Internship Report

on

Machine Learning Based Prediction of Valence Electron Concentration and Phase Classification in High-Entropy Alloys for Hydrogen Storage Applications

completed at

CSIR-INSTITUTE OF PETROLEUM (IIP) CERTIFIED INSTITUTE

(ISO 9001: 2015)

DEHRADUN (INDIA) DURING THE PERIOD

9th June, 2025 – 18th July, 2025

Submitted in partial fulfillment of the requirements for Summer Training Internship

Priyanshu Bist (102208010)

Mechanical Engineering Department
THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY, PATIALA

Under the guidance of

Dr. Shailesh Kumar Singh
Senior Scientist

Dr. Bashista Kumar Mahanta
Senior Project Associate

DECLARATION

I hereby declare that work done in the Internship Report entitled "Machine Learning Based Prediction of Valence Electron Concentration and Phase Classification in High-Entropy Alloys for Hydrogen Storage Applications" submitted towards partial fulfilment of the requirement for completion of Summer Internship Training at ***Council Of Scientific And Industrial Research–Indian Institute Of Petroleum (CSIR–IIP)*** *is an authentic record of work carried out by me under the supervision and guidance of Dr. Shailesh Kumar Singh, Senior Scientist; Dr. Bashista Kumar Mahanta, Senior Project Associate. This is the original work done by me. No part of the matter embodied in this report has been submitted to any other university or institute. I hereby declare that this written submission reflects my ideas expressed in my own words. Where I have included the ideas and words of others, I have appropriately cited and referenced the sources. I affirm that I have adhered to all principles of academic honesty and integrity. There has been no misrepresentation, fabrication, or falsification of any idea, data, fact, or source in my submission.*

Priyanshu Bist (102208010)

Date: 18-07-2025

This is to certify that the above declaration made by the student concerned is correct to the best of my/our knowledge and belief.

Dr. Shailesh Kumar Singh
Senior Scientist

Dr. Bashista Kumar Mahanta
Senior Project Associate

Acknowledgement

At the start, I would like to thank my mentors, Dr. Shailesh Kumar Singh and Dr. Bashista Kumar Mahanta, who provided crucial direction, comments, and ideas throughout the report. I would like to express my sincere gratitude and profound respect. They have been a tremendous asset to my project and a valuable source of technical expertise. I will always be grateful to them and their consistent, timely counsel and time-saving efforts. I would like to express our gratitude to everyone who helped with this report, whether directly or indirectly. Additionally, I express my sincere appreciation to the research assistants who diligently contributed to the success of this project.

Above all, I want to express my gratitude to God and my family for their unwavering support and love. I am grateful for their perseverance and sacrifice since they always wanted the best for me.

Priyanshu

Priyanshu Bist (102208010)

Abstract

High entropy and compositionally complex alloys have become attractive options for storing hydrogen as we move toward sustainable energy systems. The complex multi-elemental nature of HEAs provides notable benefits in terms of hydrogen absorption and desorption kinetics, metal-hydrogen (M-H) interactions, and the influence of external parameters like temperature, pressure, and defects. HEAs have become promising materials for hydrogen storage **[1]**. Computational modeling and machine learning are demonstrated to be crucial for accelerating alloy design and optimization. Slow kinetics, low cycling stability, and activation requirements remain problems. Multidisciplinary efforts are needed to solve these challenges in order to enhance storage capacity and kinetics, expand datasets for improved machine learning predictions, and research microstructural effects **[2]**. Despite its potential, the enormous dimensionality of the compositional space makes it difficult to anticipate the phase development and electronic behavior of HEAs. This study utilizes machine learning (ML) techniques to predict two key properties of high-entropy alloys (HEAs): (i) Valence Electron Concentration (VEC), an essential electronic descriptor influencing phase stability, and (ii) phase type (solid solution, intermetallic, or mixed), which significantly affects hydrogen absorption/desorption behavior. A curated dataset compiled from the research paper of **Iman et al. [3]** was used to train and validate the predictive models. VEC prediction was performed using Support Vector Regression (SVR), while phase classification was addressed using a stacked ensemble model integrating XGBoost, Linear Discriminant Analysis (LDA), and Gaussian Naïve Bayes classifiers. Extensive data preprocessing, including normalization, outlier removal via Isolation Forest, and correlation-based feature selection, was employed. The SVR model, using a linear kernel with $C = 0.1$, achieved excellent performance in predicting VEC, with an R^2 score of **0.9960**, a mean absolute error (MAE) of **0.0672**, and a mean squared error (MSE) of **0.0045**. The phase classification stacked ensemble model yielded high evaluation metrics, including accuracies of **86.07%** (multiclass classification of 7 different Phases BCC, FCC, BCC+FCC, BCC+IM, FCC+IM, IM) and **87.48%** (multiclass classification of SS, IM, and SS+IM), with consistently high precision, recall, F1-score, and specificity, highlighting the model's robust multiclass classification capability. In the feature importance analysis, **VEC**, **atomic radius**, **enthalpy**, and **melting point** emerged as the most influential descriptors for phase classification, while **no. of valence electrons**, **DFT calculated total energy**, and **outer shell electrons** were the key predictors for VEC in the SVR model. These insights highlight the critical role of electronic and thermodynamic parameters in accurately modeling HEA behavior using machine learning.

Keywords: Hydrogen storage, High-entropy alloys (HEAs), Valence Electron Concentration (VEC), Machine learning, Phase multiclass classification.

Nomenclature

Term	Full form
HEAs	High Entropy Alloys
SVR	Support Vector Regression
XGBoost	eXtreme Gradient Boosting
LDA	Linear Discriminant Analysis
R-squared (R^2)	Coefficient of Determination
MAE	Mean Absolute Error
MSE	Mean Squared Error
SS Phases	Solid Solution
IM Phases	Intermetallic
SS+IM Phases	Phase containing a mixture of Solid Solutions and Intermetallic
AM Phases	Amorphous Phases
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
F1 Score	F-measure
BCC Phases	Body Centered Cubic
FCC Phases	Face-Centered Cubic
Pauling_EN	Electronegativity (Pauling scale)
DFT_LDA_Etot	DFT-calculated total energy
Atomic_radius_calculated	Weighted average atomic radius
Pauling_EN_div	Standard deviation of electronegativity
Atomic_radius_calculated_dif	Standard deviation of atomic radius
E_per_el	Energy per atom

List of Figures

List of Tables

Signature

Table of Contents

1. Introduction

As the transition toward sustainable energy systems accelerates, hydrogen has emerged as a leading candidate for clean and efficient energy storage [4]. However, the safe and practical storage of hydrogen remains a longstanding challenge. Traditional hydrogen storage materials, such as metal hydrides, a class of compounds made of metal and hydrogen, have attracted a lot of interest because of their significant hydrogen-storage capabilities, making them show promise for energy systems based on hydrogen. They are appealing for applications ranging from portable devices to electric cars and renewable energy systems because of their qualities, which include high energy density, affordability, and environmental friendliness. However, there are several obstacles in the way of their development for hydrogen storage, most notably the need to find materials with a high capacity for storing hydrogen while preserving stability, safety, and economic viability [5]. Recent advances in high-entropy alloys (HEAs), characterized by their multi-principal element composition, have propelled these materials to the forefront of hydrogen storage research.

1.1.High Entropy Alloys

HEAs are a class of materials formed from five or more principal elements mixed in near-equimolar ratios, resulting in a high configurational entropy that stabilizes novel solid solution phases. Because of their remarkable mechanical qualities, wear resistance, and thermal stability, HEAs are appropriate for cutting-edge applications in biomedical, automotive, and aerospace engineering, and are also in hydrogen storage applications [6]. Body-centred cubic (BCC) shaped HEAs are of special interest because of their exceptional hydrogen storage capabilities, structural stability, and adjustable thermodynamic characteristics. The rates of hydrogen absorption and desorption are greatly impacted by high configurational entropy, severe lattice distortions, and slow diffusion. These factors also affect the modifiable interstitial sites for hydrogen incorporation [1]. Kinetic properties, specifically, the rates of hydrogen absorption and desorption, and the reversibility over repeated cycles, are critical for technological applications. The ability to synthesize and design single-phase HEAs with desirable phase types (solid solution, intermetallic, mixed) is intimately tied to their hydrogen behaviour and has fuelled a surge in experimental and computational investigations [7].

1.2. Role of Machine Learning (ML) and Potential Challenges

Despite the progress, several limitations persist. These include slow absorption/desorption kinetics, low cycling stability, and activation barriers, all of which hamper the practical adoption of hydrogen storage systems. Addressing these challenges necessitates a deep understanding of the intricate relationships between composition, electronic structure, phase stability, and microstructure. Given the enormous compositional and structural complexity of HEAs, conventional trial-and-error experimental approaches are insufficient for efficiently exploring potential design spaces. Beyond the confines of conventional computer methods, the broad impact of ML has increased their applicability and importance to a broad range of scientific domains. ML models' capabilities are constantly being improved. With the availability of strong computer resources, sophisticated algorithms, and enormous volumes of data from calculations or experiments. These characteristics make machine learning (ML) a very promising method to address the difficulties in theoretically modelling HEAs [8].

ML offers a data-driven paradigm for predicting materials properties, enabling:

1. Rapid screening of candidate compositions across vast multi-dimensional spaces,
2. Quantitative modelling of structure–property relationships,

3. Acceleration of discovery cycles by reducing the need for costly and time-consuming synthesis and testing.

The parameters, such as VEC, Phase, and H/M ratio, have been identified as integral factors for insightful predictions [9]. For instance, the successful application of ML requires robust, curated datasets often sourced from experimental literature or calculated via theoretical formulas or software. Nonetheless, significant challenges remain, such as:

1. Data scarcity: The quality and diversity of training datasets directly influence model generalizability, yet large, standardized datasets are still emerging in this field.
2. Dimensionality: The vast compositional freedom of HEAs demands advanced algorithms and robust validation frameworks to avoid overfitting and maintain prediction reliability.
3. Physical Interpretability: ML models must be integrated with domain knowledge such as phase diagrams and thermodynamic rules, to provide physically interpretable insights and actionable guidance for experimentalists.

1.3. Aims and Objectives

In this work, ML is used for two specific goals:

1. **Regression**: Predicting the **Valence Electron Concentration (VEC)** of HEAs, a critical descriptor that correlates with phase stability and hydrogen solubility.
2. **Classification**: Classification-based prediction of the **phase type** (e.g., FCC, BCC, Intermetallic) of a HEA based on compositional and other parameters.

2. Methodology

2.1. Dataset Acquisition and Feature Engineering

The initial dataset consisted of **10387 HEA compositions**, sourced from the database provided in the research paper: **Iman et al. [3]**.

Data cleaning involved:

- Removing all AM (Amorphous) Phases and a Mixture of Solid Solution and Amorphous (FCC+AM, BCC+AM, BCC+FCC+AM, and IM respectively), reducing it to **8261.**

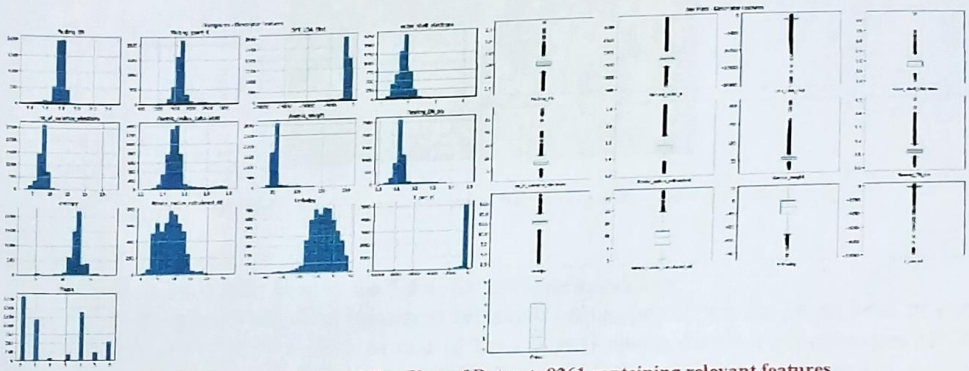


Fig 1. Histogram Plots and Box Plots of Dataset: 8261 containing relevant features

- Outlier removal using **Isolation Forest**, with a contamination rate of 5%, yielding a final dataset of **5396 samples**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	VEC	Pauling_EN	lting_point	T_LDA_Et	shell_ele	valence_el	radius_ca	omic_weig	uling_EN_	entropy	dius_calc	Enthalpy	E_per_el	Phase	
2	8.66055	1.827339	1793.838	-1452.87	1.577982	8.66055	1.552294	59.09118	0.100801	14.19763	7.034592	0.848413	-242.145	FCC+IM	
3	8.699115	1.829735	1783.858	-1428.66	1.584071	8.699115	1.547788	58.73494	0.098578	14.04436	6.493761	1.481714	-238.11	FCC+IM	
4	8.735043	1.831966	1774.56	-1406.1	1.589744	8.735043	1.54359	58.40305	0.096408	13.87132	5.926346	2.083425	-234.35	FCC+IM	
5	8.768595	1.83405	1765.877	-1385.03	1.595041	8.768595	1.539669	58.09311	0.094288	13.66823	5.323143	2.655556	-230.839	FCC	
6	8.272727	1.815455	1682.786	-1263.13	1.727273	8.272727	1.503636	55.00109	0.109326	14.69717	8.185956	-2.71074	-210.522	FCC	
7	7.615385	1.773077	1721.304	-1199.16	1.769231	7.615385	1.543077	53.90354	0.141388	16.00546	9.474733	-12.4497	-171.308	BCC+FCC+IM	
8	8.122807	1.805789	1691.571	-1248.54	1.736842	8.122807	1.512632	54.75077	0.11875	15.44525	8.580031	-5.257	-178.363	FCC	
9	7.983051	1.79678	1699.76	-1234.94	1.745763	7.983051	1.521017	54.51742	0.126242	15.76181	8.888277	-7.45763	-176.42	BCC+FCC	
10	7.852459	1.788361	1707.412	-1222.23	1.754098	7.852459	1.528852	54.29938	0.132308	15.92421	9.131192	-9.36307	-174.604	BCC+FCC	

Table 1. Snippet of csv file containing a cleaned dataset containing 5396 samples

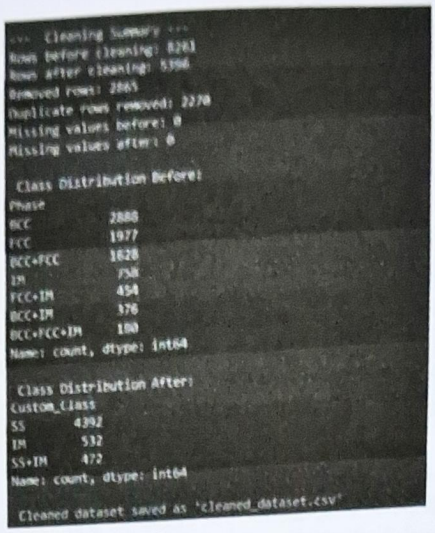


Fig 2. Snippet of Dataset cleaning (Outlier Removal and Duplicate Rows Removal)

2.2. Target Variables

a. **Regression Task**: Predict the **VEC** for each composition.

b. **Classification Task**: The classification task involves predicting the phase label of each sample at two levels: first, as one of the original **seven distinct phase types** (FCC, BCC, BCC+FCC, IM, FCC+IM, BCC+IM, and IM), and second, as part of a more logical three-class system. In the simplified classification, the seven original phases are grouped into **three broader categories**: SS (Solid Solution), which includes FCC, BCC, and BCC+FCC; IM (Intermetallic), which includes IM; and SS+IM (Mixed Phases), which includes FCC+IM and BCC+IM. Both classification levels are used to evaluate the model's performance.

2.3. Features/Attributes

Descriptor Features:

These are weighted or derived properties of the alloy composition:

1. Average Melting Point (Tm)
2. Number of Valence Electrons (NVC)
3. Mixing Entropy (ΔS_{mix})
4. Valence Electron Concentration (VEC)
5. Mixing Enthalpy (ΔH_{mix})
6. Electronegativity Deviation ($\Delta\chi$)
7. Total Energy Calculated by DFT (E)
8. Outer Shell Electrons (OSHE)
9. Atomic Size Difference (δ)
10. Phase
11. Average Electronegativity ($\bar{\chi}$)
12. Atomic weight
13. Atomic radius

2.4. Detailed Workflow of Pipelines (General)

Workflow begins with data preprocessing to ensure model reliability. The initial steps involve eliminating missing and duplicate records, as evident in Fig. 2, as their presence may introduce bias into the modeling process. Given that most machine learning algorithms require numerical input, categorical variables (such as 'Phase') were numerically encoded using label encoding, which transforms each unique category into an integer label.

To develop an understanding of the dataset and identify potential issues, such as data imbalance, an exploratory data analysis (EDA) was performed. Here, correlation matrices were plotted using heatmaps to visualize dependencies among features Fig. 3. Features with extremely high pairwise correlation coefficients (typically above 0.95) were removed, as retaining them could lead to redundancy and reduce the interpretability of downstream models. Further, data visualizations (such as box plots and histograms), Fig. 1, were employed to inspect the distribution and spread of each feature. Outlier samples deviating from the normal pattern were detected and removed through the Isolation Forest algorithm, an unsupervised technique that isolates anomalies by their attribute values, thus improving the overall quality of the dataset. Standardization of feature values followed, using z-score scaling to ensure that all variables contribute equally to model training, which is especially important for algorithms sensitive to feature scale.

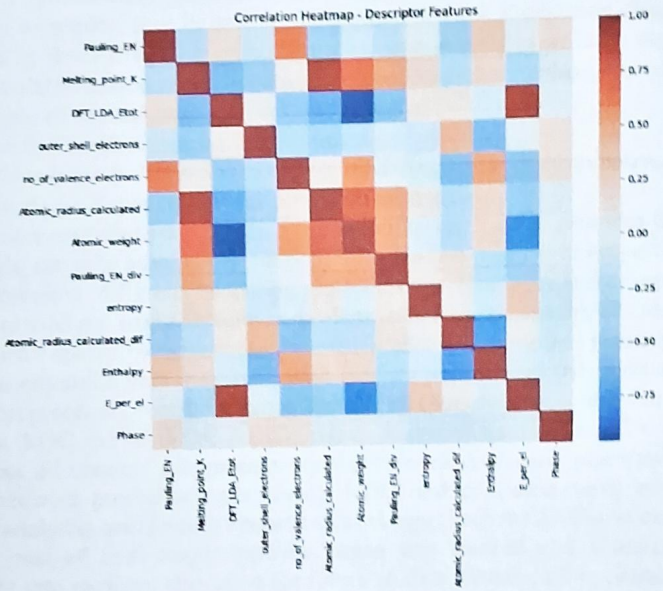


Fig 3. Co-relation Heatmap

2.4.1. Regression Pipeline

The regression model aimed to predict the Valence Electron Concentration (VEC) a critical physical parameter influencing material phase stability. Support Vector Regression (SVR) was

chosen due to its flexibility in modeling both linear and non-linear relationships. SVR extends the support vector machine (SVM) to regression tasks by constructing an optimal hyperplane (or surface) that fits the data within a specified error margin (epsilon), and penalizes deviations outside this margin using a regularization parameter (C). The use of different kernel functions (linear and radial basis function, RBF) allows the model to adapt to complex, nonlinear data structures. A grid search was conducted over SVR hyperparameters (C, kernel type, and, for the RBF kernel, gamma), utilizing three-fold cross-validation to reliably estimate model performance and avoid overfitting. Model performance was reported using standard regression metrics: R^2 score (indicating variance explained), root mean squared error (RMSE) (sensitive to large errors), and mean absolute error (MAE) (giving equal weight to all errors). To further diagnose model soundness, the residuals are checked via scatter plots (checking for randomness), and Quantile-Quantile (Q-Q) plots (for error normality).
Recognizing the importance of interpretability, permutation feature importances measuring the change in model performance when randomly shuffling each feature were computed to identify which variables most strongly influence VEC predictions. Only features with positive importance were retained for a refined model fit, and the new model's performance was compared to the model trained on all features. This allowed for a more parsimonious and possibly more generalizable predictive model.

2.4.2. Classification Pipeline

In real-world materials datasets, some phases are far less common than others, potentially leading to biased models. To address this, SMOTETomek, a hybrid resampling strategy, was applied. SMOTE synthetically generates new samples for minority classes by interpolating between existing examples, thus balancing the dataset. Tomek Links then cleans overlapping class boundaries by discarding ambiguous samples, promoting cleaner class separation.
To further boost classification performance,a stacking ensemble, a two-layer architecture that integrates strengths of several base learners was made:

- XGBoost for its efficiency with nonlinear data
- Linear Discriminant Analysis (LDA) for modeling linear discriminative boundaries
- Gaussian Naive Bayes (GNB) for probabilistic reasoning

These base learners contribute their predictions to a meta-learner that provides the final output. Stacking, though, can take more computational time has been shown to outperform individual models by harnessing different modeling perspectives. Hyperparameter optimization for XGBoost was carried out using Optuna, a modern, efficient framework for navigating large, complex parameter spaces by prioritizing promising configurations and pruning unproductive runs early. The ensemble was evaluated with metrics appropriate for multiclass problems: overall accuracy, precision, recall, F1-score, specificity (per-class true-negative rates), and the area under the ROC curve (ROC-AUC), which summarizes the model's discrimination capability across all classes. Interpretative tools included confusion matrices (summarizing correct and incorrect predictions per class), ROC and precision-recall curves (showing performance trade-offs), and feature importance rankings (from the XGBoost component). The computational cost of each major pipeline phase was tracked and visualized, providing practical insight into resource allocation for future studies. Finally, all hyperparameter search results and key best-performing configurations were recorded and reported for transparency and reproducibility.

2.5. Algorithm Summarized

2.5.1 Regression Pipeline Overview

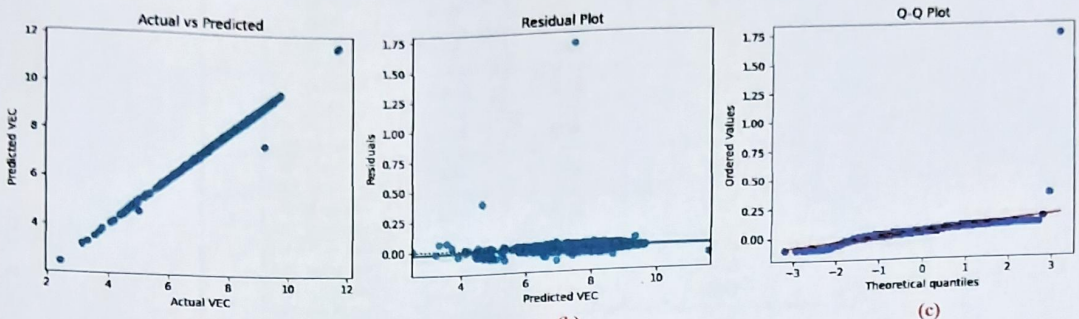
1. Data Acquisition
2. Data Cleaning
3. Label Encoding
4. Exploratory Analysis:
5. Feature Reduction
6. Outlier Detection
7. Feature Scaling
8. Train-Test Split
9. Hyperparameter Search
10. Model Training
11. Performance Evaluation
12. Permutation Importance
13. Refit & Comparison
14. Result Documentation

2.5.2 Classification Pipeline Overview

1. Data Loading
2. Data Cleaning
3. Class Relabeling & Encoding
4. EDA & Visualization
5. Outlier Removal
6. Standardization
7. Imbalance Handling
8. Hyperparameter Optimization
9. Stacked Model Construction
10. Model Training
11. Metrics Computation
12. Interpretation
13. Time Tracking
14. Result Reporting

3. Results

3.1. Regression Results:



(a) (b) (c)

Fig 4. (a) Actual vs Predicted Curve, (b) Residual Plot, and (c) Q-Q Plot of Most Optimal Hyperparameter

	feature	importance
0	no_of_valence_electrons	0.136996
1	DFT_LDA_Etot	0.107734
2	outer_shell_electrons	0.071901
3	Pauling_EN	0.045756
4	Atomic_radius_calculated	0.004532
5	Melting_point_K	0.001719
6	Atomic_radius_calculated_dif	0.001494
7	Pauling_EN_div	0.000037
8	E_per_el	0.000031
9	entropy	0.000004
10	Phase	0
11	Enthalpy	-0.00003

(a)

	feature	importance
0	no_of_valence_electrons	0.136996
1	DFT_LDA_Etot	0.107734
2	outer_shell_electrons	0.071901
3	Pauling_EN	0.045756
4	Atomic_radius_calculated	0.004532
5	Melting_point_K	0.001719
6	Atomic_radius_calculated_dif	0.001494
7	Pauling_EN_div	0.000037
8	E_per_el	0.000031
9	entropy	0.000004

(b)

Table 2. (a) Feature Importance Score, (b) Feature Importance Score (Positive Features Only)

	Metric	All Features	Positive Features
0	R2	0.995178	0.99589
1	RMSE	0.07843	0.072412
2	MAE	0.050483	0.040424

Table 3. Evaluation Scores of the Most Optimal Hyperparameter

	params	R2	MAE	MSE
0	**{'C': 0.1, 'kernel': 'linear'}**	**0.996**	**0.0672**	**0.0045**
1	{'C': 1, 'kernel': 'linear'}	0.996	0.0672	0.0045
2	{'C': 10, 'kernel': 'linear'}	0.996	0.0672	0.0045
3	{'C': 100, 'kernel': 'linear'}	0.996	0.0672	0.0045
9	{'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}	0.9932	0.0878	0.0077
11	{'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}	0.9928	0.0901	0.0081
7	{'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}	0.9898	0.1076	0.0116
8	{'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}	0.9801	0.15	0.0225
10	{'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}	0.9801	0.15	0.0225
6	{'C': 1, 'gamma': 0.01, 'kernel': 'rbf'}	0.9752	0.1673	0.028
5	{'C': 0.1, 'gamma': 0.001, 'kernel': 'rbf'}	0.95	0.2375	0.0565
4	{'C': 0.1, 'gamma': 0.01, 'kernel': 'rbf'}	0.9404	0.2594	0.0673

Table 4. SVR Hyperparameter Results (Sorted by Highest R2 and Lowest MSE)

3.2. Classification Results:

Class Distribution After:
SS 4392
IM 532
SS+IM 472

Fig 5. Custom Class Distribution

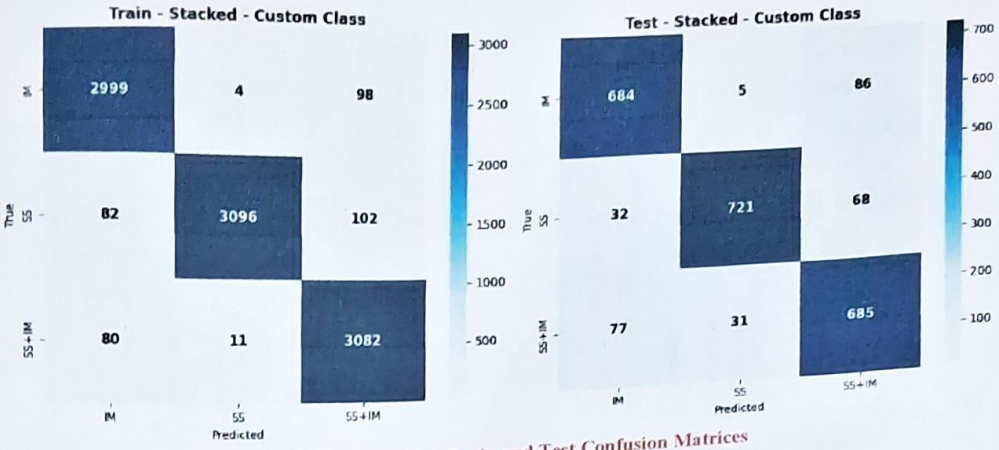


Fig 6. Custom Class Train and Test Confusion Matrices

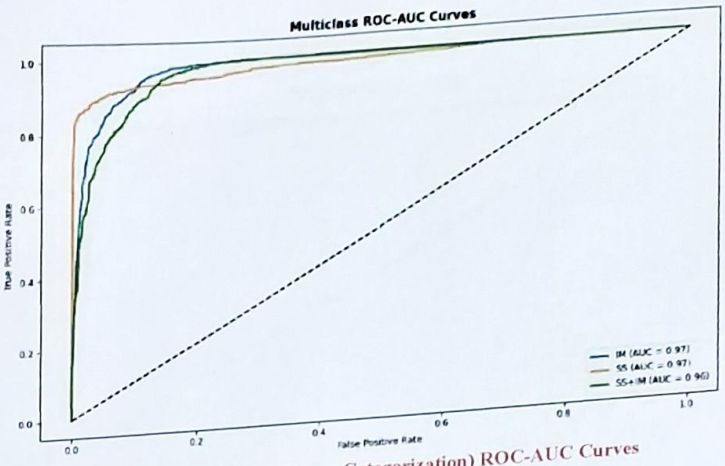


Fig 7. Custom Class (3 Phases Categorization) ROC-AUC Curves

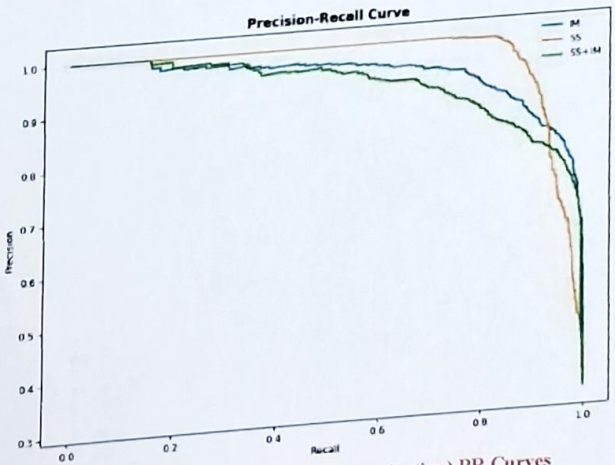


Fig 8. Custom Class (3 Phases Categorization) PR Curves

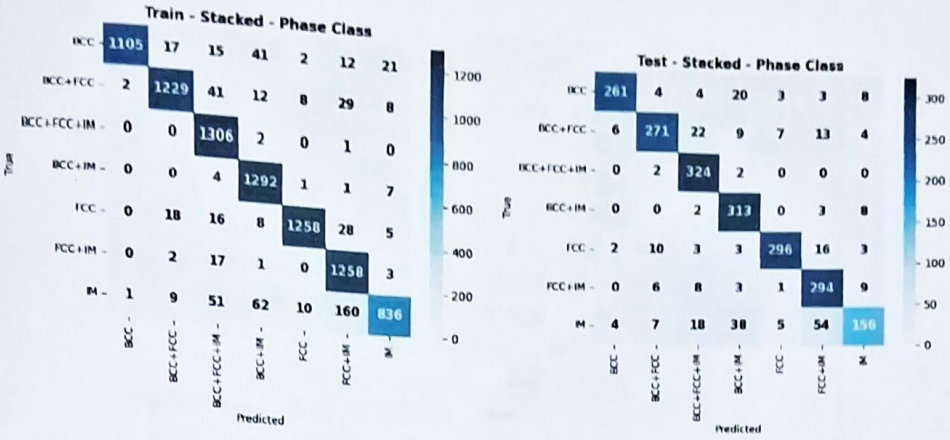


Fig 9. Phase Class (7 Phases Categorization) Train and Test Confusion Matrices

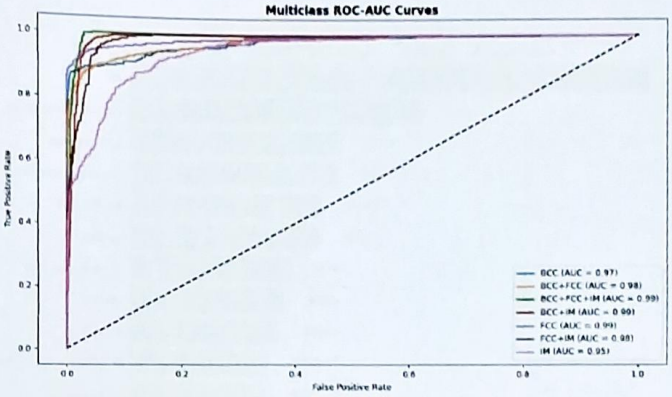


Fig 10. Phase Class (7 Phases Categorization) ROC-AUC Curves

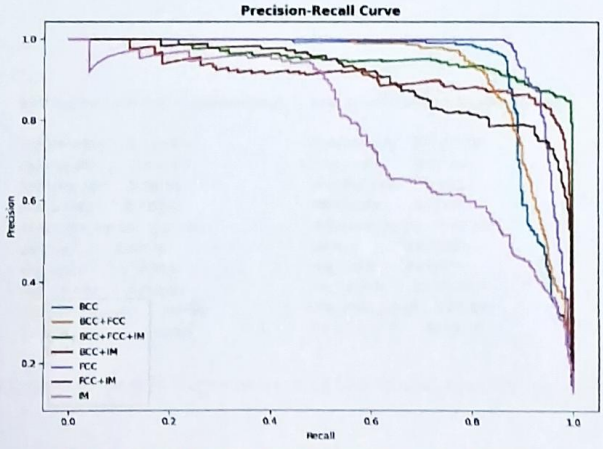


Fig 11. Phase Class (7 Phases Categorization) PR Curves

Fig 12. Custom Class Feature Importance Graph

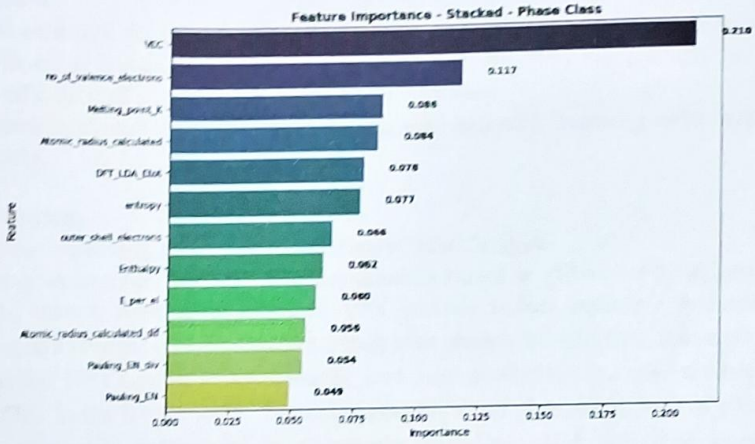


Fig 13. Phase Class Feature Importance Graph

Best Hyperparameters for Custom Class:

n_estimators	223.000000
max_depth	11.000000
learning_rate	0.281300
subsample	0.779347
colsample_bytree	0.937751
gamma	0.174443
reg_alpha	1.077931
reg_lambda	0.494468
min_child_weight	5.000000
CV_ROC_AUC	0.963055

Best Hyperparameters for Phase Class:

n_estimators	284.000000
max_depth	8.000000
learning_rate	0.145269
subsample	0.739186
colsample_bytree	0.932262
gamma	0.002906
reg_alpha	0.262775
reg_lambda	33.125132
min_child_weight	3.000000
CV_ROC_AUC	0.978377

Table 5. Most Optimal Hyperparameter for Classification-Based Prediction Results

Final Evaluation Metrics:

--- Stacked - Custom Class ---	--- Stacked - Phase Class ---
Accuracy: 0.8748	Accuracy: 0.8607
Precision: 0.8781	Precision: 0.8664
Recall: 0.8748	Recall: 0.8607
F1-Score: 0.8757	F1-Score: 0.8566
Specificity: 0.9377	Specificity: 0.9767
ROC AUC: 0.9697	ROC AUC: 0.9785

Table 6. Final Evaluation Scores (Classification Based Prediction)

4. Conclusion and Future Work

This work showcases the practical integration of ML in predicting crucial properties of solid-state hydrogen storage alloys. By targeting both continuous (VEC) and categorical (phase) properties, the models provide a dual pathway for fast screening and design of HEA compositions.

Key Outcomes:

- SVR achieved R2 of ~**99.6%** and MSE as low as **0.0045** in VEC regression.
- XGBoost achieved accuracies of ~**87.48%** and ~**86.07%** respectively for multi-class classification of 3 categorical Phases and 7 Phases.
- Feature engineering, correlation filtering, and ensemble learning were instrumental in success.

Future Extensions:

- Further improving the accuracy for Phase Classification.
- Further testing ML models for newer datasets based on HEAs for hydrogen storage.
- In the feature importance analysis, VEC, atomic radius, enthalpy, and melting point emerged as the most influential descriptors for phase classification, while no. of valence electrons, DFT calculated total energy, and outer shell electrons were the key predictors for VEC in the SVR model. These insights highlight the critical role of electronic and thermodynamic parameters in accurately modeling HEA behavior using machine learning.

5. References

[1] Ramatsoma, B. S., Makhatha, M. E., Klenam, D. E. P., & Bodunrin, M. O. (2025). **Role of compositionally complex and high entropy alloys for hydrogen storage: A bibliometric and mechanistic assessment.** *Renewable and Sustainable Energy Reviews, 222,* **115903.**

[2] Qureshi, T., Khan, M. M., & Pali, H. S. (2024). The future of hydrogen economy: Role of high entropy alloys in hydrogen storage. Journal of Alloys and Compounds, 1004, 175668.

[3] Peivaste, I., Jossou, E., & Tiamiyu, A. A. (2023). Data-driven analysis and prediction of stable phases for high-entropy alloy design. Scientific Reports, 13(1), 22556.

[4] Bhuiyan, M. M. H., & Siddique, Z. (2025). Hydrogen as an alternative fuel: A comprehensive review of challenges and opportunities in production, storage, and transportation. International Journal of Hydrogen Energy, 102, 1026-1044.

[5] Osman, A. I., Nasr, M., Eltaweil, A. S., Hosny, M., Farghali, M., Al-Fatesh, A. S., ... & Abd El-Monaem, E. M. (2024). Advances in hydrogen storage materials: harnessing innovative technology, from machine learning to computational chemistry, for energy storage solutions. International journal of hydrogen energy, 67, 1270-1294.

[6] Salam, M. Y. A., Ogunmuyiwa, E. N., Manisa, V. K., Yahya, A., & Badruddin, I. A. (2025). Effect of fabrication techniques of high entropy alloys: a review with integration of machine learning. Results in Engineering, 104441.

[7] Kong, L., Cheng, B., Wan, D., & Xue, Y. (2023). A review on BCC-structured high-entropy alloys for hydrogen storage. Frontiers in Materials, 10, 1135864.

[8] Soni, V. K., Sanyal, S., Rao, K. R., & Sinha, S. K. (2021). A review on phase prediction in high entropy alloys. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 235(22), 6268-6286.

[9] Radhika, N., Niketh, M. S., Akhil, U. V., Adediran, A. A., & Jen, T. C. (2024). High entropy alloys for hydrogen storage applications: A machine learning-based approach. Results in Engineering, 23, 102780.

6. Plagiarism Check

6% of your text matches external sources

Matches were found on the web or in academic databases