Deep learning the formation energy of Heusler alloys

Submitted by

Li Yifan

U0219075

Department of Mechanical Engineering

National University of Singapore

Session 2019/2020

Abstract

Many industries use deep learning algorithms to assist in the research and development of products. This project focuses on using deep learning to help predict the parameters of new materials (full heusler alloys) and to help determine whether the material is likely to exist.

This project explores the establishment of AI models through different algorithms for machine learning and deep learning. Obtain the important features in the prediction process from the machine learning part, so as to provide a reference for actual research work. The accuracy of the model is analyzed and compared to determine the final algorithm and build the model. Using this model, the formation energy and stability of ternary compounds under a certain structure are predicted, to determine whether they can form a complete Hossler alloy. The project integrates functions into a GUI, so that researchers can use this GUI to make stable predictions of any unknown new materials easily and conveniently. Throughout the project, the accuracy of the model and the GUI have been continuously improved. And show these results in the following part.

Acknowledgment

This project deserves a few acknowledgments to those who have contributed to its success. The author would also like to express his sincere gratitude to everyone including his supervisor, Dr Shen Lei who provided feedback and advice with his experience working in many projects and provided resourceful reading materials.

Contents

List of Figures

List of Tables

2. Test Results of the CGCNN model with the best validation accuracy.

3. Comparison of ML and DL results.

List of Abbreviation

ML	Machin Learning
DL	Deep Learning
CGCNN	Crystal Graph Convolutional Neural Networks
GUI	Graphical User Interface
KNN	K-Nearest Neighbors
ANN	Artificial Neural Network
CPU	Central Processing Unit
OQMD	Open Quantum Materials Database
DFT	Density Functional Theory
SQL	Structured Query Language
MAE	Mean Absolute Error
RDBM	Relational Database Management System
LGPL	GNU Lesser General Public License
GDB	GNU Debugger
WP	Word Press
CSV	Comma-Separated Values

1.Introduction

1.1 What is Deep learning?

Deep learning is part of an overall community of machine learning approaches focused on representation-learning artificial neural networks. Learning may be managed, semi-supervised, or unmonitored. (Bengio et al., 2013)(Schmidhuber. 2015)(LeCun Yann et al., 2015)

Deep learning frameworks including deep belief networks， deep neural networks, convolutional neural networks, and recurrent neural networks have already been adapted to several fields, where they have provided outcomes that are comparable to and often surpass actual specialist output. (Ciresan D, 2012)(Krizhevsky Alex, 2012)(Jon Russell, 2017)

Artificial neural networks were inspired by the processing of information and the distribution of communication nodes within biological systems. ANNs have different biological brain variations. Neural networks, in particular, appear to be rigid and conceptual whereas the physiological mind is dynamic and analogous to most living organisms. (Marblestone Adam H, 2016)(Olshausen B. A, 1996)(Bengio Yoshua, 2015)

Deep neural networks are widely used in the field of materials and have become an important driving factor for accelerating the development of new materials, processes, and technologies. Thousands of research results have been published in this field in the past five years, which have achieved good application results in crystal structure

prediction, stability analysis, and chemical synthesis research, and gradually expanded to the material's equation of state, mechanical properties, optical properties, high-temperature superconductivity, corrosion, and many other fields.

1.2 Development of Deep learning

Deep learning technology has been going through a lengthy process. The growth of neural networks was very slow from the advent of artificial neural networks in 1943 through the 1980s. The backpropagation algorithm, which perfectly solved the nonlinear classification problem, was proposed by Jeffrey Sinton until 1986 and the artificial neural network attracted people's attention once again. However, the technology re-entered a bottleneck phase because of the restricted hardware constraints at the time. Until 2006, Jeffrey Sinton and his student were providing a solution to the "disappearing gradients" problem, which allowed deep learning to flourish. Deep learning is very advanced and widespread today.

1.3 Differences between deep learning and machine learning

Machine learning's most simple approach is to use machines to process data, improve, and then make actual-world judgments and predictions of occurrences. Machine learning is about "playing" with vast quantities of data, and learning how to complete computer tasks across different algorithms.

Deep learning is indeed not necessarily an individual method of learning, but it both incorporates supervised and unsupervised methods of learning to train deep neural networks. Nonetheless, due to the rapid growth of the field in recent years, several new

learning methods have been introduced one after the other, and more and more people see it as a learning method alone.

1.4 Current deep learning applications

People usually use some translation tools in their everyday lives. The photo demonstrates Google Translate, which is widely known today. A couple of years ago Google's translated material was for comparison only. But now Google Translate is able to pre-process the sequence and allow the algorithm to learn the relationship between the text and the pointing language, which makes the translated content compatible with local accurate and true language habits.

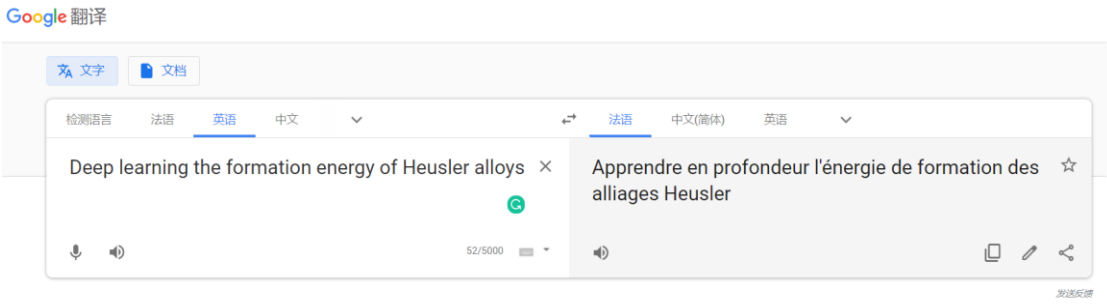


Figure 1: Demonstration of Google Translate.

Moreover, image recognition used to be a problem that has troubled people for many years because while the same kind of objects has the same characteristics, there are many discrepancies in individual information and there are space, angle, light, contrast, and other issues. But now, it is possible to create an image recognition model in 18 minutes (Pulkit Sharma, 2019).

1.5 Motivation

Heusler alloy is a special type of intermetallic composite material with over one hundred and sixty thousand members. The chemical composition is XYZ (commonly called semi-Heusler) or X2YZ (full Heusler), where X and Y are metal elements of transformation, and Z is the main element. (Daobin Lin, 2015) And several Heusler alloys can predict the properties by simply measuring the number of valence electrons. Heusler alloys besides that have other useful chemical properties. For instance, a few are Shape memory alloys while others are Superconductors. That is because the intermetallic material is a mixture of various metal components and various electronic structures. Such materials often have complicated energy band structures and exhibit many diverse physical phenomena due to the intense mixed orbital effect between limited D electrons and unrestricted s or p electrophones. They're also of great practical interest. These are used as magnetic alloys for the manufacture of CPU and hard disks, for example, and many of them are predicted to have strong thermoelectric efficiency and will be used in the future as renewable energy. Fritz Heusler is known as the material scientist who first discovered this form of compound and Heusler alloys were named after him as well. Eighty years later, after the first Heusler alloy was found in 1903, Batches of new Heusler alloys were confirmed to exist. In addition, the total number of discovered Heusler alloys in a century is only about 1500, less than one percent of the total. Over a long time, predicting new materials requires certain techniques, over example: experimental error process, computer simulation, and even expert experience-based predictions, which is very time-consuming and expensive content.

Deep learning has demonstrated many benefits today, which are difficult to equate with other approaches. It's unsupervised, for example, it doesn't have to offer it features manually so it has good portability, saves time and energy, and is really accurate. This is, without a doubt, very suitable for processing compounds of three elements with an enormous volume of data and many characteristics. The use of profound learning is, therefore, a way to help predict new Heusler alloys.

2. Literature review

2.1 Using machine learning to discover Heusler alloys

A machine-learning model was trained to discover compounds of Heusler, which are intermetallics that exhibit various physical properties that are attractive for applications in thermoelectric and spintronic material. Improving these properties includes knowledge of crystal structures, which exist in three subtle variations (Heusler, reverse Heusler, and CsCl-type structures) that are difficult to discern through diffraction techniques, and sometimes even impossible. The Heusler discovery engine operates exceptionally well in comparison with alternative approaches, allowing for fast and accurate forecasts of the incidence of Heusler or non-Heusler compounds for an unspecified mixture of elements with no structural input on more than 400000 candidates. The model has a true positive score of 0.94 (and 0.01 for false positives). The flagging of dubious entries in crystallographic databases is also useful for data sanitizing. It was applied to test candidates with the formula AB2C and predict the presence as Heusler compounds of 12 novel glides MRu2Ga and RuM2Ga (M = Ti – Co), which were

experimentally verified. One member, TiRu2Ga, exhibited peaks of the diagnostic superstructure that supported an ordered Heusler's adoption as opposed to a disordered CsCl-type structure. (Anton O. Oliynyk, 2016)

Besides this, the development of new, multi-component crystalline materials is a complex activity due to the vast space of feasible structures. Scientists demonstrate a method for dramatically accelerating the discovery of materials by using a machine learning (ML) model based on Open Quantum Materials Database (OQMD) density functional theory (DFT) results. This ML model predicts a material's stability based on its crystal structure and chemical composition, and demonstrates the method's efficacy by applying it to finding. This ML-based approach can identify new stable materials at a rate 30 times faster than undirected searches and this approach is used to predict 55 stable QH compounds previously known. Research shows that when trained using the variety of crystal structures available in the OQMD, the accuracy of this ML model is higher than it was when training on well-curated databases that contain merely one single-family of crystal structures. The benefit of using numerous training data shows how large datasets, such as OQMD, are particularly useful for the exploration of materials and that people do not need to train separate ML models to predict each particular family of crystal structures. This method works better for the small (< 103) and wide (> 105) training set sizes compared to other proposed ML approaches. The excellent versatility and precision of the method described here can be extended easily to other forms of crystals. (Kyoungdoc Kim, 2018)

2.2 Deep neural networks to correct crystal stability predictions

One of the main problems of materials science is predicting the stability of crystals. Today calculations of density functional theory (DFT) remain comparatively costly and scale poorly with the size of the device. Research indicates that deep neural networks using only two descriptors — the Pauling electronegativity and ionic radii — can predict separately the C3A2D3O12 garnets and ABO3 perovskites DFT formation energies with low mean absolute errors (MAEs) of 7 to 10 meV / atom and 20 to 34 meV / atom, well inside the DFT precision limits. Using a binary encoding scheme, further extension to mixed garnets and perovskites can be accomplished with little loss in accuracy, overcoming a critical gap in the extension of machine-learning models from fixed stoichiometry crystals to infinite multi-species crystal universes. (Weike Ye, 2018)

2.3 Convolutional neural network for predicting material properties

Using machine learning methods to speed up the design of crystalline materials typically involves manually built feature vectors or complex atom coordinate transformation to input the crystal structure, which either limits the model to certain types of crystals or makes it difficult to provide chemical insights. Scientists construct a crystal graph convolutional neural networks framework that offers a universal and interpretable representation of crystalline materials to directly learn material properties from the interaction of atoms inside the crystal. This approach provides a highly accurate prediction of measured properties of the density functional theory for eight different properties of crystals with various types of structure and compositions after training with 104 data points. Therefore, the structure is interpretable as one can derive the

contributions to global properties from local chemical environments. (Tian Xie *et al*., 2018)

3. Project

3.1 Project Objective

Based on background knowledge and literature review, deep learning is found to be useful in helping material scientists predict various parameters and properties of compounds. The project will, therefore, concentrate on using profound learning to predict the strength of formation and the stability of unknown full heusler alloys, helping to assess if they have the potential of real formation. Most complete heusler alloys need to be formed when the energy of formation is less than 1eV and the stability is greater than 0, so the likelihood of formation is very small, which enables researchers to screen out potential compounds of three elements without repeating experiments quickly. Additionally, after training, the deep learning model is very fast, has high accuracy, and doesn't require artificial features to be given, and has good portability. In addition, three conventional machine learning algorithms will be used by the project to construct a model as a comparison with the accuracy of the deep learning model and to obtain the features that account for the largest proportion in the prediction process. The project should have: (1) Comparing the accuracy of machine learning and deep learning, and obtaining important physical features (2) The function of predicting the formation energy and stability of unknown full Heusler alloys; (3) GUI interface, researchers can use this GUI to efficiently and conveniently make stable predictions for any unknown full Heusler alloy.

3.2 Software

MySQL is a framework for the management of relational databases. Relational databases store data in various tables instead of storing all data in a large warehouse which improves speed and flexibility. Developed by the Swedish company MySQL AB, it is a product of Oracle. As for WEB applications, MySQL is one of the popular computer tools for RDBMS (Relational Database Management System). MySQL software adopts a dual-licensing approach and is split into a version of the Group and a commercial version. It is widely used in small and medium-sized websites on the Internet due to its small size, fast speed, the low total cost of ownership, and particularly open source, it has become the most popular open-source database. As MySQL continues to grow, it is also being used slowly on larger websites and applications, such as Wikipedia, Twitter, and Facebook. The "M" refers to MySQL in the very famous open-source software combination LAMP. The SQL language used by MySQL is the structured language used most to access databases. (Asher Feldman, 2013)

Python is a commonly used language of description, high-level programming, general-purpose programming, developed by Guido van Rossum, the first version was released in 1991. Python is the ABC language's successor, which can also be called a LISP dialect utilizing conventional forms of infixes. (Peter Norvig, 2000) The design philosophy of Python emphasizes the readability and concise syntax of code (especially the use of space indentation to separate blocks of code, instead of using braces or keywords). Python helps developers to express ideas with less code, compared to C++ or Java.

Whether it is a small program or a big one, the language helps to make the program structure transparent. Unlike dynamic programming languages including Scheme, Ruby, Perl, Tcl, Python has a dynamic type system and garbage collection feature can handle memory use automatically and supports various programming paradigms like object-oriented, imperative, functional, and procedural programming. It has an enormous and comprehensive regular library itself. On almost all operating systems, the Python interpreter itself can operate.

Qt is a cross-platform application development system C++ which Qt Company founded in 1991. Developing GUI programs is commonly used for this. It's also called a part toolbox, in this case. These can also be used for creating applications other than GUI, such as computer software and servers. Qt uses common extensions for C + + and special code creation, as well as some macros. Other programming languages can also use Qt can language binding. Qt is free, open-source software, published under the GNU Lesser General Public License (LGPL) terms. Both versions support a wide variety of compilers, including C + + compiler from GCC and Visual Studio. In 2008, Nokia acquired Qt Company Technology, and Qt became Nokia's language programming tool. In 2012 Digia acquired Qt. The cross-platform optimized development framework Qt Creator 3.1.0 was officially launched in April 2014, achieving complete iOS support, introducing plugins such as WinRT and Beautifier, removing GDB debugging support without Python interface and incorporating Clang-based C / C++ application module, and making changes to Android support to ensure full iOS, Android, WP compatibility. (Pintscher Lydia, 2011)

4. Experimental Procedure

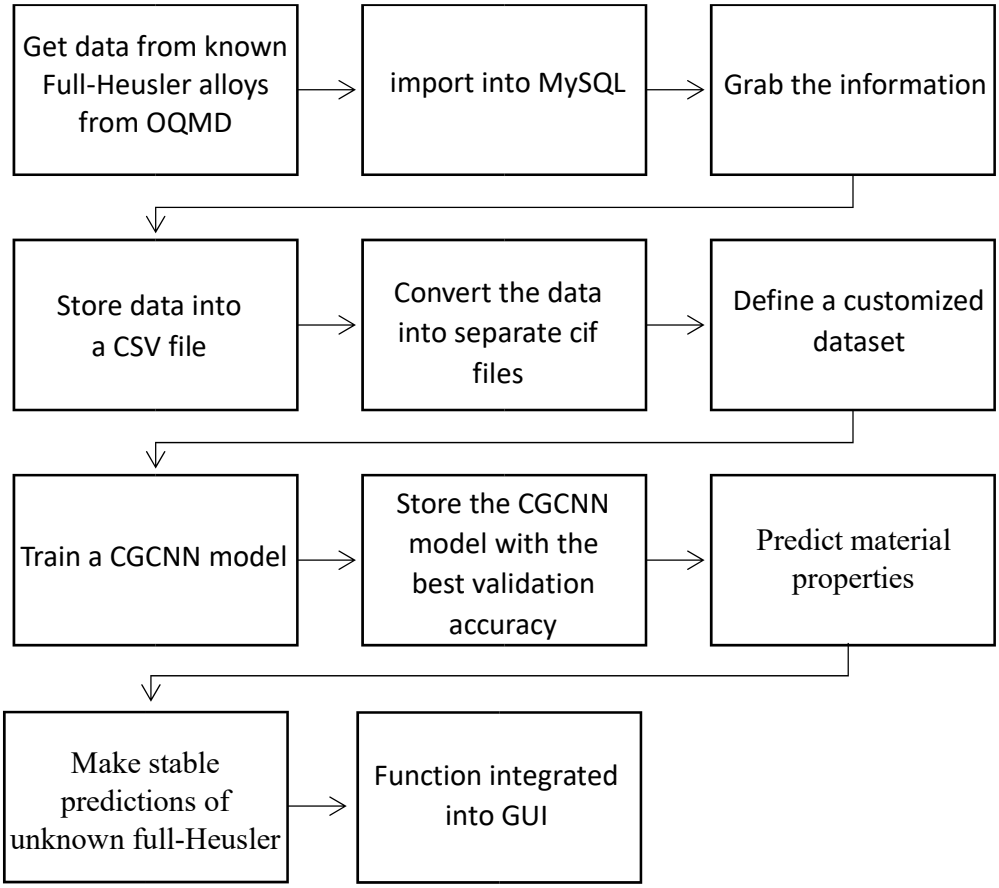


Figure 2: Project flowchart involving tasks to be completed.

4.1 Machine learning

This research uses a study of three machine learning algorithms. The first is the K-nearest neighbor, which for classification and regression tests the distance between different characteristic values. The concept is that if a certain category belongs to most of the most similar k samples in the feature space, the sample also belongs to this group, where K is

typically no greater than 20. In the KNN algorithm, the selected neighbors are all properly classified objects. This approach defines only the category to which the samples to be assigned belong in the classification judgment, based on the category of the nearest sample or samples. The second is linear regression, which is a regression analysis to model the relationship between one or more independent and dependent variables using the least square function called the linear regression equation. The third approach is gradient boosting regression. The idea is to first assign the sample weights the same value, train a poor learner, then measure their error rate and redistribute the weights according to whether the sample classification is correct and repeat this process until there are few samples wrongly divided and finally, each learner is combined with weights to form a decision tree classifier.

4.1.1 Data retrieval

The first step is to obtain the data from known Full-Heusler alloys from OQMD and import them into MySQL. And Grab the id, structure, and formation energy information of Heusler alloys by MySQLdb. Then Store these data in a data frame file. Figure 2 shows that I let the MySQL loading all material data of the OQMD database.

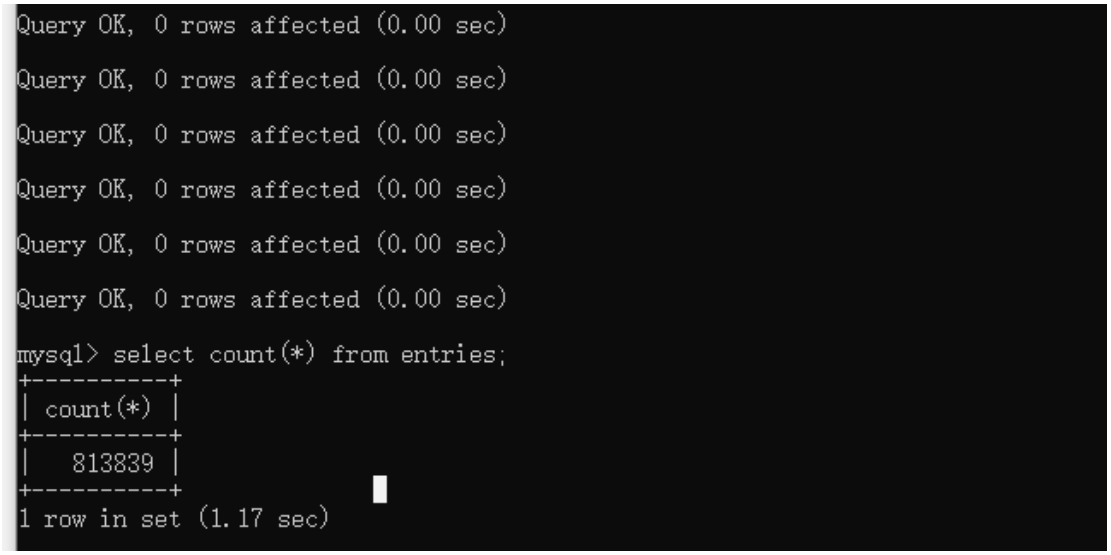


Figure 3: MySQL loading all material data of the OQMD database.

4.1.2 Feature extraction

The second step is feature extraction, first I filter the data, Because the formation energy of some materials in the database is unknown, if not filtered, a Value Error will occur when training, which causes the machine to fail to recognize the dtype. Then store the features into a data frame .and save the feature data in a CSV file to facilitate the reading of data by the machine. Figure 3 shows that the CSV files saving the feature data.

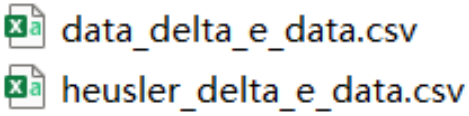
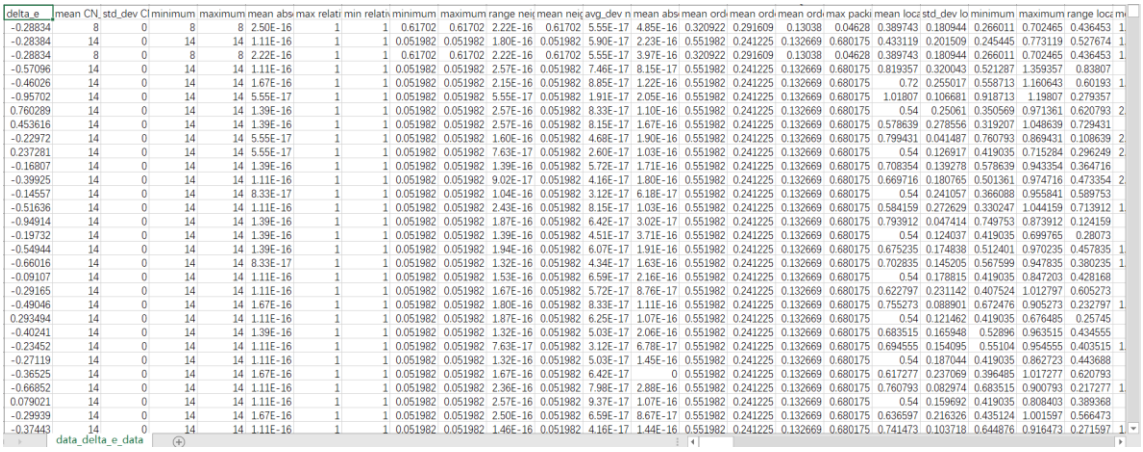


Figure 4a: The CSV files saving the feature data.



delta_e	mean CN	std_dev Cl	minimum	maximum	mean abs	max relati	min relati	minimum	maximum	range nei	mean nei	avg_dev n	mean abs	mean ord	mean ord	mean ord	max packi	mean loc	std_dev lo	minimum	maximum	range loci	mi
-0.28834	8	0	8	8	2.50E-16	1	1	0.61702	0.61702	2.22E-16	0.61702	5.55E-17	4.85E-16	0.320922	0.291609	0.13038	0.04628	0.389743	0.180944	0.266011	0.702465	0.436453	1
-0.28384	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	1.80E-16	0.051982	5.90E-17	2.23E-16	0.551982	0.241225	0.132669	0.680175	0.433119	0.201509	0.245445	0.773119	0.527674	1
-0.28834	8	0	8	8	2.22E-16	1	1	0.61702	0.61702	2.22E-16	0.61702	5.55E-17	3.97E-16	0.320922	0.291609	0.13038	0.04628	0.389743	0.180944	0.266011	0.702465	0.436453	1
-0.57096	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	2.57E-16	0.051982	7.46E-17	8.15E-17	0.551982	0.241225	0.132669	0.680175	0.819357	0.320043	0.521287	1.359357	0.83807	
-0.46026	14	0	14	14	1.67E-16	1	1	0.051982	0.051982	2.15E-16	0.051982	8.85E-17	1.22E-16	0.551982	0.241225	0.132669	0.680175	0.72	0.255017	0.558713	1.160643	0.60193	1
-0.95702	14	0	14	14	5.55E-17	1	1	0.051982	0.051982	5.55E-17	0.051982	1.91E-17	2.05E-16	0.551982	0.241225	0.132669	0.680175	1.01807	0.106681	0.918713	1.19807	0.279357	
0.760289	14	0	14	14	1.39E-16	1	1	0.051982	0.051982	2.57E-16	0.051982	8.33E-17	1.10E-16	0.551982	0.241225	0.132669	0.680175	0.54	0.25061	0.350569	0.971361	0.620793	2
0.453616	14	0	14	14	1.39E-16	1	1	0.051982	0.051982	2.57E-16	0.051982	8.15E-17	1.67E-16	0.551982	0.241225	0.132669	0.680175	0.578639	0.278556	0.319207	1.048639	0.729431	
-0.22972	14	0	14	14	5.55E-17	1	1	0.051982	0.051982	1.60E-16	0.051982	4.68E-17	1.90E-16	0.551982	0.241225	0.132669	0.680175	0.799431	0.041487	0.760793	0.869431	0.108639	2
0.237281	14	0	14	14	5.55E-17	1	1	0.051982	0.051982	7.63E-17	0.051982	2.60E-17	1.03E-16	0.551982	0.241225	0.132669	0.680175	0.54	0.126917	0.419035	0.715284	0.296249	2
-0.16807	14	0	14	14	1.39E-16	1	1	0.051982	0.051982	1.39E-16	0.051982	5.72E-17	1.71E-16	0.551982	0.241225	0.132669	0.680175	0.708354	0.139278	0.578639	0.943354	0.364716	
-0.39925	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	9.02E-17	0.051982	4.16E-17	1.80E-16	0.551982	0.241225	0.132669	0.680175	0.669716	0.180765	0.501361	0.974716	0.473354	2
-0.14557	14	0	14	14	8.33E-17	1	1	0.051982	0.051982	1.04E-16	0.051982	3.12E-17	6.18E-17	0.551982	0.241225	0.132669	0.680175	0.54	0.241057	0.366088	0.955841	0.589753	
-0.51636	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	2.43E-16	0.051982	8.15E-17	1.03E-16	0.551982	0.241225	0.132669	0.680175	0.584159	0.272629	0.330247	1.044159	0.713912	1
-0.94914	14	0	14	14	1.39E-16	1	1	0.051982	0.051982	1.87E-16	0.051982	6.42E-17	3.02E-17	0.551982	0.241225	0.132669	0.680175	0.793912	0.047414	0.749753	0.873912	0.124159	
-0.19732	14	0	14	14	1.39E-16	1	1	0.051982	0.051982	1.39E-16	0.051982	4.51E-17	3.71E-16	0.551982	0.241225	0.132669	0.680175	0.54	0.124037	0.419035	0.699765	0.28073	
-0.54944	14	0	14	14	1.39E-16	1	1	0.051982	0.051982	1.94E-16	0.051982	6.07E-17	1.91E-16	0.551982	0.241225	0.132669	0.680175	0.675235	0.174838	0.512401	0.970235	0.457835	1
-0.66016	14	0	14	14	8.33E-17	1	1	0.051982	0.051982	1.32E-16	0.051982	4.34E-17	1.63E-16	0.551982	0.241225	0.132669	0.680175	0.702835	0.145205	0.567599	0.947835	0.380235	1
-0.09107	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	1.53E-16	0.051982	6.59E-17	2.16E-16	0.551982	0.241225	0.132669	0.680175	0.54	0.178815	0.419035	0.847203	0.428168	
-0.29165	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	1.67E-16	0.051982	5.72E-17	8.76E-17	0.551982	0.241225	0.132669	0.680175	0.622797	0.231142	0.407524	1.012797	0.605273	
-0.49046	14	0	14	14	1.67E-16	1	1	0.051982	0.051982	1.80E-16	0.051982	8.33E-17	1.11E-16	0.551982	0.241225	0.132669	0.680175	0.755273	0.088901	0.672476	0.905273	0.232797	1
0.293494	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	1.87E-16	0.051982	6.25E-17	1.07E-16	0.551982	0.241225	0.132669	0.680175	0.54	0.121462	0.419035	0.676485	0.25745	
-0.40241	14	0	14	14	1.39E-16	1	1	0.051982	0.051982	1.32E-16	0.051982	5.03E-17	2.06E-16	0.551982	0.241225	0.132669	0.680175	0.683515	0.165948	0.52896	0.963515	0.434555	
-0.23452	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	7.63E-17	0.051982	3.12E-17	6.78E-17	0.551982	0.241225	0.132669	0.680175	0.694555	0.154095	0.55104	0.954555	0.403515	1
-0.27119	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	1.32E-16	0.051982	5.03E-17	1.45E-16	0.551982	0.241225	0.132669	0.680175	0.54	0.187044	0.419035	0.862723	0.443688	
-0.36525	14	0	14	14	1.67E-16	1	1	0.051982	0.051982	1.67E-16	0.051982	6.42E-17	0	0.551982	0.241225	0.132669	0.680175	0.617277	0.237069	0.396485	1.017277	0.620793	
-0.66852	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	2.36E-16	0.051982	7.98E-17	2.88E-16	0.551982	0.241225	0.132669	0.680175	0.760793	0.082974	0.683515	0.900793	0.217277	1
0.079021	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	2.57E-16	0.051982	9.37E-17	1.07E-16	0.551982	0.241225	0.132669	0.680175	0.54	0.159692	0.419035	0.808403	0.389368	
-0.29939	14	0	14	14	1.67E-16	1	1	0.051982	0.051982	2.50E-16	0.051982	6.59E-17	8.67E-17	0.551982	0.241225	0.132669	0.680175	0.636597	0.216326	0.435124	1.001597	0.566473	
-0.37442	14	0	14	14	1.11E-16	1	1	0.051982	0.051982	1.46E-16	0.051982	4.16E-17	1.44E-16	0.551982	0.241225	0.132669	0.680175	0.741473	0.103718	0.644876	0.916473	0.271597	1

data_delta_e_data

Figure 4b: The CSV files saving the feature data.

4.1.3 Train the machine learning models

The third step is building the Machine learning model. After Importing different algorithms, features and labels are assigned to x and y, respectively. Then, split the data into train and test set. (4:1) and tune some model parameters, such as estimators and max_depth, to achieve a high score.

4.1.4 Define a function to plot the relative feature importance

The fourth step is defining a function to plot the relative feature importance. Use feature importance to calculate the percentage of each feature. After simple data processing, use plt to draw it into a histogram and only take the top five features with the largest proportion.

4.2 Deep learning

This research uses crystal graph convolutional neural networks (CGCNN). We can learn material properties directly from the association of atoms in the crystal, providing a universal and interpretable representation of crystalline materials. The central idea is to describe the crystal structure by a crystal graph encoding both atomic details and bonding interactions between atoms, and then building up a convolutionary neural network on top of the graph to automatically extract representations that are optimized for predicting target properties by learning with calculated DFT data. (Weike Ye, 2018)

4.2.1 Define a customized dataset

First, a customized dataset is defined similarly to the previous part. Some Heusler alloys data is stored in a CSV file. Then the data from each Heusler alloy is converted into a separate cif file.

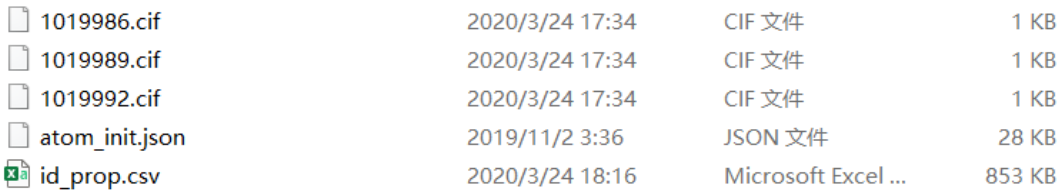


Figure 5: Demonstration of the customized dataset.

4.2.2 Train a CGCNN model

The second step is training the CGCNN model. Let the machine Get data from the customized dataset. And Train the model multiple times. Store some files, including the

CGCNN model with the best validation accuracy, the CGCNN model at the last epoch, the ID, target value, and predicted value for each crystal in the test set.

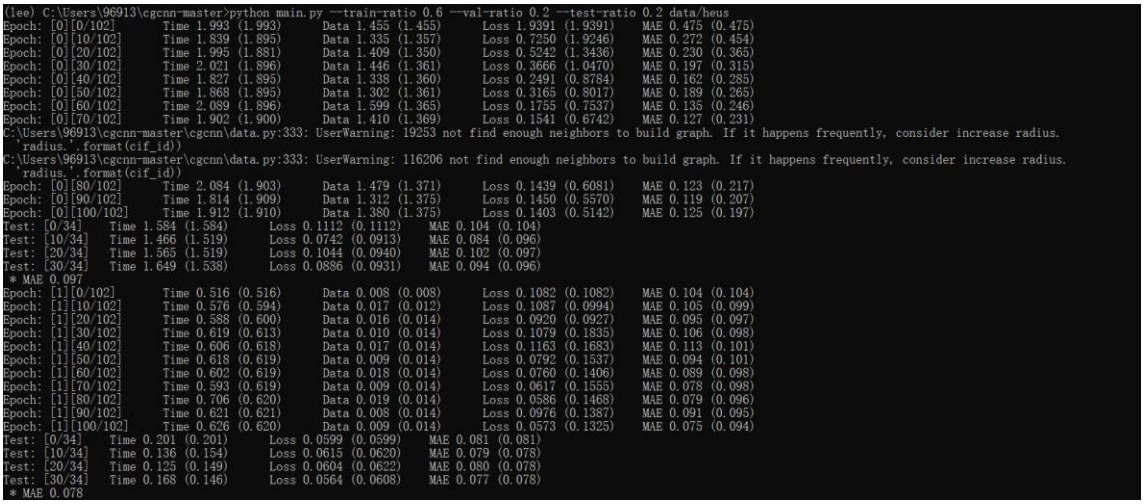


Figure 6a: Process of training the CGCNN model.

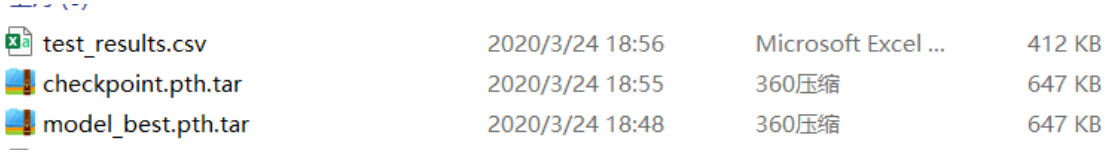


Figure 6b: Result of training the CGCNN model.

4.2.3 Predict material properties with a pre-trained CGCNN model

The third part is predicting material properties with a pre-trained CGCNN model. First, redefine a customized dataset. The cif file and CSV file at this time are the relevant parameters of the unknown full-Heusler alloys to be predicted. At this time, the true values of the formation energy and stability can be set to any value, using the highest

accuracy obtained before The CGCNN model predicts it, and the obtained test results file is the final predicted value.

4.3 Compare machine learning and deep learning results

The comparison is taken into account because it is a reliable indicator of the value of the project. At the same time, in order to compare the contingency caused by the differences in algorithms, after obtaining all the results, compare the accuracy of the three algorithms of machine learning (K nearest neighbor, linear regression, gradient boosting regression) and the deep learning model to verify whether the deep learning method is more accurate high.

4.4 GUI

Because directly using the trained CGCNN model needs to redefine a database, this requires users to have a certain basis of use. It also requires a certain degree of python language mastery to run the program, which also puts forward certain requirements for the user's language foundation. Therefore, in order to avoid this situation, so that the project can be used more widely and more simply, a GUI interface is considered to be constructed. This project uses PYQT as a GUI programming solution for the Python language. To construct a cif file of unknown full heusler alloys that only needs to be predicted, you can get their formation energy, stability, and judgment of whether they can exist with one click Usually the formation energy <1eV, stability> 0). In this project,

first use QTdesigner to design the interface layout, as shown in the figure, convert it into python code, and then encode the function of each part.

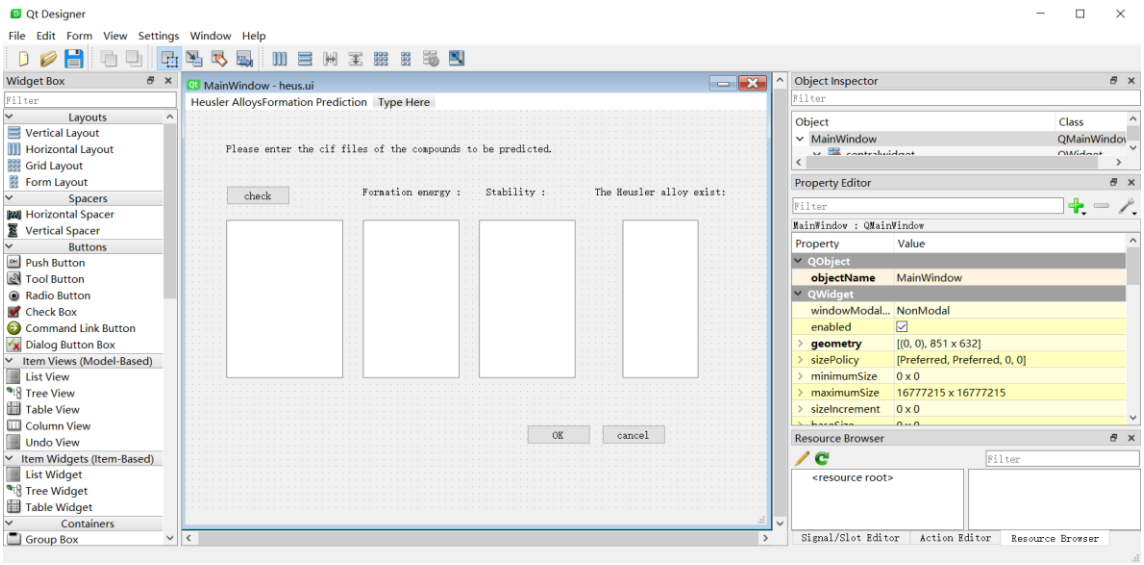


Figure 7: Demonstration of QT designer.

5. Results and discussions

This chapter will detail the final results of each part of the project (including machine learning and deep learning), and analyze and discuss it. Since the gradient boosting regression model performs well in the experiment and has high accuracy, it is mainly used to compare with the deep learning model, and the relative feature importance table is obtained from the model. The entire process is based on python, and mainly uses numpy, pandas, json and pymatgen modules to process data, and matminer to process data to get features. In the machine learning part, use sklearn to call each algorithm, and

use the matplotlib module to get the fitting chart of each part and the relative feature importance table.

5.1 Machine learning

This section is divided into three parts: K-nearest neighbor regression, linear regression, and gradient boosting regression. In each part, the MAE and fitting chart are obtained, and the relative feature importance table is obtained in the gradient boosting regression part (due to its highest accuracy).

5.1.1 K-neighbors Regression

The k-nearest neighbor regression algorithm has too many calculations (more than 40,000 samples), sample imbalance, and other problems, and it cannot give the inherent meaning of the data, so it has not been used as the focus of machine learning in the experiment algorithm so that the final accuracy is not high. The MAE of the algorithm is shown in the figure, and the fitting chart is shown. Obviously, its accuracy is far from meeting the expectations of the project.

```
>>> #calculate and print test error
>>> print("MAE:  ",np.mean(abs(y_pred-y_test)))
MAE:   0.188794068781082
```

Figure 8a: MAE of K-neighbors Regression.

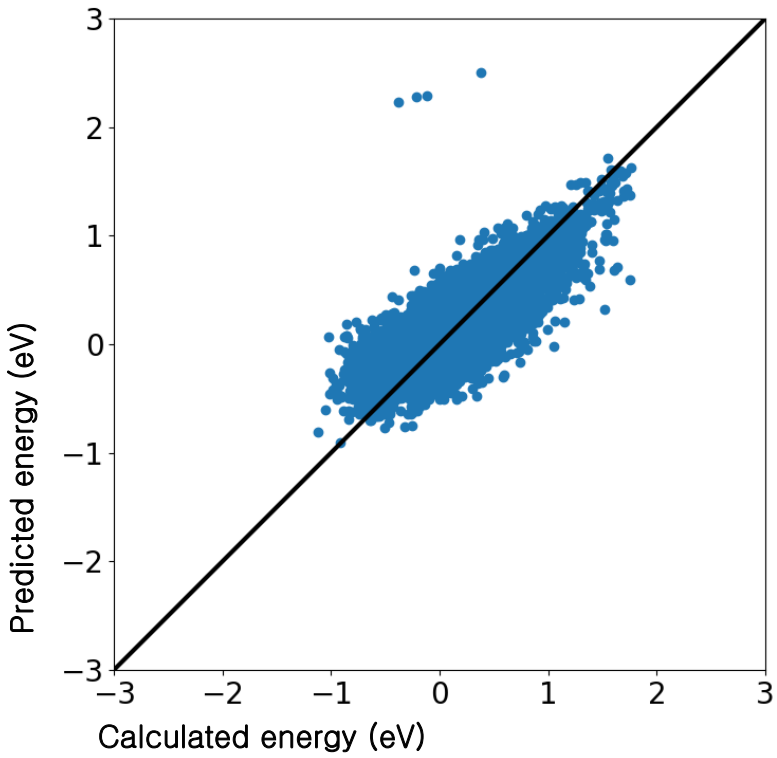


Figure 8b: Fitting chart of K-neighbors Regression.

5.1.2 Linear Regression

The linear regression algorithm found that it is not suitable for this project after the experiment (non-linear accuracy is not high), so the final accuracy is only slightly higher than K nearest neighbor regression. The MAE of the algorithm is shown in the figure,

and the fitting chart is shown. Obviously, its accuracy has not met the expectations of the project.

```
>>> print("MAE:    ",np.mean(abs(y_pred-y_test)))
MAE:    0.16986995715141445
```

Figure 9a: MAE of Linear Regression.

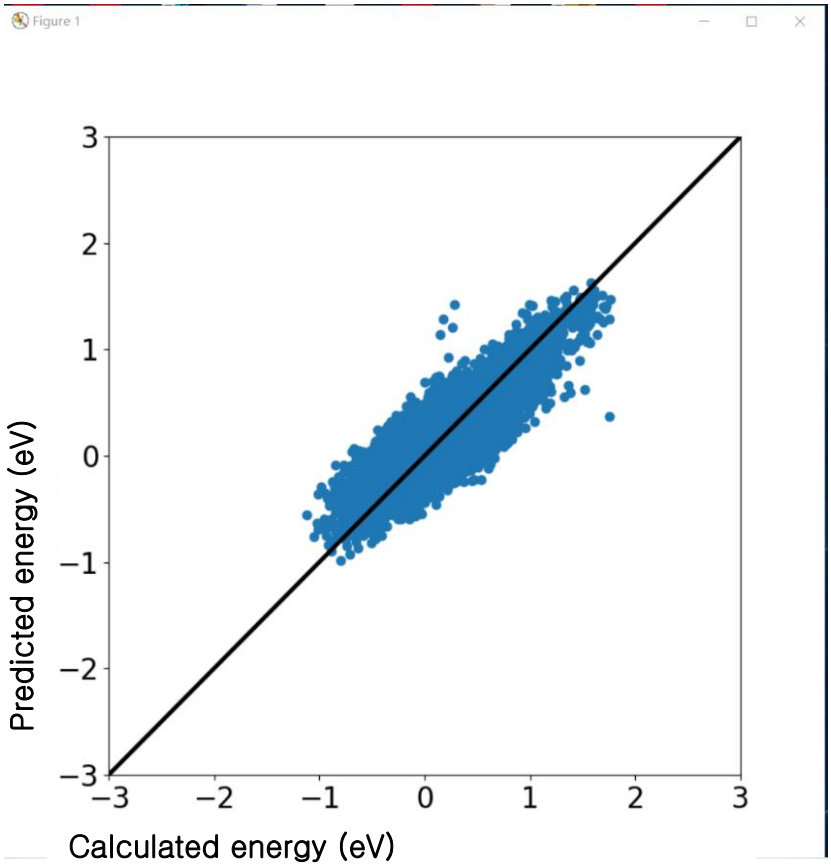


Figure 9b: Fitting chart of Linear Regression.

5.1.3 Gradient Boosting Regression

The gradient boosting regression algorithm has high accuracy and can deal with nonlinear data, and it is very robust to outliers. This overcomes the problem of missing many sample data in the experiment and is very suitable for this project. The accuracy of the model is very high. The MAE of the algorithm is shown in the figure, and the fitting chart is shown in the figure. Its accuracy is already very high, about 95.3%.

```
>>> print("MAE:  ",np.mean(abs(y_pred-y_test)))
MAE:    0.04688001203959118
```

Figure 10a: MAE of Gradient Boosting Regression.

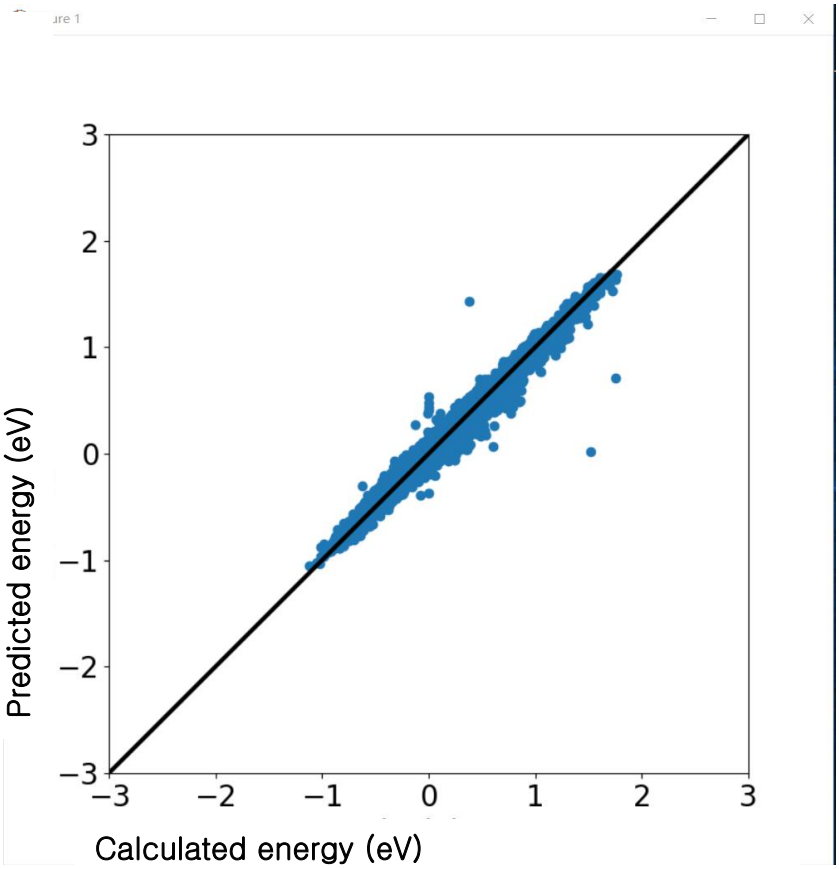


Figure 10b: Fitting chart of Gradient Boosting Regression.

5.1.4 The relative feature importance table

The features of the machine learning part have actual physical meanings, and the relative feature importance obtained can correspond to the actual physical parameters, providing a reference for the actual work of the materials scientist and a reference for its working direction. This experiment obtains the relative feature importance table according to the gradient boosting regression algorithm with higher accuracy, and outputs the top five features, as shown in the figure.

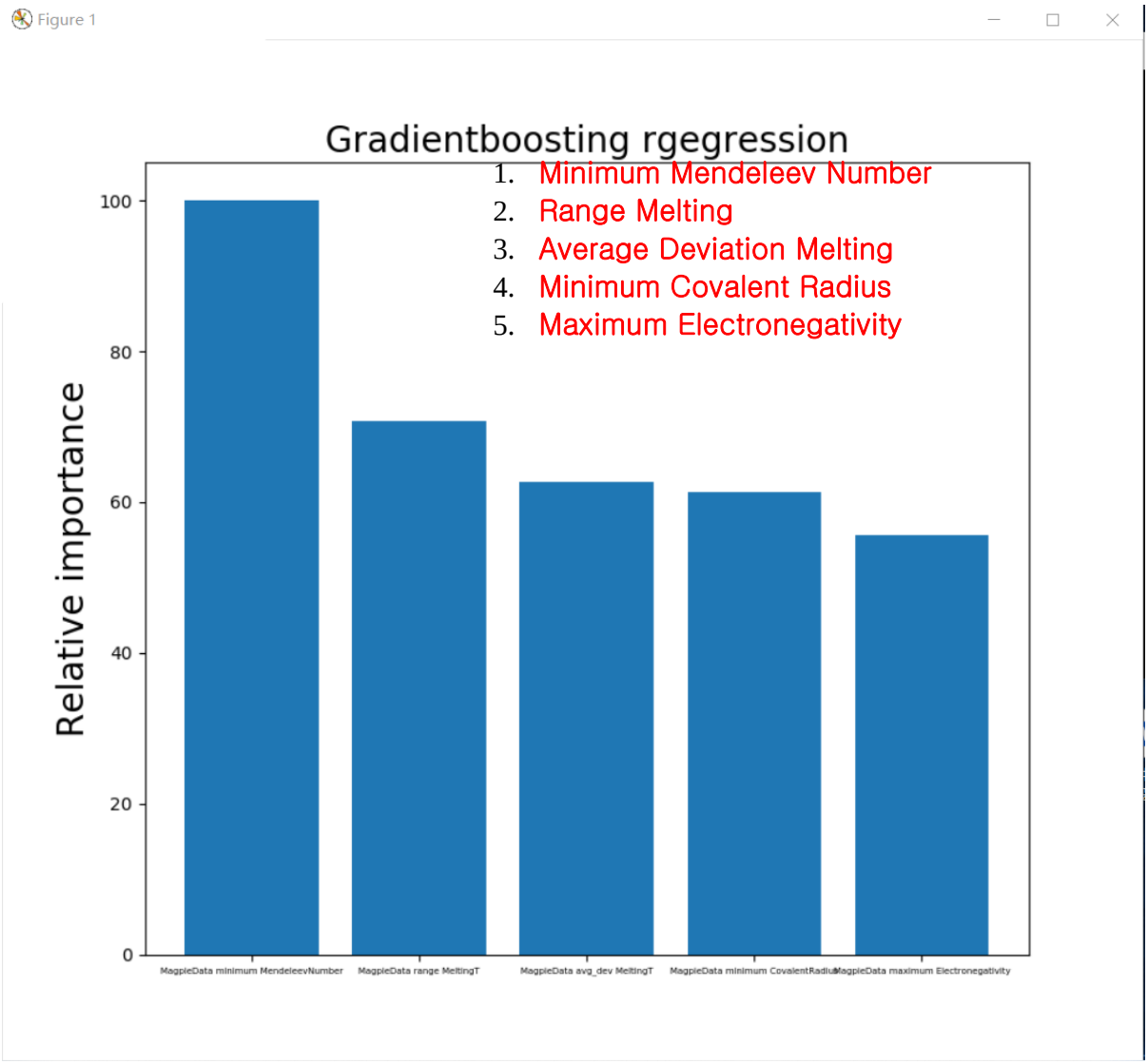


Table 1: The relative feature importance table.

The largest proportion is the Minimum Mendeleev Number. The Mendeleev number is an order number ascribed in the periodic system to each chemical element.

Second is Range Melting, which is the temperature range from the beginning of melting to complete melting. The third is Average Deviation Melting.

Then there is the Minimum Covalent Radius, the Covalent radius is half the internuclear separation of two single-bonded atoms of the same type between the nuclei.

Finally, Maximum Electronegativity, electronegativity is a measure of element atoms' ability to attract electrons in a compound. The higher the number of related electronegativity, the more electrons are drawn towards themselves by an atom or a substituent group.

5.2 Deep learning

In the deep learning part, after 30 trainings, this experiment takes the model with the highest accuracy to build a model and obtains its test results as verification. Among them, the result with the highest accuracy (the smallest MAE) is shown in the figure below.

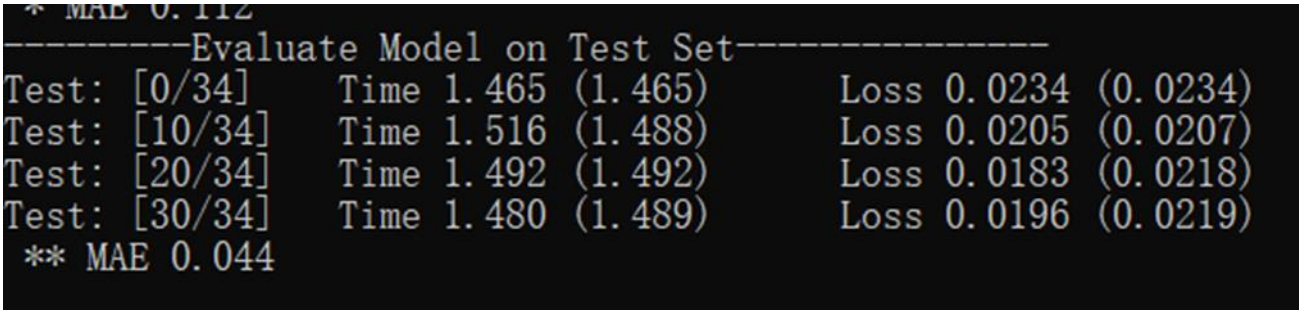


Figure 11: MAE of CGCNN model with the best validation accuracy.

Then there are some test results. From the table, it can be seen that the accuracy of the overall data is relatively high, and the error from the real data is small, which can be used as a reference for real data.

ID	TURE	TEST
484568	0.28342	0.283712
384360	0.057244	0.065705
527168	1.014224	1.06547
413411	0.360041	0.362412
539670	0.077494	0.0768
411297	-0.09186	-0.08909
478053	1.412591	1.427123
529754	0.112818	0.124845
410202	0.236581	0.242029
522418	0.049278	0.052572
521452	0.310112	0.350011
384802	-0.50797	-0.51902
554791	0.738813	0.706566
504642	0.355937	0.345865

Table 2: Test Results of CGCNN model with the best validation accuracy.

5.3 Comparison of ML and DL results

After building all the models, calculate their accuracy for comparison, and get the following table.

Machine Learning	K-Neighbors Regression	81.1%
	Linear Regression	85.0%
	Gradient Boosting Regression	95.3%
Deep Learning	CGCNN	95.6% √

Table 3: Comparison of ML and DL results.

From the table, it can be seen that the highest accuracy in machine learning is the gradient boosting algorithm, which is about 95.3%. Due to a large amount of data, overall deep learning is more accurate, whose accuracy is about 95.6%.

5.4 GUI

In this part, to realize the functions of the GUI, first use QT designer to design the layout of the GUI, use the pyuic5 module to convert it into a python file, and then define the function of each window and button. After completion, as shown below.

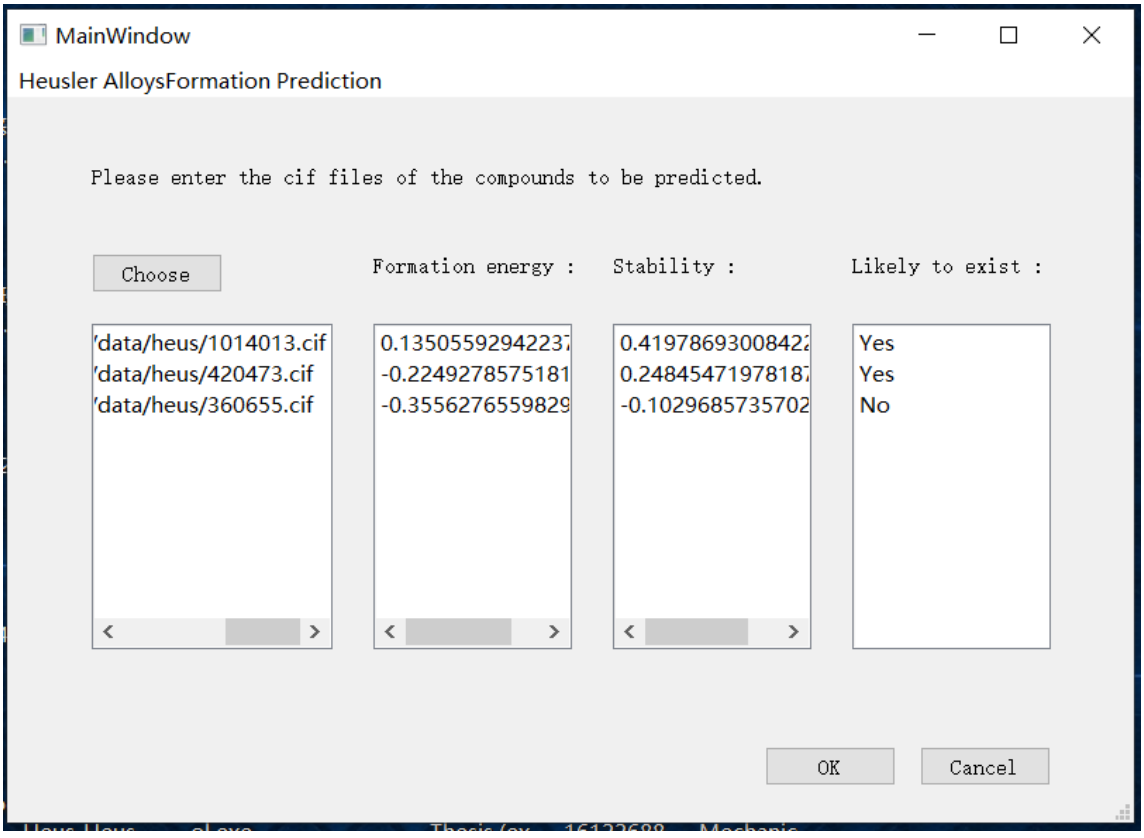


Figure 12: Demonstration of GUI.

6. Future Work

Due to time constraints, more work needs to be done in order to maximize the functionality of this project. This project demonstrates the basic usage of deep learning to predict the parameters of heusler alloys to help determine whether three-element compounds are likely to form full heusler alloys. The accuracy of the project can be further improved by fine-tune the parameters. At the same time, due to the good portability of deep learning, this model can be used to predict more material data, which needs to be further developed.

The GUI helps to transform the deep learning model into a simple and easy-to-use software tool. These features are useful for the widespread use of the project. Therefore, with the continuous improvement of project functions, the GUI can also be further optimized and upgraded to add more functions.

7. Conclusion

This project intends to establish AI models through different algorithms of machine learning and deep learning. The important features in the prediction process are obtained from the machine learning part, so as to provide a reference for the actual research work. After analyzing and comparing the accuracy of the models, determine the final algorithm and construct the model, use this model to predict the formation energy and stability of the three-element compounds under a certain structure, so as to determine whether they may form a full Heusler alloy.

So far, the project has obtained the relative feature importance table, and after comparing the accuracy of the four algorithms, the model is basically constructed using CGCNN and the function is integrated into the GUI. Researchers can use this GUI to easily and conveniently make stable predictions for any unknown full Heusler alloy.

References

Bengio, Y.; Courville, A.; Vincent, P. (2013). *Representation Learning: A Review and New Perspective*. Retrieved from

https://ieeexplore.ieee.org/document/6472238/

[Assessed on 24 Jun. 2012].

Schmidhuber, J. (2015). *Deep Learning in Neural Networks: An Overview.*

Retrieved from

https://www.sciencedirect.com/science/article/abs/pii/S0893608014002135?via%3Dihub

[Assessed on 30 Apr. 2014].

Bengio, Yoshua; LeCun, Yann; Hinton, Geoffrey (2015). *Deep Learning*. Retrieved from

https://www.nature.com/articles/nature14539

[Assessed on 27 May. 2015].

Ciresan, D.; Meier, U.; Schmidhuber, J. (2012). *Multi-column deep neural networks for image classification*. Retrieved from

https://ieeexplore.ieee.org/document/6248110

[Assessed on 13 Feb. 2012].

Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffry (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. Retrieved from

http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networ

[Assessed in 2012].

Jon Russell (2017). *Google's AlphaGo AI wins three-match series against the world's best Go player*. Retrieved from

https://techcrunch.com/2017/05/24/alphago-beats-planets-best-human-go-player-ke-jie/amp/

[Assessed on 24 May. 2017].

Marblestone, Adam H.; Wayne, Greg; Kording, Konrad P. (2016). *Toward an Integration of Deep Learning and Neuroscience*. Retrieved from

https://www.frontiersin.org/articles/10.3389/fncom.2016.00094/full

[Assessed on 14 Sept. 2016].

Olshausen, B. A. (1996). *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*. Retrieved from

ttps://www.frontiersin.org/articles/10.3389/fnco

[Assessed on 13 Jun. 1996].

Bengio, Yoshua; Lee, Dong-Hyun; Bornschein, Jorg; Mesnard, Thomas; Lin, Zhouhan (2015). *Towards Biologically Plausible Deep Learning*. Retrieved from

https://arxiv.org/abs/1502.04156

[Assessed on 13 Feb. 2015].

Pulkit Sharma (2019). *Build your First Image Classification Model in just 10 Minutes*. Retrieved from

https://www.analyticsvidhya.com/blog/2019/01/build-image-classification-model-10-minutes/

[Assessed on 10 Jan. 2019].

Daobin Lin，Baisheng Sa，Jian Zhou，Chen Si，Zhimei Sun (2015). *Ab Initio Methods of Heusler alloy M2CoA (M = Mn, Ti; A = Al, Si)*. Retrieved from

http://kns.cnki.net/kcms/detail/detail.aspx?filename=BJHK201603024&dbcode=CJFD&dbname=CJFD2016&v=

[Assessed on 8 Apr. 2015].

Anton O. Oliynyk; Erin Antono; Taylor D. Sparks; Leila Ghadbeigi; Michael W. Gaultois; Bryce Meredig; Arthur Mar (2016). *Ab Initio Methods of Heusler alloy M2CoA (M = Mn, Ti; A = Al, Si)*. Retrieved from

https://pubs.acs.org/doi/10.1021/acs.chemmater.6b02724

[Assessed on 8 Sept. 2016].

Kyoungdoc Kim; Logan Ward; Jiangang He; Amar Krishna; Ankit Agrawal; C. Wolverton (2018). *Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds*. Retrieved from

https://www.scholars.northwestern.edu/en/publications/machine-learning-accelerated-high-throughput-materials-screening-

[Assessed on 4 Dec. 2018].

Weike Ye, Chi Chen; Zhenbin Wang; Iek-Heng Chu; Shyue Ping Ong (2018). *Deep neural networks for accurate predictions of crystal stability*. Retrieved from

https://www.nature.com/articles/s41467-018-06322-x

[Assessed on 18 Sept. 2018].

Tian Xie; Jeffrey C. Grossman (2017). *Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties*. Retrieved from

https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.145301

[Assessed on 18 Oct. 2017].

Asher Feldman (2013). *Wikipedia Adopts MariaDB*. Retrieved from

https://blog.wikimedia.org/2013/04/22/wikipedia-adopts-mariadb/

[Assessed on 22 Apr. 2013].

Peter Norvig (2000). *Python for Lisp Programmers*. Retrieved from

https://norvig.com/python-lisp.html

[Assessed on May. 2000].

Pintscher, Lydia (2011). *KDE Applauds Qt's Move to Open Governance*. Retrieved from

https://dot.kde.org/2011/10/21/kde-applauds-qts-move-open-governance

[Assessed on 21 Oct. 2011].

Appendix

Project code: https://github.com/IvanLee858/DLFNOHA