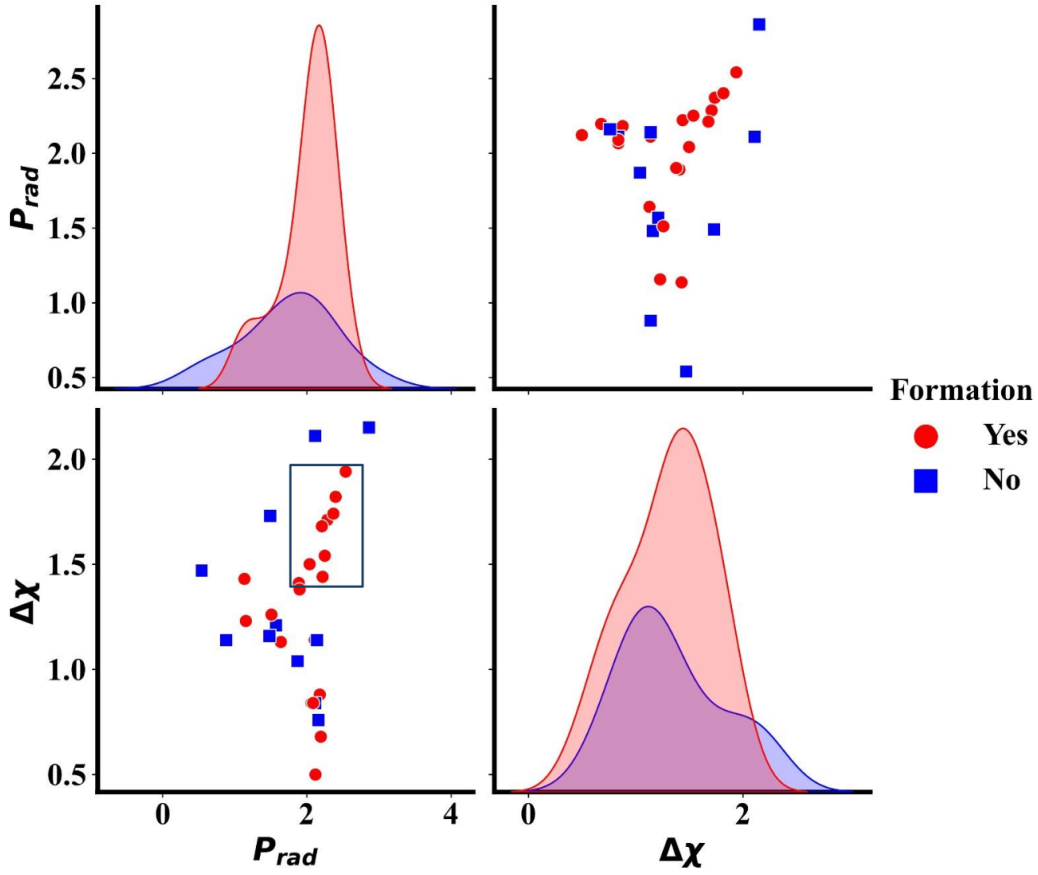

Figure 1: Correlation analysis between atomic radius mismatch (P_{rad}) and electronegativity difference ($\Delta\chi$). Red circles indicate successful single-phase formations, while blue squares denote multi-phase compositions.

4.3 Feature Selection

To develop effective **machine learning models** for phase prediction and stability classification of High Entropy Nitrides (HENs), a comprehensive set of **structural** and **thermodynamic descriptors** was curated. These features were selected based on their known relevance to phase formation, thermodynamic behavior, and configurational complexity in multi-component systems.

The input features used for training the machine learning models were meticulously designed through careful analysis of structural and thermodynamic attributes derived from **atomic environment mapping** and existing datasets. These features directly influence the stability and phase formation in HENs. The final feature pool includes:

- **Atomic Environment Mapping Parameters:** Extracted from structural plots representing local atomic arrangements.
- **Atomic Packing Efficiency (X_p):** Reflects how densely atoms are packed in the lattice.
- **Radius Ratio ($P_{\mathbf{rad}}$):** Accounts for mismatch in atomic sizes among constituent elements.
- **Enthalpy of Mixing ($\Delta H_{\mathbf{mix}}$):** Indicates thermodynamic stability of the mixture.
- **Electronegativity (Pauling and Mulliken):** Captures variations in bonding tendencies between elements.
- **Atomic Mismatch:** Quantifies lattice distortion due to size differences.
- **Pseudo Potential Ratio:** Encodes relative electronic interactions among atoms.
- **Entropy and Configurational Parameters:** Derived from mixing complexity across quaternary and quinary compositions.

These features were systematically curated from the generated structural mappings and semi-synthetic datasets developed during this study. The selection process was driven by their observed influence on phase differentiation and structural behavior across quaternary and quinary nitride compositions. Emphasis was placed on incorporating both **atomic-scale descriptors** and **thermodynamic estimators**, which collectively capture the configurational complexity and interaction dynamics within High Entropy Nitrides.

These engineered variables constituted the foundational input set for training the classification algorithms, enabling the models to effectively learn and generalize the underlying phase formation patterns in a high-dimensional compositional space.

5 Methodology

5.1 Handling Class Imbalance

In the dataset constructed for this study, **single-phase** High Entropy Nitride (HEN) compositions were significantly outnumbered by **multi-phase** samples. This inherent class imbalance presents a critical challenge to the performance and generalizability of machine learning classifiers, as models tend to exhibit bias toward the majority class.

To mitigate this issue, the **Adaptive Synthetic Sampling (ADASYN)** technique was employed. ADASYN generates synthetic samples for the minority class by adapting the density distribution in feature space. The algorithm prioritizes **harder-to-learn instances**, thereby enhancing the classifier's sensitivity to underrepresented data. This approach ensures a more balanced training dataset, facilitating improved phase discrimination in high-dimensional compositional spaces.

By employing ADASYN, the study aims to enhance the robustness of the machine learning models, allowing for more accurate predictions and better understanding of the structural stability and phase classification of HENs.

- **K-Nearest Neighbors (KNN)**: A non-parametric algorithm that classifies new samples based on the majority class among their k nearest neighbors in the feature space. Its simplicity and interpretability make it well-suited for structured datasets.

- **Random Forest (RF)**: An ensemble learning method based on decision trees, where multiple classifiers are trained on bootstrapped subsets of the data. The aggregated predictions help reduce variance and prevent overfitting.

- **Support Vector Machine (SVM)**: A robust classifier that constructs an optimal hyperplane to maximize the margin between different classes. It is particularly effective in high-dimensional spaces and scenarios with clear class separability.

- **Gaussian Naive Bayes (GNB)**: A probabilistic model based on Bayes' theorem with an assumption of feature independence and Gaussian distribution. Despite its simplicity, it provides competitive results on small- to medium-scale datasets.

Each model was trained and validated on both the original and the ADASYN-augmented datasets. Their performance was assessed using standard classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, allowing a comprehensive evaluation of each algorithm's predictive power in the context of HEN phase classification.

6 Model Comparison

The evaluation of various **machine learning models** is crucial for determining their effectiveness in classifying High Entropy Nitrides (HENs). The comparison of evaluation metrics and **ROC curves** for different models is illustrated in Figure 2. Key observations include:

- **Evaluation Metrics:** The bar chart on the left presents scores for multiple evaluation metrics, including **accuracy**, **precision**, **recall**, **F1 score**, and **Kappa**. All models, including **Support Vector Machine (SVM)**, **Decision Tree**, **Random Forest**, **Naive Bayes**, and **K-Nearest Neighbors (KNN)**, demonstrate comparable performance, with scores consistently above 0.85. This indicates that each model is capable of effectively classifying the phases of HENs.

- **ROC Curve Analysis:** The ROC curves on the right provide insights into the **true positive rates** against **false positive rates** for each model. The **Random Forest** model exhibits the highest area under the curve (AUC) at 0.97884, indicating superior performance in distinguishing between single-phase and multi-phase compositions. In contrast, the **Decision Tree** model shows the lowest AUC at 0.90248, suggesting it may be less effective in this classification task.

- **Model Robustness:** The close proximity of the ROC curves for **SVM**, **Naive Bayes**, and **KNN** indicates that these models are robust and reliable for phase classification, making them suitable candidates for further exploration in material discovery applications.

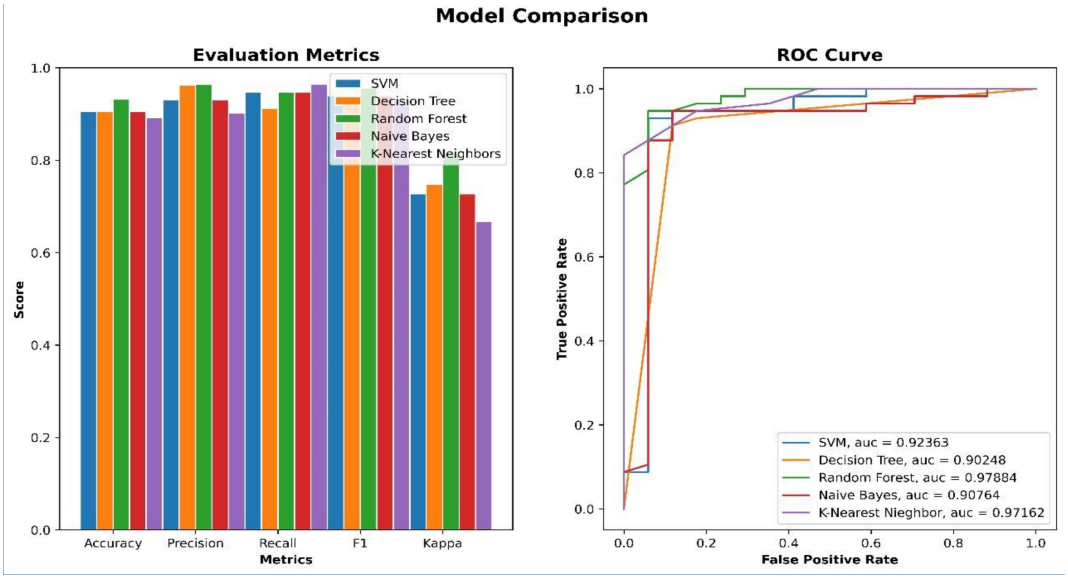

Figure 2: Comparison of evaluation metrics and ROC curves for various machine learning models used in classifying High Entropy Nitrides. The left panel shows evaluation scores, while the right panel illustrates the ROC curves for each model.

7 Analysis of Learning Curves and ROC Metrics

The analysis of **learning curves** and **Receiver Operating Characteristic (ROC)** metrics is essential for evaluating the performance of machine learning models. The left panel of Figure 3 displays the learning curves, which plot the **training** and **cross-validation accuracy scores** against the number of training examples.

The training accuracy, represented by the **red line**, remains consistently high, indicating that the model is effectively fitting the training data. In contrast, the cross-validation accuracy, shown in **green**, exhibits variability, suggesting that the model's **generalization capability** may be limited. This discrepancy indicates potential **overfitting**, where the model performs well on training data but struggles with unseen examples.

Such insights are crucial for refining model parameters and improving overall predictive performance, guiding future iterations of model development.

The right panel of Figure 10 presents the **ROC curves**, which provide valuable insights into the model's classification performance across various thresholds. The **area under the curve (AUC)** values for the **Multi Phase** and **Single Phase** classes are both 0.97,

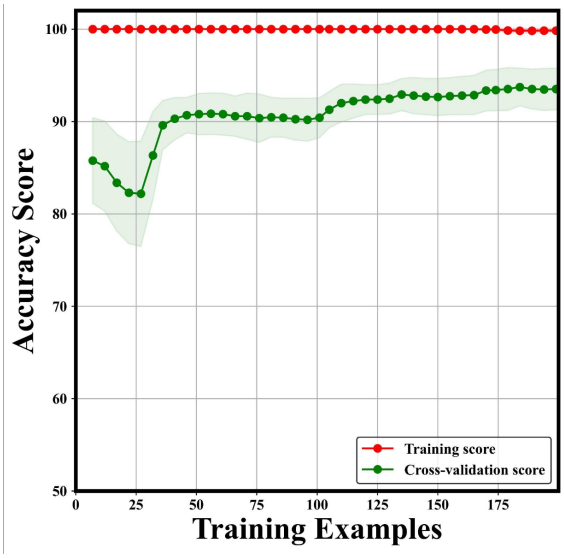

Figure 3: Learning curves showing training and cross-validation accuracy scores.

indicating strong discrimination ability.

Additionally, the **micro-average AUC** of 0.98 and **macro-average AUC** of 0.97 further confirm the model's robustness in distinguishing between classes. These metrics suggest that the model is well-suited for the classification task; however, continuous monitoring and validation are recommended to ensure sustained performance.

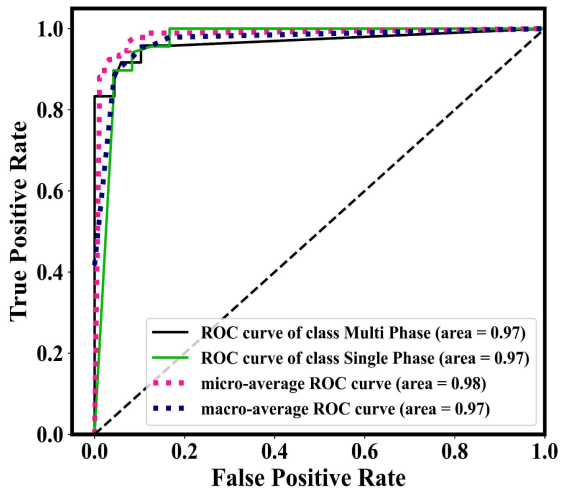

Figure 4: ROC curves for different classes and average metrics.

In summary, the analysis of the **learning curves** and **ROC metrics** underscores the importance of balancing **model complexity** and **generalization capabilities**. The findings indicate that while the model demonstrates **strong performance**, further refinement

may be necessary to enhance its generalization to new data. Continuous evaluation and adjustment will be essential to optimize the model's effectiveness in practical applications.

8 Results and Evaluation

8.1 Model Performance Before ADASYN

KNN achieved the highest training and test accuracy prior to oversampling:

- Training Accuracy: 99.63%
- Test Accuracy: 93.4%

8.2 Feature Importance Analysis

The bar chart presented in Figure 5 illustrates the **mean absolute differences** for various structural and thermodynamic descriptors relevant to High Entropy Nitrides (HENs). This analysis provides valuable insights into the relative importance of each feature in predicting **phase stability**. Key observations include:

- **Dominant Feature:** The descriptor r_A/r_C exhibits the highest mean absolute difference, indicating its significant influence on phase formation. This suggests that variations in **atomic radius ratios** are critical for determining stability in HENs.
- **Comparative Importance:** Other notable features include $\Delta\chi_{\text{mulliken}}$ and δ, which also show substantial mean absolute differences. Their contributions highlight the importance of **electronic** and **structural factors** in phase stability.
- **Less Influential Features:** Conversely, descriptors such as ΔH_{mix} and VEC demonstrate lower mean absolute differences, suggesting that they may have a lesser impact on the classification of **single-phase** versus **multi-phase** compositions.
- **Implications for Model Training:** The findings underscore the necessity of prioritizing features with higher mean absolute differences during model training, as they are

likely to enhance the **predictive accuracy** of the machine learning models employed in this study.

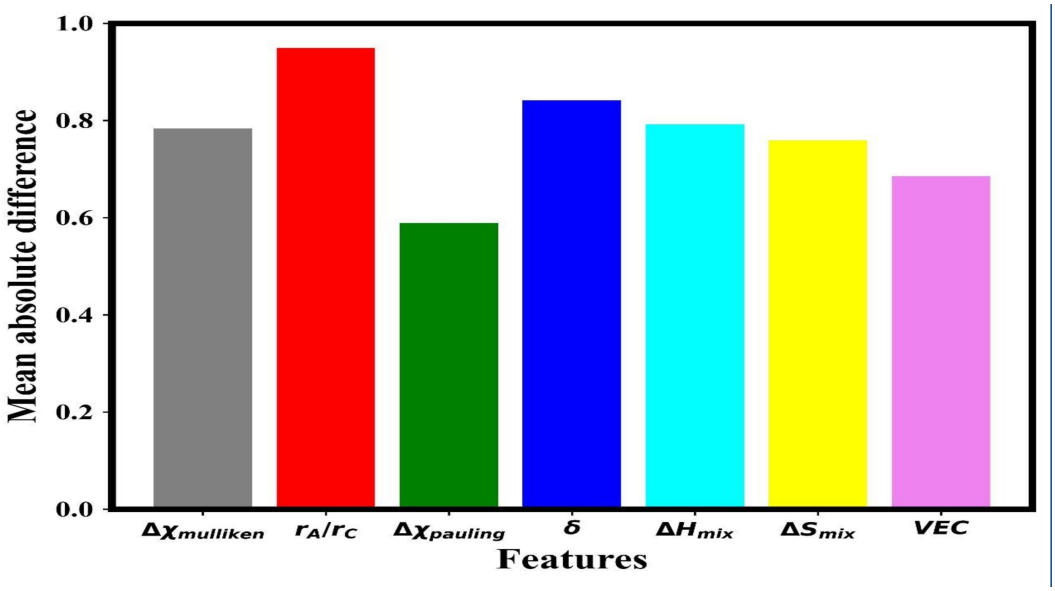

Figure 5: Mean absolute differences for various structural and thermodynamic descriptors. The height of each bar indicates the relative importance of the feature in predicting phase stability.

8.3 Pairwise Correlation Analysis

The pairwise correlation plots illustrated in Figure 6 provide a comprehensive examination of the interrelationships among various structural and thermodynamic descriptors pertinent to High Entropy Nitrides (HENs). The following observations can be drawn from the analysis:

- **Distinct Clusters:** The plots reveal clear clusters for **single-phase** (blue) and **multi-phase** (orange) compositions. This differentiation suggests that specific descriptors can effectively categorize the two phases, which is crucial for accurate classification.
- **Significant Correlations:** Notable positive correlations are observed between $\Delta\chi_{\text{mulliken}}$ and ΔH_{mix}, as well as between δ and S_{mixing}. These relationships indicate that these features are instrumental in influencing phase stability.
- **Variability in Multi-Phase Compositions:** The distribution of multi-phase data points across various descriptors highlights a greater variability in their properties. This

variability may complicate the classification process, necessitating advanced modeling techniques.

- **Feature Selection Importance:** The analysis underscores the critical need for judicious **feature selection** in model training. Certain descriptors exhibit stronger correlations with phase formation, which can enhance the predictive power of the model.

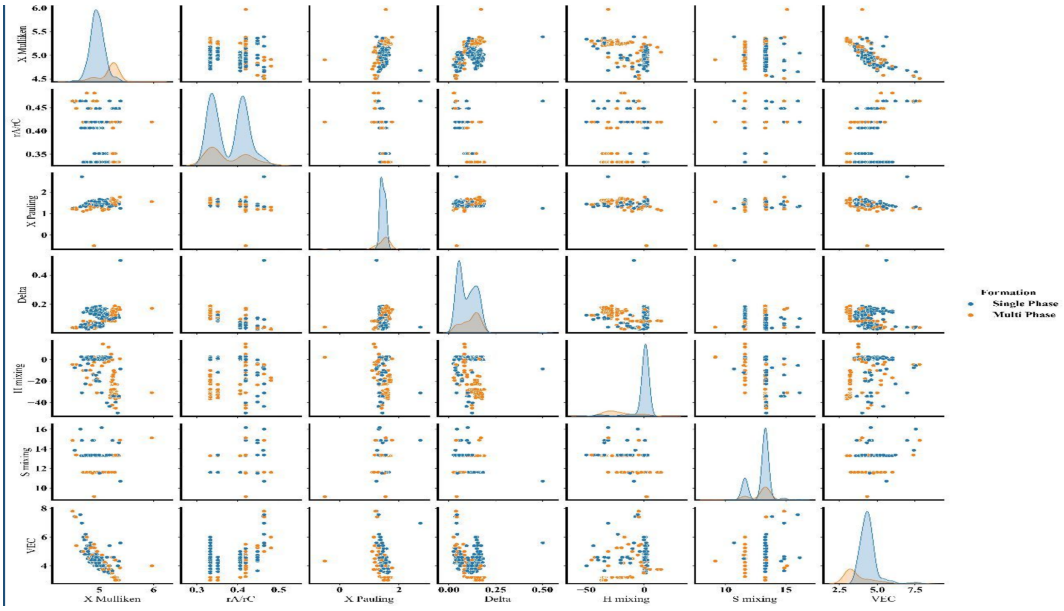

Figure 6: Pairwise correlation plots depicting the relationships between structural and thermodynamic descriptors for High Entropy Nitrides. Blue points represent single-phase compositions, while orange points denote multi-phase compositions.

8.4 Model Performance Metrics

8.4.1 Correlation Analysis

The correlation matrix presented in Figure 7 illustrates the relationships between various structural and thermodynamic descriptors used in the classification of High Entropy Nitrides (HENs). Key observations include:

- **Significant Correlations**: The descriptor $\Delta\chi_{\text{pauling}}$ shows a strong positive correlation with δ (0.52) and ΔH_{mix}(0.54).

- **Negative Correlations**: Notably, $\Delta\chi_{\text{mulliken}}$ exhibits a strong negative correlation with ΔH_{mix} (-0.39) and δ (-0.27), suggesting that as one parameter increases, the other tends to decrease, which could be pivotal for model training.

- **Diverse Relationships**: The varying degrees of correlation among descriptors highlight the complexity of the compositional space in HENs, emphasizing the need for careful feature selection in model development.

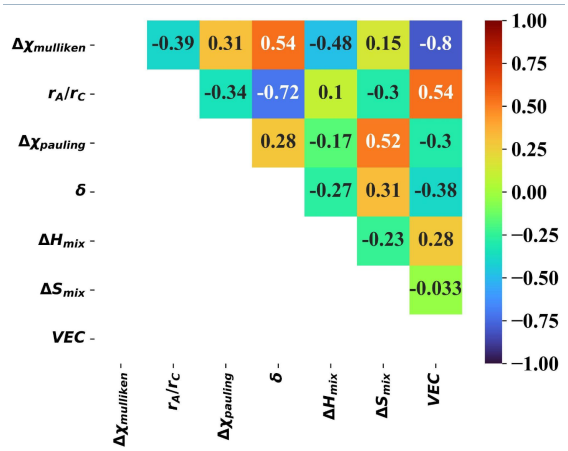

Figure 7: Correlation matrix of structural and thermodynamic descriptors.

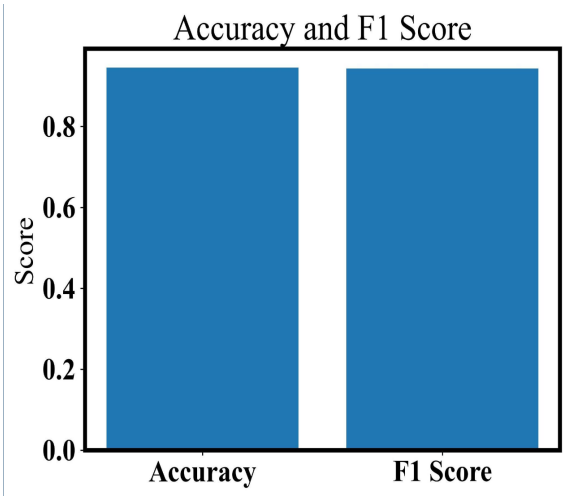

Figure 8: Model performance metrics: Accuracy and F1 Score.

8.5 Observations from the Normalized Confusion Matrix

The normalized confusion matrix provides critical insights into the performance of the machine learning model for classifying High Entropy Nitrides (HENs):

1. **Single Phase Accuracy**: The model exhibits a **99% accuracy** in predicting single-phase compositions, indicating a strong capability to identify stable materials.

2. **Multi Phase Performance**: The accuracy for multi-phase predictions stands at **76%**, suggesting that while the model is generally effective, it faces challenges in distinguishing complex multi-phase systems.

3. **False Negatives**: A mere **1% false negative rate** for single-phase classifications highlights the model's reliability in correctly identifying stable single-phase materials.

4. **False Positives**: The **24% false positive rate** for multi-phase predictions indicates a significant number of incorrect classifications, pointing to areas for improvement in model training and feature selection.

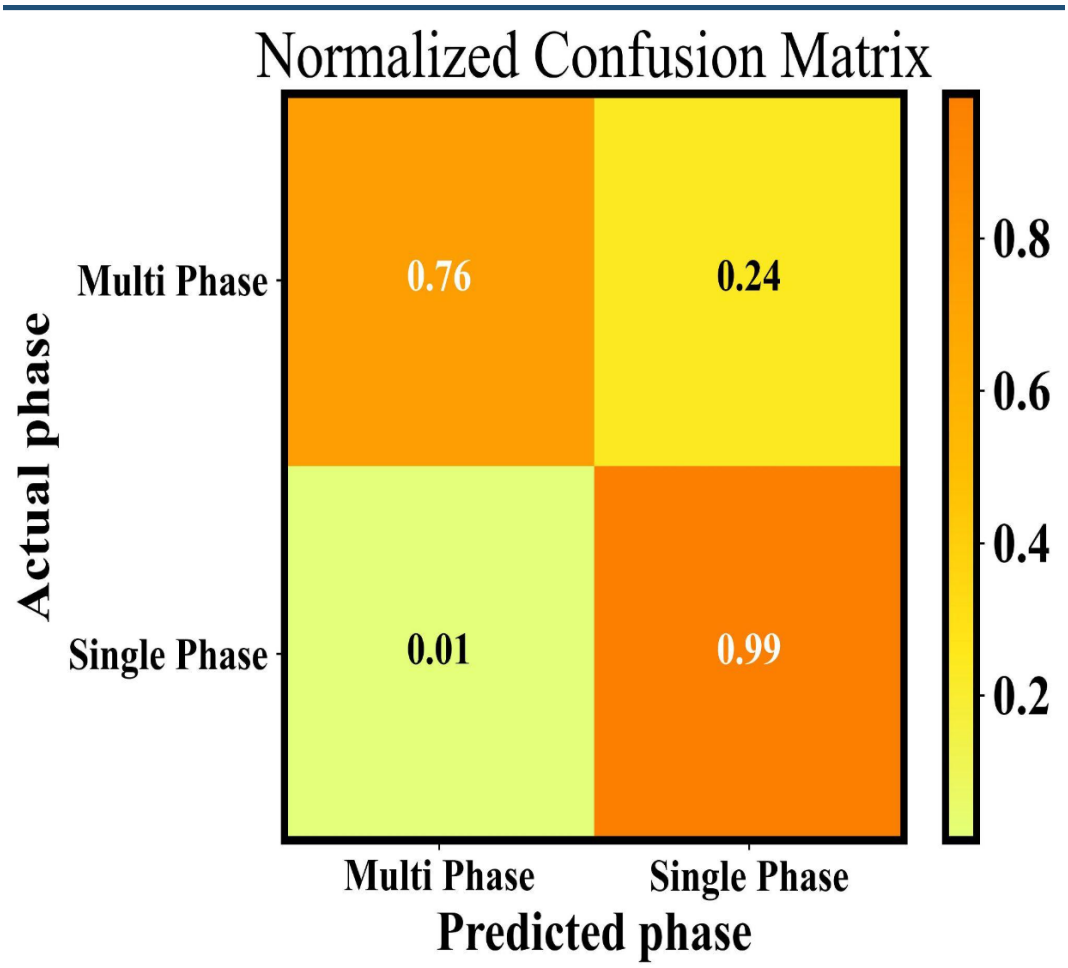

Figure 9: Correlation Matrix of Engineered Features

8.6 Model Performance After ADASYN

8.6.1 Observations from the ROC Curve

The Receiver Operating Characteristic (ROC) curve analysis provides valuable insights into the performance of the machine learning model for classifying High Entropy Nitrides (HENs). The following observations can be made from the ROC curves presented in Figure 10:

1. **True Positive Rate**: The ROC curves for both single-phase and multi-phase classifications demonstrate high true positive rates, with areas under the curve (AUC) of **0.95** for each class. This indicates that the model is effective in distinguishing between stable single-phase and multi-phase materials.

2. **Micro-Average Performance**: The micro-average ROC curve, with an AUC of **0.98**, suggests that the model performs exceptionally well across all classes, effectively capturing the overall performance in a multi-class setting.

3. **Macro-Average Performance**: The macro-average ROC curve, with an AUC of **0.95**, indicates that the model maintains a balanced performance across different classes, reinforcing its reliability in classifying both single-phase and multi-phase compositions.

4. **Model Robustness**: The proximity of the ROC curves to the top-left corner of the plot signifies a robust model with a low false positive rate, further validating its utility in practical applications for material classification.

These findings highlight the model's strong predictive capabilities, particularly in the context of high-dimensional compositional spaces inherent in HENs. Continuous refinement and validation of the model will be essential for enhancing its accuracy and applicability in material discovery.

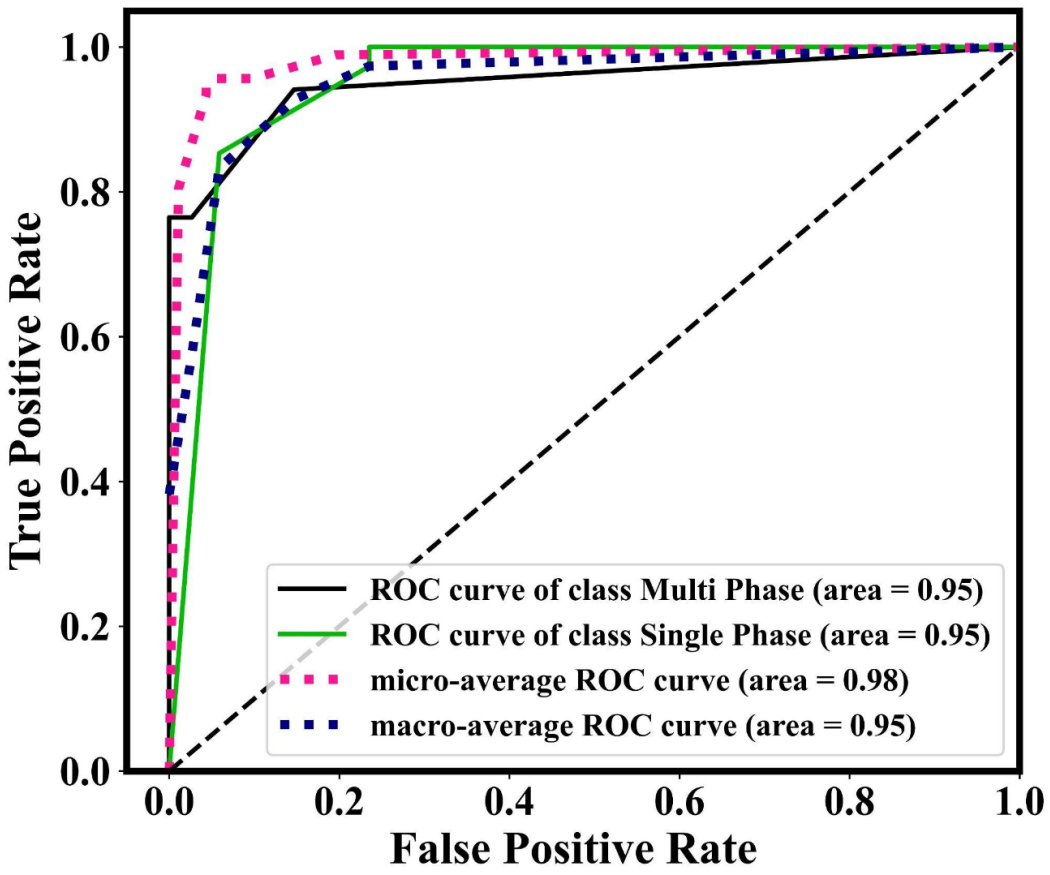

Figure 10: ROC curves for classifying High Entropy Nitrides, illustrating the performance of the model across different phases.

9 Discussion

The **KNN model** consistently outperformed other models in terms of both **accuracy** and **generalizability**. The implementation of **ADASYN** proved effective in mitigating the imbalance problem, leading to more reliable classification of **single-phase compositions**.

While the **Random Forest** and **Support Vector Machine (SVM)** models also demonstrated strong performance, they were slightly more prone to **overfitting**. In contrast, the **Gaussian Naive Bayes** model, although fast and efficient, exhibited relatively lower accuracy due to its strong distributional assumptions.

These findings highlight the importance of selecting appropriate models and techniques for the classification of High Entropy Nitrides (HENs), emphasizing the need for ongoing

refinement and validation to enhance predictive performance in practical applications.

10 Conclusion

This study effectively demonstrates the application of **machine learning models** for **phase prediction** and **stability analysis** of **High Entropy Nitrides (HENs)**. By leveraging **structural** and **thermodynamic features**, the proposed framework accurately identifies **stable compositions**, particularly after class rebalancing using the **ADASYN technique**. This research not only facilitates **data-driven material discovery** but also lays the groundwork for future investigations into **high-entropy ceramics** and **alloys**.

11 Impact

The **K-Nearest Neighbors (KNN)** model achieved impressive **performance metrics**, with an **accuracy** of **99.63%** on the **training dataset** and **93.4%** on the **test dataset** following **cross-validation**. These results underscore the model's **effectiveness** in predicting the **structural stability** of **High Entropy Nitrides (HENs)**.

Acknowledgments

The author gratefully acknowledges the **Department of Materials Science and Engineering** at **IIT Kanpur** for their **guidance** and **support**.

References

1. Yeh, J. W., et al. "Nanostructured **high-entropy alloys** with multiple principal elements: Novel alloy design concepts and outcomes." *Advanced Engineering Materials*, 2004.

2. Zhang, Y., et al. "Microstructures and properties of **high-entropy alloys**." *Progress in Materials Science*, 2014.

3. He, F., et al. "Machine learning approaches in the exploration of **high entropy alloys**." *npj Computational Materials*, 2021.

4. Haibo He, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." *IEEE International Joint Conference on Neural Networks*, 2008.

5. Zhang, Y., et al. "A comprehensive review on **high-entropy alloys**: Design, processing, and properties." *Materials Science and Engineering: R: Reports*, 2019.

6. Miracle, D. B., and J. Y. Zhang. "A fundamental look at **high-entropy alloys**." *Nature*, 2019.

7. Cantor, B., et al. "Microstructural development in **high-entropy alloys**." *Materials Science and Engineering: A*, 2004.

8. Liu, C., et al. "High-entropy alloys: A new era of **materials science**." *Materials Today*, 2016.