# PREDICTION OF AL ALLOYS' HARDNESS USING MACHINE LEARNING TECHNIQUES

Danny Lipson Thomas Reesar (1006892374), Wenyue Sun (1007233209) Project Report

MSE 1065 - Application of Artificial Intelligence in Materials design

**Abstract:**

**In this project, a variety of machine learning models are used to speed up the design of Aluminum alloys by improving the precision of the prediction of the alloys' hardness. The raw data is extracted from existent literature, containing chemical compositional and physical metallurgical parameters, along with aging conditions and hardness. By employing more refined machine learning and data processing techniques, XG Boost regression model reported the best $R^2$ value, pushing the value to be 6.2% higher than that from existing literature.**

**Keywords: Feature Selection, Correlation matrix, Hardness, ML functions, PCA, VIF.**

## I. INTRODUCTION:

It is difficult to overstate the importance of developing quantitative property relationship models in the scope of material science. Machine learning (ML), as a part of Artificial Intelligence, is defined as the study of computer algorithms that improve automatically through experience. Being seen as a part of artificial intelligence, machine learning algorithms are built on sample data to make predictions or decisions without being explicitly programmed to do so. The algorithms are used in a variety of applications, ranging from helping solve engineering problems to banking, online advertising and market analysis.

With a density of 2.7g/cm3 - about three times lighter than steel - Aluminum alloys, coupled with high strength, allow for design and construction of strong, lightweight structures [2]. Moreover, thanks to the inert oxide film formed on the exposed Al surface that blocks further oxidation, Al has resistance to the progressive oxidation that would affect steel. If alloyed and treated properly, Al can resist corrosion of water, salt and other environmental factors, as well as some chemical and physical agents [2]. These advantages make Al alloys particularly suitable for space vehicles, aircrafts, and all types of land and water borne vehicles.

Since the conventional approach of materials design and discovery relied heavily on chemical intuition and perfected by laborious trial-and-error based optimization cycles, a data driven route offers a much more efficient and focused approach [3]. The informatics-based tools can help greatly reduce the time and risk required to develop, produce and deploy a new material, which generally takes many years. It is also for that reason, a machine learning integrated approach is currently being pursued in the material community [3], which necessitates more relevant research and experiments.

In this report, the hardness of different Al alloys will be predicted by a machine learning approach, which will contribute to the computer aided design of Aluminum alloys with superior properties from compositional and processing perspectives. A dataset from [6] containing Al alloys' chemical composition (CC), physical metallurgical (PM) properties and age hardening (AH) parameters is selected for the project. Also, the effectiveness of the input parameters will also be analyzed with respect to predicting the hardness of Al-Alloy.

## II. LITERATURE REVIEW:

The unique combinations of properties offered by Al and its alloys makes it one of the most versatile, attractive metallic materials for a broad range of applications [2]. However, in engineering applications pure aluminum and its alloys have shortcomings such as relatively low strength and unstable mechanical properties. While the traditional approach of altering the compositions or processing parameters, then testing the properties of modified alloys by trial and error has no doubt contributed to the evolution of Al alloys, it is a rather time-consuming process. With the help of machine learning techniques, the design of Al alloys can be effectively accelerated.

The prominent characteristics of a materials informatics approach and the methods used to complete the tasks can be concluded into three processing steps. The first step is to build a dataset that is representative of the problem, after which the data would be pre-processed by removing the unintended bias inherent to the dataset [3]. After defining the problem and boundary conditions, the next task would be learning [3]. Supervised learning can be divided into regression and classification learning. Supervised learning learns a function that maps an input to an output based on example input-output pairs. Some of the common examples are KNN's and ANN's Meanwhile, unsupervised learning, whose objective is to assign a label to each data point in X without the explicit knowledge of the target variable, is used to find correlations and similarities in datasets and detects anomalous and outlier data points. Principal component analysis (PCA) is a common unsupervised learning methods [3].

ML techniques have found a growing usage in materials research in recent years. For example, applying a quantitative statistical learning model to estimate the targeted properties of new NiTi based shape memory alloys [11], and using different ML models for the predictions of physical properties and composition of steel [12]. In addition, there are a few experiments done regarding the aided design of Al alloys with a materials informatics approach.

Jiaheng Li et al [4] have investigated the Al-Zn-Mg-Cu alloy system (7xxx series) by ML based composition and process optimization. In their research ML was used to discover new 7xxx alloys with desired ultimate tensile strength (UTS). As a result, a 7xxx Al alloy with 950 MPa grade strength was developed. Research [5] has also been conducted using machine learning techniques in the prediction of Al alloys' mechanical properties. Linear regression, kNN and ANN models were selected to fit the training data, and RMSE (Root mean square error) and R-squared values were used to evaluate model performance. Reasonable predictive precision for all three models were achieved [5].

Umer Masood Chaudry et al [6] proposed a design of aluminum alloys with high hardness. The same data provided in the paper will be used in this project to improve the result by incorporating more sophisticated methods for data cleaning and model generation. For the experiment, chemical composition, physical metallurgical properties, age hardening parameters of Al- Cu-Mg-x alloy were chosen as input parameters. In the paper, feature selection was done using filter and wrapper method. The dataset has been split into training and testing set and six models were selected - LR, DL, DT, RF, GBT, and SVR. The performance of the models was analyzed using MSE and $R^2$ score [6].

Chunguang shen et al [7] used physical metallurgical parameters for the prediction of hardness or other output variables with SVM for ultra-strength stainless steel. Chemical composition, hardening parameters and physical metallurgical properties were taken as input. It was concluded that using physical metallurgical (PM) properties decreased the standard deviation and overfitting in training, as well as giving more information regarding phase formation [7]. Amir Kordijazi et al [8] reviewed the application of machine learning in design, synthesis and characterization of metal matrix composites. To design a metal matrix composite by expected properties, ANN is the most commonly used model [8], but other ML models like KNN, RF, GBT, SVR are also used.

Zheng-hua Deng et al [9] predicted the mechanical properties of Cu-Al alloy with machine learning. By setting a target in the material's mechanical properties, chemical composition was found using the best performed model among six different algorithms, out of which sequential minimal optimization

algorithm for SVR with puk kernel performed the best. By experimenting the same with the powder metallurgy, the results were close to that of the predicted values [9], bridging the gap between the computational design and experimental design of new Cu-Al alloy.

Omid khalaj et al [10] discussed the potential role of machine learning techniques in modelling the hardness of Oxidation precipitation hardened (OHP) alloys. The hardness of OHP alloys were predicted using three ML models - ANN, Adaptive Neuro Fuzzy information system (ANFIS) Model and SVM Regression (SVMR). The input features were normalized before training and mean square error (MSE), RMSE, mean absolute error (MAE) and $R^2$ value were used as performance criteria [10].

## III. PROJECT OBJECTIVES:

The main idea behind the project is to build an efficient ML model to predict the hardness of Al alloys. Usually, an inverse model is preferred to get the chemical composition required to achieve the desired hardness [8]. However, research [6][7] suggested that physical metallurgical properties are favored as input. While there are numerous papers related to using chemical composition as input, which helped obtain relevant output [4][5][9][10], there is very little research related to the physical metallurgy guided ML models in predicting the characteristics of a metallic material.

In this project, since the data of both physical metallurgical properties and chemical composition are available, there will be three sets of ML models generated to check the input feature importance with the target variable. Three Input feature datasets are a) complete dataset (PM, CC & AH), b) physical metallurgical properties and age hardening parameters together, and c) chemical composition and age hardening parameters together. From the literature review of similar work, few ML models were chosen, they are Linear regression model (LR), Random Forest (RF), Gradient Boosted Tree (GBT), XG boost regressor (XGB), Support Vector Mechanism (SVM) with different kernels and K-Nearest Neighbor Regressor (KNN).

By comparing models' performance after optimization across various datasets through performance metrics such as R-squared values, MAE and RMSE, the best performing model to predict the hardness will be identified and we will be able to understand the relevance of input features with respect to the target variable. Moreover, by using a different feature engineering technique and selecting other types of ML models, this project aims to improve the outcome of Al hardness prediction.

## IV. METHODOLOGY:

The methodology in this project adheres to common ML approaches, starting with data collection, data pre-processing, then the application of a variety of ML models, the tuning and optimization of initial models, ending with model-validation and comparison between different models, as well as the comparison with literature values [6]. Figure 1 demonstrates the flow chart of major steps in this project. More specifically, they can be divided into three steps:

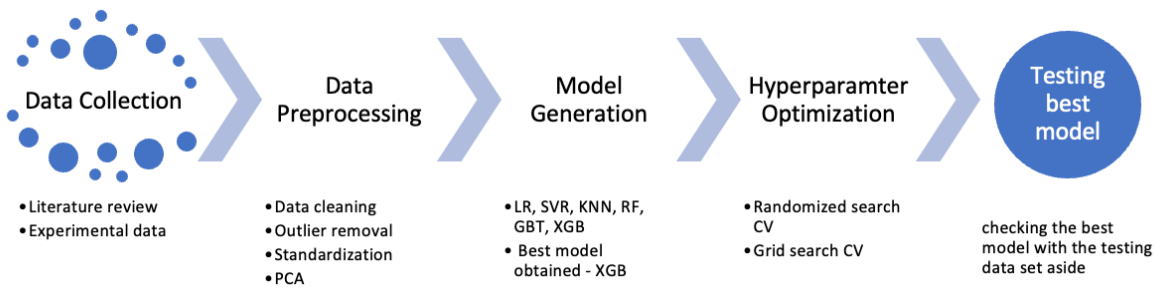


Figure 1. Flow chart for the project

Step 1: Data collection and pre-processing

This project's dataset is retrieved from literature [6] as discussed above, which contains a range of different types of Al alloys' compositions, processing conditions and hardness values. The data will be split into training and testing sets, after which it will be standardized in order to filter out the influence of differences in scales. Then feature engineering will be deployed to help reduce the number of input variables and select the most suitable characteristics for the prediction of target – hardness. In addition to the correlation values between target and input features, Variance Inflation Factor (VIF) will be used to delineate multicollinearity from the less correlated features, for that correlation matrix and scatter plots only show the bivariate relationship between variables while VIF can show the correlation of a variable with a group of other variables by quantifying the severity of multicollinearity in an ordinary least square regression analysis. Afterwards, Principal Component Analysis (PCA) will be applied to reduce dimensionality and collinearity of the data.

Step 2: Model generation and optimization

After data pre-processing, a variety of ML models are selected and applied to the training set. For this project's data, after inspecting the available models, there are in total six models selected: Linear regression, Support Vector Machines with different kernels, Random Forest, k Nearest Neighbors, Gradient Boosted Trees, and XGBoost regressor. Considering the size of the dataset and the poor results obtained by prior research [6], DL model is not selected. Below are advantages of choosing some of the models:

SVM: An algorithm that supports both linear and non-linear regression problems, and is robust to outliers, since it SVR works on the principle of reducing the error rate within a range, it should be an ideal pick for the experiment with rbf Kernal.

KNN: Not making any assumption on the data, KNN is crucial when studying data with little or no prior knowledge, which makes KNN more advantageous than models that assume linearity before training.

RF: The algorithm operates by constructing a multitude of trees running in parallel at training time and its output is either the mode of classes (classification) or the mean prediction (regression) of individual trees. As one of the most accurate learning algorithms available, RF has an effective method for estimating missing data, runs effectively on large datasets, and can handle thousands of input variables without variable deletion.

GBT: It generates a prediction model in the form of an ensemble of weak prediction models (i.e. Decision Trees), who are trained in a gradual, additive and sequential manner. GBT sometimes shows a higher accuracy than RF.

XGBoost: As an efficient and easy to use algorithm delivering high performance and accuracy, XGBoost is also known as the regularized version of GBM. With the ability to handle missing values and prune trees effectively, XGBoost is a versatile learning model that should work well with the dataset.

The rationale behind the utilization of a range of ML algorithms is that it's uncertain which exact model would be the best for the dataset selected for this specific project. In addition, for each of the machine learning models applied, after its generation, hyperparameter tuning via cross validation will be performed to help optimize the models by determining the best hyperparameter combinations.

Step 3: Selection of best models and testing

After sorting out the best models, they will be applied to the test set and have their corresponding RMSE, MAE and $R^2$ values reported. Those values will serve as metrics of the models' performance. Finally, there will be a discussion about the algorithms' efficiency in regards to their metrics scores.

## V. RESULTS AND DISCUSSIONS

In this section, the results obtained by implementing six ML models were analyzed in detail. The python build-in libraries such as numpy, pandas, sklearn, plotly, seaborn, matplotlib etc. are used, and important functions such as LR, KNN, SVR, RF, GBT, XGB are imported to understand the model performance with the three datasets.

A. Correlation Matrix:

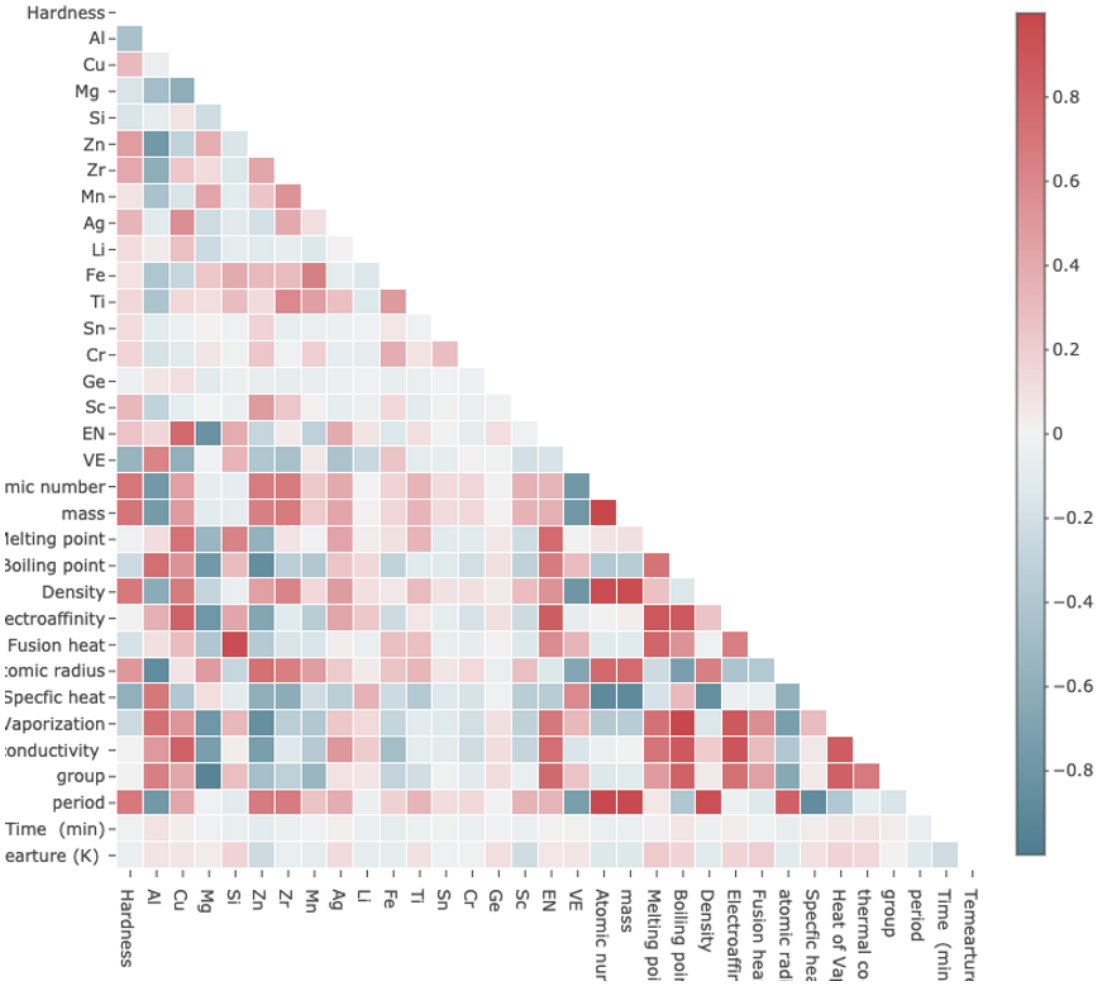


Figure 2. Correlation heatmap for the entire dataset

From the correlation plot it can be seen that in our dataset, most of the physical metallurgical parameters are heavily correlated with each other and moreover, some physical metallurgical parameters such as mass and atomic number, boiling point and heat of vaporization are completely correlated with each other, which will not add any value to the model generation and prediction. On the other hand, chemical compositional parameters are not heavily correlated with each other, but they are correlated with physical metallurgical parameters as silicon and fusion heat, Magnesium and Electronegativity etc.

It is to be noted that the addition of the physical metallurgical parameters will introduce collinearity in the dataset, when the same is checked with a statistical method known as variation inflation factor

analysis, we found out that most of the variables are dependent on each other and this causes the efficiency in prediction to drop. In order to avoid that the data is standardized with a mean of zero and a variance equal to one. Afterwards, we have used eigen vectors to completely describe the dataset (i.e. PCA – Principal component analysis). With this method only reasonable eigenvectors which can explain about 95% of variance in the data are selected and further analysis was made.

B. Dataset Analysis

By importing the data as obtained, the dataframe is cleaned entirely and the columns which didn't have any values in them are removed. After data cleaning, the dataset is converted into training and testing sets with a split of 4:1 for predicting the hardness of the material. We based our evaluation on three metrics, $R^2$ score, mean absolute error (MAE) and Root mean squared error (RMSE). The results of three different sets of models for training and testing set are discussed below.

a) Analysis on the entire dataset:

First, the results of the experiment conducted on the entire dataset are analyzed. Table 1 & 2 summarize the results for using the entire dataset with three metrics in relation to the Hardness on training and testing set. The linear models such as ordinary least squares or LR and SVR didn't perform well because of the nonlinear relationship of the features with the target variable. However, $R^2$ values of all the other non-linear models were high and those models performed well in the 10-fold cross validation (it can be seen in Figure 3).

Table 1. The metrics scores for the entire dataset before hyperparameter tuning

Output of Intial model development

| Models | MAE | RMSE | R_Squared | cv_score_val |
|---|---|---|---|---|
| LR | 16.4944 | 20.8818 | 0.1538 | 0.5307 |
| SVR | 15.3681 | 20.2614 | -0.1537 | 0.5489 |
| KNN | 4.7994 | 7.3592 | 0.9395 | 0.8801 |
| RF | 1.7849 | 3.004 | 0.9902 | 0.9429 |
| GBT | 7.3895 | 9.5871 | 0.8751 | 0.8663 |
| XGB | 7.3863 | 9.6897 | 0.8723 | 0.8641 |

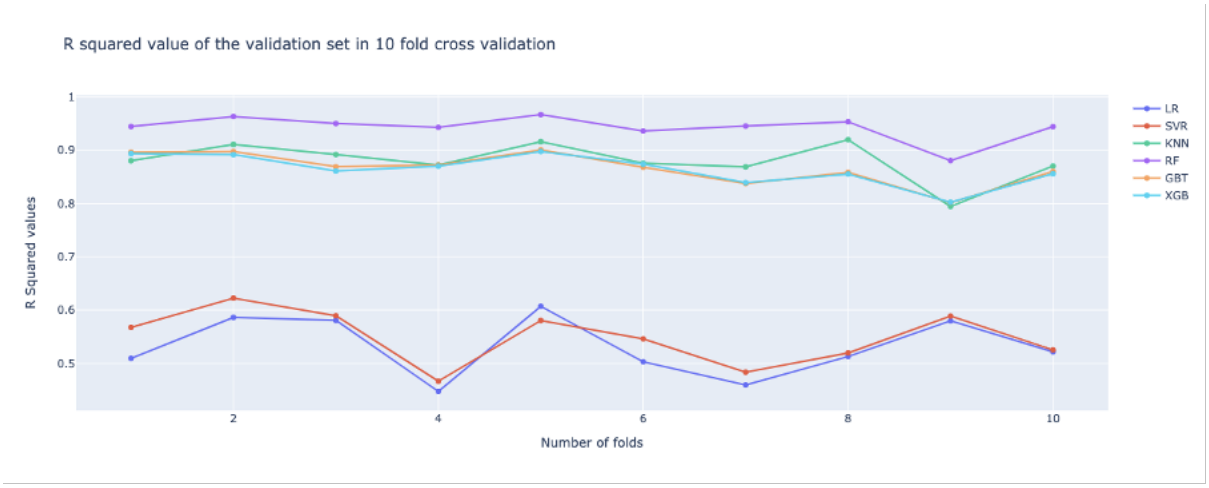


Figure 3. $R^2$ values of the models in 10-fold cross validation

Since above results are from models without hyperparameter tuning, in order to increase the efficiency of the models, hyperparameter optimization using Random search CV and Grid search CV is utilized to provide best parameters. Although RF performed better on the training set prior to hyperparameter tuning, all the other nonlinear models also performed well after hyperparameter optimization. From the result tabulated below it can be seen that since for RF, GBT and XGB the values

of MAE, RMSE and R2 values are comparable to one another, the model selection is done based on cross validation score, which can unequivocally tell us about the generalization capability of the model.

Therefore, XGB Regressor with learning rate: 0.1, 200 estimators, subsample: 0.5, minimum child weight: 3 and column sample by tree: 0.7 is used for further testing and implementation, the final result of which would be discussed in the conclusion section.

Table 2. The metrics scores for the entire dataset after hyperparameter tuning

Output of Final model development

| Models | MAE | RMSE | R_Squared | cv_score_val |
|---|---|---|---|---|
| LR | 16.4944 | 20.8818 | 0.1538 | 0.5307 |
| SVR | 15.3681 | 20.2614 | -0.1537 | 0.5489 |
| KNN | 2.3294 | 3.8357 | 0.9844 | 0.9358 |
| RF | 1.7871 | 2.9999 | 0.9903 | 0.944 |
| GBT | 1.0126 | 2.0303 | 0.9956 | 0.9388 |
| XGB | 1.2655 | 2.294 | 0.9944 | 0.9472 |

b) Analysis on CC, AH and Hardness as dataset

Similarly, analysis of Chemical composition and age hardening parameters were done excluding physical metallurgical parameters, unlike the entire dataset, dataset with only CC and AH parameters has very little collinear data, only two features, namely, Al content and Temperature, are heavily dependent upon other independent variables. Since the value of Al content is heavily dependent on the %wt. increase of other alloying elements, collinearity of Al was very high and same case is spotted for Temperature as well. In order to avoid multicollinearity in our dataset, similar to the operations done on case a), in this dataset standardization is done followed by PCA. 12 eigen vectors making up 95% of the cumulative explained variance were chosen as basis vectors for this experiment and the results of the training set before and after hyperparameter optimization is provided below in Figure 4.

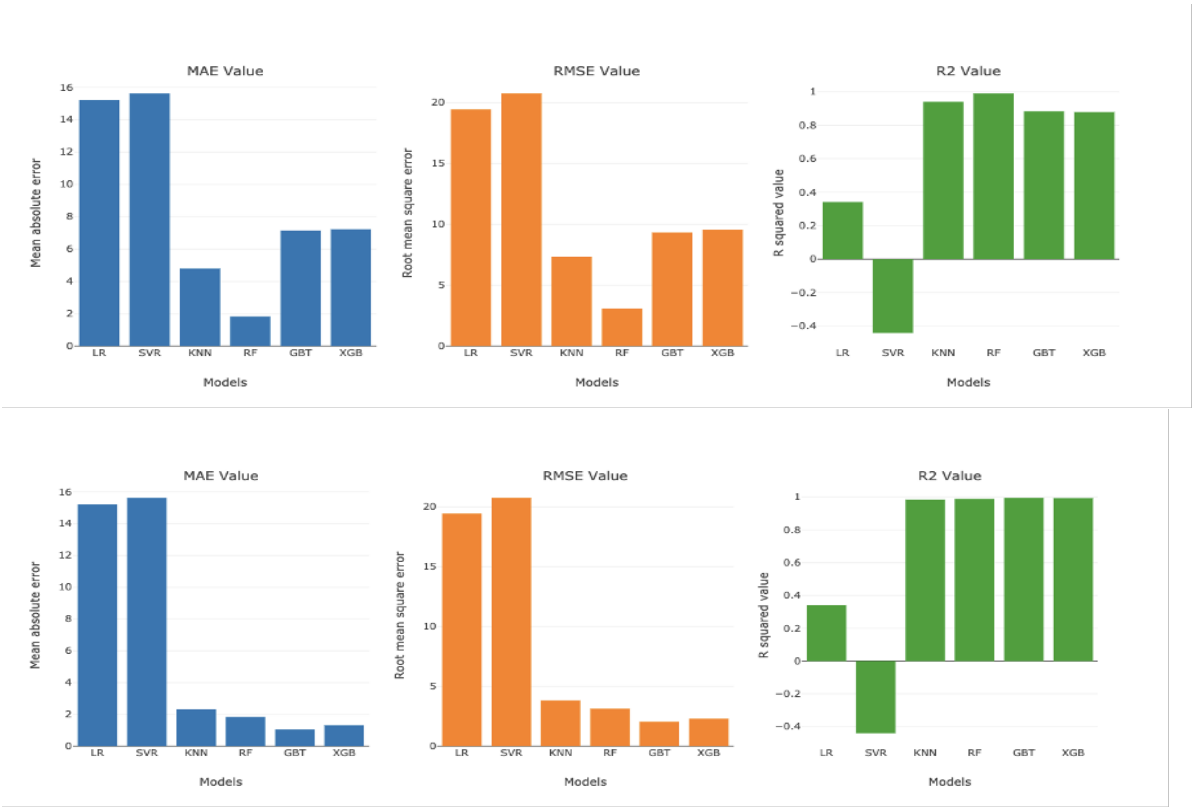


Figure 4. Evaluation metrics before (a) and after (b) Hyperparameter optimization

From the graphs it can be understood that similar to models generated using the entire dataset, RF performed better before hyperparameter optimization. However, after Hyperparameter optimization using Random search CV and Grid Search CV, even though MAE, RMSE values of GBT are slightly lower compared to XGB, $R^2$ values of RF, KNN, GBT and XGB are similar. In order to identify the better performing model, 10-fold cross validation scores are scrutinized. Chart provided below.

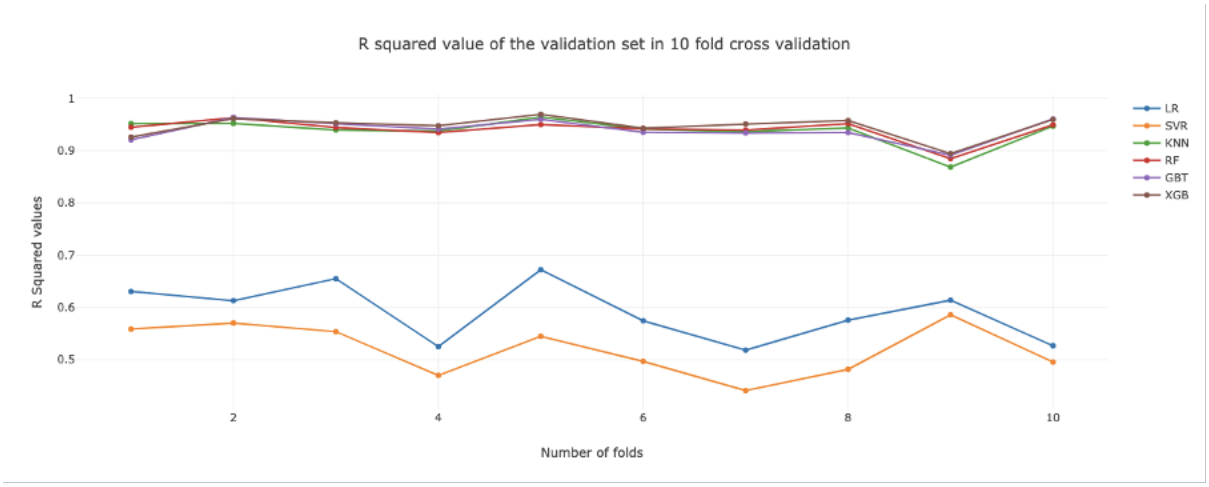


Figure 5. $R^2$ values of the models in 10-fold cross validation

Ten-fold cross validation proved that XGB is the best suited model. Out of the 10 folds R2 test on the validation set, XGB model out performed other models in seven folds out of ten, while GBT model only out performed others twice and KNN once. Since generalization of the model is important for our prediction, XGB Regressor is selected for testing.

c) Analysis on PM, AH and Hardness as input

In order to find the relevance of physical metallurgical properties with the prediction of hardness, the dataset is separated from chemical compositional parameters. Similar to the earlier case, since most of the physical metallurgical parameters in the dataset are heavily correlated with each other, it is imperative to remove the heavily correlated input features. In addition, when checking for the VIF for each input variable, even though the values didn't rise to infinity as it did when checking for complete dataset, the values for this dataset were also very high compared to dataset in case b). Then PCA was applied as well, and models are trained and hyperparameters are tuned, the results are as follows.

Table 3. The metrics scores for dataset before hyperparameter tuning

Output of Initial model development with PM and age harening parameters

| Models | MAE | RMSE | R_Squared | cv_score_val |
|---|---|---|---|---|
| LR | 16.3556 | 20.9364 | 0.1456 | 0.5332 |
| SVR | 15.0732 | 19.7654 | 0.0602 | 0.576 |
| KNN | 4.8266 | 7.3426 | 0.9396 | 0.8779 |
| RF | 1.8406 | 3.0122 | 0.9901 | 0.9422 |
| GBT | 7.7927 | 10.3059 | 0.8552 | 0.8505 |
| XGB | 7.9542 | 10.5352 | 0.8482 | 0.8475 |

Table 4. The metrics scores for dataset after hyperparameter tuning

Output of Final model development - Input parameter (PM & age hardening parameters)

| Models | MAE | RMSE | R_Squared | cv_score_val |
|---|---|---|---|---|
| LR | 16.3556 | 20.9364 | 0.1456 | 0.5332 |
| SVR | 15.0732 | 19.7654 | 0.0602 | 0.576 |
| KNN | 2.3294 | 3.8234 | 0.9845 | 0.9357 |
| RF | 1.7899 | 2.9119 | 0.9908 | 0.9475 |
| GBT | 1.2584 | 2.202 | 0.9948 | 0.9385 |
| XGB | 1.3162 | 2.3271 | 0.9942 | 0.9453 |

Since physical metallurgical parameters is a derivative of chemical composition of the alloys. In all three cases XGB was carried out as it generalized the output more than other models. Additionally, the same hyperparameter produced the best result in all three different sets of input parameters. XGB Regressor with 200 estimators, maximum depth of 10 is selected for all the cases and the same is given for the testing set as well. The R2 value obtained in all three cases was nearly close to 1 which is almost like predicting the exact value without any error. The plot of actual values versus predicted for all three cases provided below.

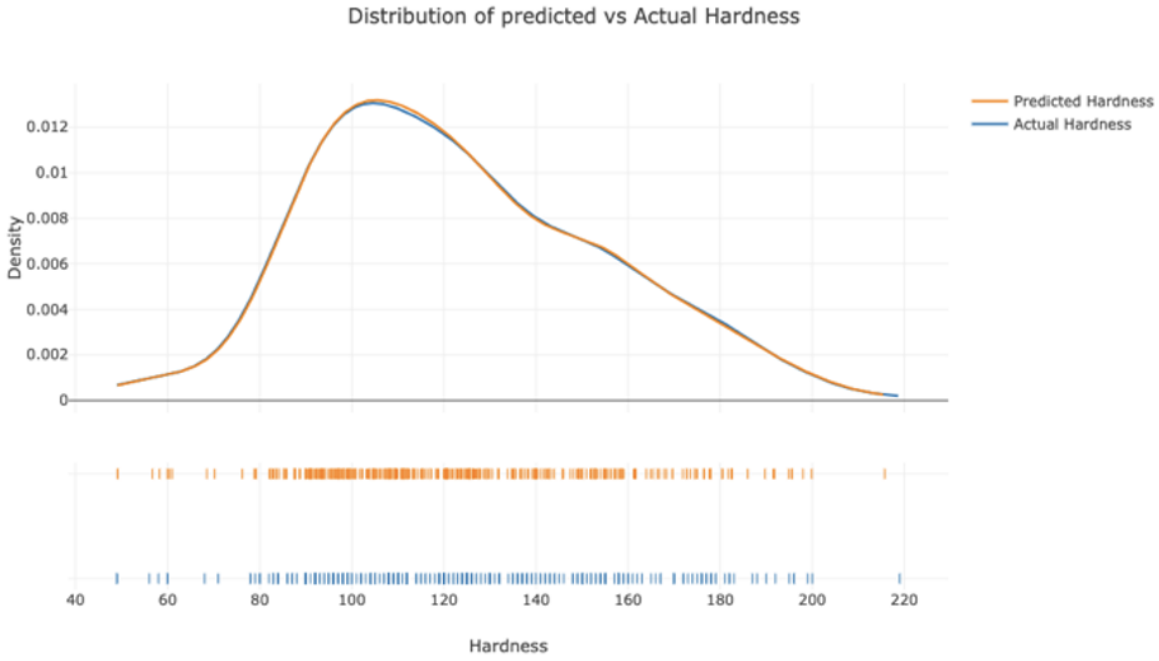


Figure 6. True value vs. predicted values

## VI. CONCLUSION AND FUTURE WORK SUGGESTIONS

In all three datasets, the evaluation metrics scores were very close. Nevertheless, as results suggested, the dataset with chemical composition and age hardening parameters performed better compared to the entire dataset as well as dataset with physical metallurgical and age hardening parameters alone. The plots indicating the metrics scores between three cases are shown in Figure 7 below, which help analyze the importance in terms of the combinations of input features.

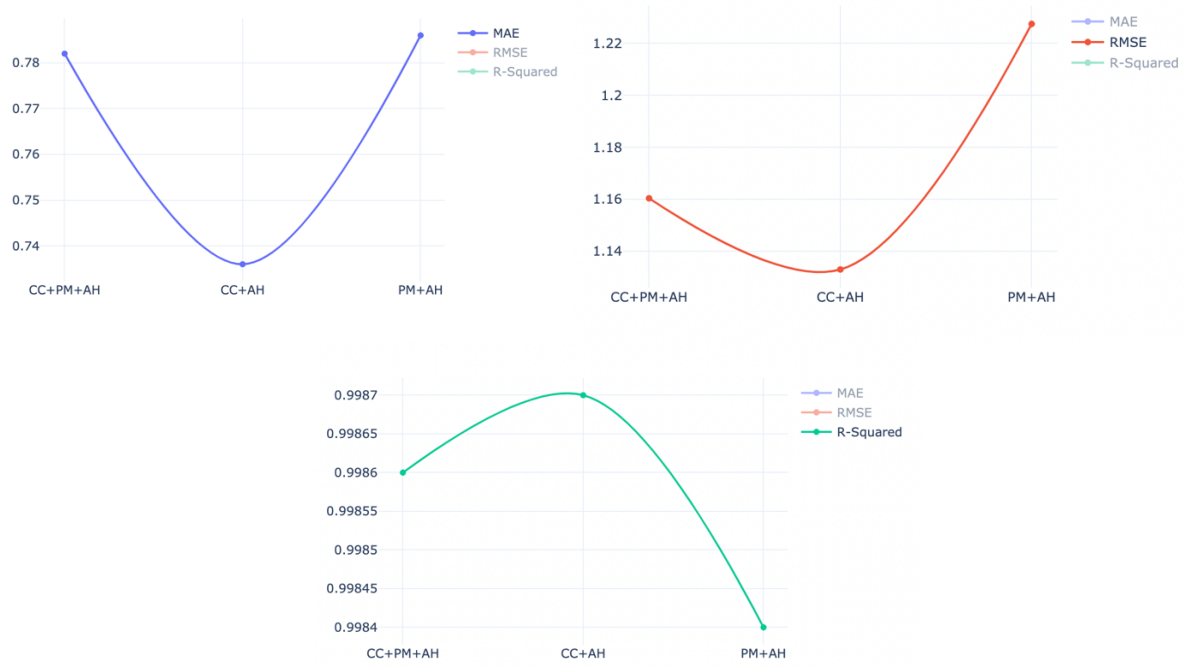


Figure 7. Metrics scores on testing sets

From Figure 7 it can be clearly observed that input parameters with chemical composition and age hardening parameters performed better than other datasets. Since physical metallurgical parameter can be seen as a derivative of the chemical composition to a certain degree - as the parameters can be set based upon the chemical elements' combination - the dataset with chemical composition and age hardening alone performed better. Despite not being able to get a direct correlation between input features and target variables because the use of PCA, with dataset (b) the $R^2$ test result was increased by 6.2% from 0.94 to 0.9987, and RMSE values were reduced by 57.98%.

With the help of the finalized XGB model with high testing precision, for new chemical combinations that may theoretically lead to Al alloys with robust hardness, the trial-and-error process necessary for current researches can be skipped especially when attempting to narrow down the compositional combinations for materials scientists and innovators to choose from. However, because the use of PCA makes it difficult to trace the eigen vectors back to initial input features, researches still need to be done pertaining to a more refined ML pipeline that avoids this shortcoming while still maintaining the extremely high predicting precision.

**REFERENCES:**

[1] Rana, R.S., Purohit, R. and Das, S., 2012. Reviews on the influences of alloying elements on the microstructure and mechanical properties of aluminum alloys and aluminum alloy composites. *International Journal of Scientific and research publications*, *2*(6), pp.1-7.

[2] Davis, J.R., 1993. *Aluminum and aluminum alloys*. ASM international.

[3] Vasudevan, R., Pilania, G. and Balachandran, P.V., 2021. Machine learning for materials design and discovery.

[4] Li, J., Zhang, Y., Cao, X., Zeng, Q., Zhuang, Y., Qian, X. and Chen, H., 2020. Accelerated discovery of high-strength aluminum alloys by machine learning. *Communications Materials*, *1*(1), pp.1-10.

[5] Devi, M.A., Prakash, C.P.S., Chinnannavar, R.P., Joshi, V.P., Palada, R.S. and Dixit, R., 2020, September. An Informatic Approach to Predict the Mechanical Properties of Aluminum Alloys using Machine Learning Techniques. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 536-541). IEEE.

[6] Chaudry, U.M., Hamad, K. and Abuhmed, T., 2021. Machine learning-aided design of aluminum alloys with high performance. *Materials Today Communications*, *26*, p.101897.

[7] Shen, C., Wang, C., Wei, X., Li, Y., van der Zwaag, S. and Xu, W., 2019. Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel. *Acta Materialia*, *179*, pp.201-214.

[8] Kordijazi, A., Zhao, T., Zhang, J., Alrfou, K. and Rohatgi, P., 2021. A Review of Application of Machine Learning in Design, Synthesis, and Characterization of Metal Matrix Composites: Current Status and Emerging Applications. *JOM*, pp.1-15.

[9] Deng, Z.H., Yin, H.Q., Jiang, X., Zhang, C., Zhang, G.F., Xu, B., Yang, G.Q., Zhang, T., Wu, M. and Qu, X.H., 2020. Machine-learning-assisted prediction of the mechanical properties of Cu-Al alloy. *International Journal of Minerals, Metallurgy and Materials*, *27*(3), pp.362-373.

[10] Khalaj, O., Ghobadi, M., Zarezadeh, A., Saebnoori, E., Jirková, H., Chocholaty, O. and Svoboda, J., 2021. Potential role of machine learning techniques for modeling the hardness of OPH steels. *Materials Today Communications*, *26*, p.101806.

[11] Xue, D., Xue, D., Yuan, R., Zhou, Y., Balachandran, P.V., Ding, X., Sun, J. and Lookman, T., 2017. An informatics approach to transformation temperatures of NiTi-based shape memory alloys. *Acta Materialia*, *125*, pp.532-541.

[12] Bélisle, E., Huang, Z., Le Digabel, S. and Gheribi, A.E., 2015. Evaluation of machine learning interpolation techniques for prediction of physical properties. *Computational Materials Science*, *98*, pp.170-177.

[13] Platt, J., 1998. Sequential minimal optimization: A fast algorithm for training support vector machines.