Research Article

Prediction of critical temperature and new superconducting materials

Anton Matasov[1,2] · Varvara Krasavina[3]

Received: 16 April 2020 / Accepted: 27 July 2020 / Published online: 7 August 2020

Abstract
Models for predicting the critical temperature are constructed on the widest base of superconductors. Based on the selection of features, the most important physical parameters were obtained for determining the critical temperature. Models of the best quality were built on the divided data on the number of chemical elements. The resulting models are used to model new superconducting materials. The resulting materials exceed the known used superconductors at a critical temperature.

Keywords Superconductivity · Superconductor · Machine learning · Critical temperature

1 Introduction

At present, there is no complete theory of superconductivity. What follows is the impossibility of predicting the critical temperature of most known superconducting materials. The critical temperature is the most important parameter that determines the superconducting state, the economic feasibility of using superconducting materials. Also, due to the lack of a general theory, the search for new superconducting materials with a higher critical temperature is mostly intuitive.

The generally accepted theory of superconductivity is the Bardeen–Cooper–Schrieffer (BCS) theory [1, 2]. The critical temperature in this theory in the weak coupling limit depends on the Debye temperature, the electron–phonon interaction potential, and the density of electronic states at the Fermi level. These parameters cannot always be accurately measured, and it is also shown that the electron–phonon interaction is insufficient for the appearance of superconducting properties in high-temperature superconductors [3, 4]. Another general theoretical approach to determining the critical temperature is the effect of zero-point oscillations on the formation of superconducting particles [5–7]. But this theory works only for ordinary metals and at the moment does not allow predicting the critical temperature of more complex materials.

Another alternative approach in predicting the critical temperature may be the use of machine learning methods. There are several works in this direction, where statistical methods are considered and models are constructed to determine the critical temperature of high quality [8–12]. Thus, in the present work, machine learning methods will be used to solve the problem of determining the critical temperature and searching for new superconducting materials.

The work [12] uses the most comprehensive data on superconducting materials that contain all classes of superconductors. The disadvantage of [12] is that when creating the database, the influence on the quality of the model of the presence in the data of materials with the same name, input features, but different critical temperatures was not taken into account. This, apparently, is due to the fact that when creating the database, the authors did not take into account the different oxygen content in cuprate and other superconductors, which negatively affects the predictive and interpretative capabilities of the

✉ Anton Matasov, matasov_av93@mail.ru; Varvara Krasavina, varvara.krasavina@gmail.com | [1]Department of Physics and Technologies of Electrical Materials and Components, National Research University "MPEI", Moscow, Russia. [2]MIREA - Russian Technological University (RTU MIREA), Moscow, Russia. [3]Epoch8, Moscow, Russia.

model. This problem was solved in [13], where a complete preprocessing and visualization of the data was carried out; the obtained data will be used in this work.

Data represent three datasets with unique names of materials with minimum (sc_min), mean (sc_mean), maximum (sc_max) experimental values. The splitting of the data into three datasets is due to the fact that the critical temperature significantly depends in particular on the size of the sample and the number of defects; therefore, several values of the critical temperature are possible for the same material.

Each dataset contains 15,542 unique materials with values of 83 features. Features were formed based on the values of the parameters of the atoms (Table 1) that make up the compound (which are parameters of simple substances that are measured under normal conditions.), their weights, and expressions for calculating the initial features (Table 2). Also added to the data is the feature "the number of chemical elements" and the critical temperature.

In this paper, for the first time, taking into account the problem of work [12]: we developed models for determining the critical temperature for all classes of superconductors based on the most comprehensive database of superconducting materials, identified classes of materials based on dividing data by the number of chemical elements, developed and applied feature selection algorithms, created databases for modeling new possible superconducting materials, modeling of new possible superconducting materials with determination of their critical temperature.

Based on the objectives, the structure of this work is: a description of the basic model of a random forest, the results of dividing the data by the number of chemical elements, the algorithm and the results of the selection of attributes, the results of predicting the critical temperature of new possible superconducting compounds, discussion and conclusions.

Table 1 Initial features of atoms and their units

Feature	Units
Atomic mass	Atomic mass unit (amu)
Ionization energy	Kilojoule per mole (kJ/mol)
Atomic radius	Picometer (pm)
Density	Kilogram per cubic meter (kg m^3)
Electron affinity energy	Kilojoule per mole (kJ/mol)
Heat of fusion	Kilojoule per mole (kJ/mol)
Thermal conductivity	Watt per Meter-Kelvin (W/(m * K))
Valence	–

2 Critical temperature prediction model

2.1 Random forest regression

It is proposed to consider the random forest regression [14] as a model for predicting the critical temperature. For comparison with other studies, the coefficient of determination or r-square is used as the main metric for the quality of the model. It is important to improve the model quality criterion in order to increase the applicability of

Table 2 An example of calculation formulas for the formation of characteristic values of features for the number of chemical elements equal to 2

(t_1, t_2—feature value for the atoms in the compound;
p_1, p_2—ratios of the atoms in the compound.)

$w_1 = \frac{t_1}{t_1+t_2}$	$A = \frac{p_1 w_1}{p_1 w_1 + p_2 w_2}$
$w_2 = \frac{t_1}{t_1+t_2}$	$B = \frac{p_2 w_2}{p_1 w_1 + p_2 w_2}$

Feature	Expression
Mean	$\mu = \frac{t_1+t_2}{2}$
Weighted mean	$\nu = p_1 t_1 + p_2 t_2$
Geometric mean	$\sqrt{t_1 t_2}$
Weighted geometric mean	$t_1^{p_1} * t_2^{p_2}$
Entropy	$-w_1 \ln(w_1) - w_2 \ln(w_2)$
Entropy weighted	$-A\ln(A) - B\ln(B)$
Range	$t_1 - t_2 (t_1 > t_2)$
Weighted range	$p_1 t_1 - p_2 t_2$
Standard deviation	$\left(0.5(t_1 - \mu)^2 + (t_2 - \mu)^2\right)^{0.5}$
Weighted standard deviation	$\left(p_1 (t_1 - \nu)^2 + p_2 (t_2 - \nu)^2\right)^{0.5}$

the model to all superconducting materials and reduce the difference between the predicted critical temperature and the experimentally measured one.

The algorithm for constructing a model and assessing its quality is standard:

(1) Divide the data into attributes X and the target variable y. The critical temperature of superconducting materials is y. As X - data from which the features are removed: critical temperature, material name.
(2) Divide our data into training and test samples, where the test sample is 0.3 of the total and the training sample is 0.7, respectively.
(3) Train the model on the training set.
(4) Estimate the coefficient of determination using a trained model on test data.

Hyper parameters of the model were selected manually to improve the quality of the model. As a result, the following parameters were obtained: max_depth = 22, n_estimators = 119, min_impurity_decrease = 0.0007.

Outliers may still be present in the data. In order to get rid of them, we will remove all materials for which the values of the features are not included in the range of values of 3 standard deviations from the mean value of any feature. The obtained values of quality metrics for three datasets are presented in Table 3.

To further improve the quality of the models, the feature selection algorithm was used. Characteristic selection methods are required when there can be too many features, more than what is really needed for the model. The data may contain noise signs that spoil the quality of the model.

The feature selection algorithm used in the work is as follows:

(1) Take the first feature of data X, train the model and calculate the coefficient of determination r^2.
(2) Add the following feature, calculate r^2. If the correlation coefficient is greater, then as a result add a feature, if not, then delete.
(3) Begin to delete the features and calculate the correlation coefficient after deletion. If the correlation coefficient becomes larger, then we delete the attribute; if not, then add it.
(4) Add features that are not in the current data. Calculate the correlation coefficient, if it becomes larger, as a result add a feature, if not, then delete it.
(5) If the quality metric does not change, then complete the process of removing and adding features. If not, return to step 3.

The results of applying the feature selection by the feature selection algorithm for each dataset before removing outliers are presented in Table 4.

Also, for feature selection, we used the method "feature_importances_" from RandomForestRegressor in Python. This method allows to determine which features when training the model most strongly affect the quality of the model. When a feature is used to split a node, the quality of the model is calculated and compared with the quality of the original node without breaking. Changes in the quality value are summarized for each characteristic and normalized at the end of the calculation. The higher the value of the obtained coefficient for a characteristic, the more important it is for constructing a random forest.

For the three models built on sc_min, sc_mean, sc_max data, we select the 10 most important features (Fig. 1).

2.2 Dividing data by the number of chemical elements

The resulting random forest models provide good-quality metrics. In order to further increase the quality of the models, it is proposed to break the data by the number of chemical elements. It can be more convenient application of models and the possible division of superconductors into classes.

The number of chemical elements in datasets varies from 1 to 9. A complete enumeration of possible options (for example, materials with the number of elements 1,2,3

Table 3 Quality metrics for test data for each dataset before and after outlier removal

Model quality parameter	sc_min	sc_mean	sc_max
Before removing outliers			
r^2 score	0.9101	0.9190	0.9150
Mean absolute error	5.66	5.32	5.48
Mean squared error	98.19	89.98	96.66
After removing outliers			
r^2 score	0.8880	0.9134	0.8953
Mean absolute error	6.87	6.20	6.74
Mean squared error	129.57	101.61	124.36

Table 4 Quality metrics for test data for each dataset after selection of features

Model quality parameter	sc_min	sc_mean	sc_max
r^2 score	0.9152	0.9239	0.9200
Mean absolute error	5.63	5.33	5.43
Mean squared error	92.69	84.55	91.08
Number of features after selection	39	38	38

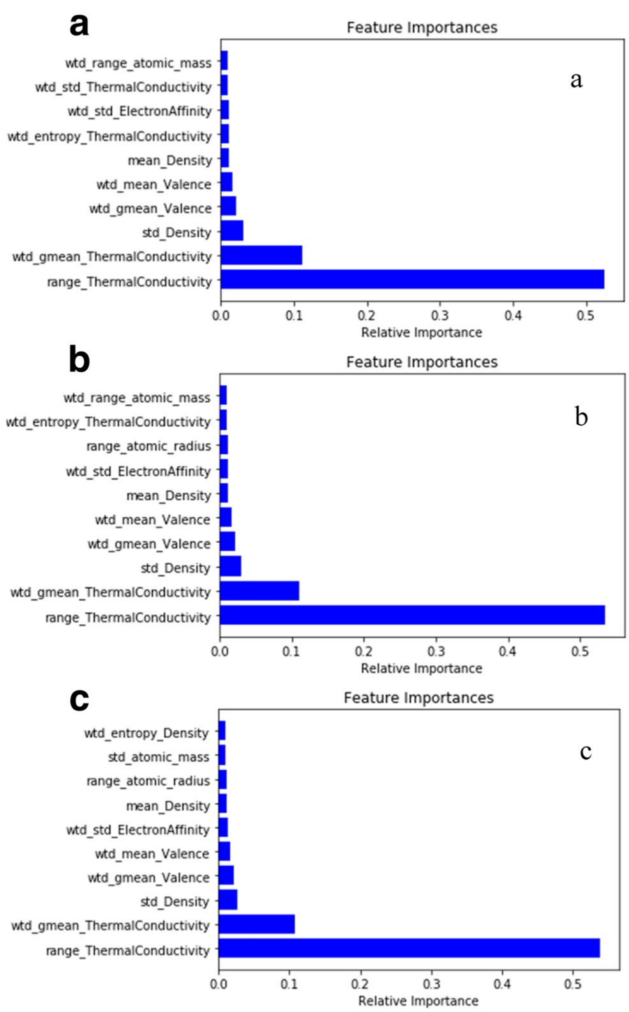

Fig. 1 The ten most important features for random forest models based on sc_min (**a**), sc_mean (**b**), sc_max (**c**) data

or 2,9,4, etc.) is ineffective. To determine the best material ratios by the number of chemical elements, the following algorithm is proposed:

(1) Select all materials for which the number of chemical elements is equal to one.
(2) Add materials with a different number of elements to this data (Example: 1,2; 1,3; 1,4…1,9) and calculate the coefficient of determination for each received dataset. Select a dataset with the best coefficient of determination.
(3) Again add materials to the selected dataset with a different number of elements (Example: 1,2,3; 1,2,4…1,2,9) and calculate the coefficient of determination for each dataset.
(4) When the coefficient of determination ceases to increase when adding new data, stop the process and record the resulting ratio of materials for various amounts of chemical elements.
(5) Start the process anew for materials with two or more elements.

The algorithm was applied to sc_min, sc_mean, sc_max data, and the results are presented in Table 5.

Next, we apply the feature selection algorithm for the data in Table 5 and chose the best ratio of materials by the number of chemical elements for a certain value of a given attribute (Table 6).

3 Prediction of the critical temperature of new superconducting materials

3.1 Databases for modeling superconductive materials

In order to apply the model to new superconducting materials, it is first necessary to model their attributes. For this purpose, the corresponding algorithms described below were developed and databases were formed. The purpose of the algorithms developed here is to obtain material with the same characteristics as in the source data.

Table 5 The best ratios of materials by the number of chemical elements for sc_min, sc_mean, sc_max data

sc_min		sc_mean		sc_max	
Number of chemical elements	r^2 score	Number of chemical elements	r^2 score	Number of chemical elements	r^2 score
1,8,2	0.9492	1,8	0.9328	1,8	0.9061
2,7,9	0.9531	2,8,9	0.9430	2,6,9,1	0.9384
3,7,2	0.9416	3,7,2	0.9439	3,7,2	0.9418
4,5,2,9,3	0.9164	4,7,8,2,1	0.9238	4,7,8,2	0.9251
5,2,3	0.9227	5,2,3	0.9320	5,2,3	0.9296
6,2,1,9	0.9274	6,2,1,9	0.9376	6,2,9,1	0.9384
7,3,2	0.9416	7,3,2	0.9439	7,3,2	0.9418
8,1,2	0.9492	8,2,9	0.9430	8,2,9,6	0.9328
9,1,8,2,7	0.9440	9,2,7	0.9456	9,4,2,5,3	0.9225

Table 6 The best ratio of materials by the number of chemical elements for the data after the selection of features

Number of chemical elements	r^2 score	Data
1,8,2	0.9587	sc_min
2,7,9	0.9628	sc_min
3,7,2	0.9576	sc_mean
4,7,8,2	0.9350	sc_max
5,2,3	0.9371	sc_mean
6,2,1,9	0.9439	sc_max
7,3,2	0.9576	sc_mean
8,2,9	0.9595	sc_mean
9,2,7	0.9537	sc_mean

Note once again that the initial features are formed on the basis of 8 physical parameters from Table 1. For each feature, based on the composition of the material, features were formed: the mean, weighted mean, geometric mean, geometric mean weighted, entropy, entropy weighted, range, weighted range, standard deviation and weighted standard deviation.

The function *priznak*, which allows to get 10 features for each of the 8 physical parameters, is as follows:

(1) The input of the function is a vector of the weights of the components of the material *p*, a vector of values of a certain physical parameter of the components *t*, the number of chemical elements in the material *n*.
(2) According to the corresponding formulas (Table 2), values are calculated for the mean, weighted mean, geometric mean, geometric weighted mean, entropy, weighted entropy, range, weighted range, standard deviation and weighted value of the standard deviation of the characteristic.
(3) At the output of the function, 10 obtained characteristic values are supplied.

The function *material*, which allows to get all material parameters in accordance with the source data, is as follows:

(1) The input of the function is the vector of the weights of the components of the material *p*, 8 vectors with the values of the physical parameters of each component, the number of chemical elements *n*, the name of the material.
(2) A vector of initial features is formed. The first value of the vector is assigned *n*. From 1 to 11, assign the values of the first attribute t_1, from 11 to 22, of the second attribute t_2, and so on, respectively, using the function *priznak*, assign the name of the material to the last attribute.
(3) The output of the function is a generated vector of attribute values.

Using the function *material*, databases of positive and negative ions with the corresponding values of physical parameters were formed. As a result, we obtained a library (one_plus) of positive ions, consisting of 301 ions, positive ions with the most frequent valency, found in nature (one_chast_plus)—132 ions, negative ions (one_minus)—53 ions, a library of metal ions (one_metals)—75 ions.

The resulting libraries were used to model superconducting materials.

3.2 Algorithm for determining the critical temperature of new possible superconductors

Using the obtained libraries, the general algorithm for determining the critical temperature of simulated materials can be represented as follows:

(1) Set the number of ions in compound *n*, an empty database, where the simulated materials are added.
(2) Determine the possible combinations between positive ions (one_plus, one_chast_plus) and negative (one_minus) for materials with covalent and ionic bonds; for alloys, possible combinations between metal ions (one_metals).
(3) For each combination, set or calculate the vector of weights *p*. In the case of a material with an ionic or covalent bond, the calculation is based on the valency of the ions in the compound.

Expressions for weights for the case $n = 2$:
Compound: $A_a^{+\alpha} B_b^{-\beta}$
Weights: $p_A = \frac{\beta}{\alpha+\beta}$; $p_B = \frac{\alpha}{\alpha+\beta}$
Expressions for weights for the case $n > 2$:
Compound: $X1_a^{+\alpha} X2_b^{+\beta} \ldots Xn-1_x^{+\chi} Y_y^{-\psi}$
For this case, need to know the ratio between the ions in the material: $x_1 : x_2 : \ldots x_{n-1} : y$, where $y = \sum_{i=1}^{n-1} x_i$, x_i is the fraction of the *i*th positive ion, *y* is the fraction of the negative ion.

Weights: $p_Y = \frac{\alpha * x_1 + \beta * x_2 + \ldots + \chi * x_{n-1}}{\alpha * x_1 + \beta * x_2 + \ldots + \chi * x_{n-1} + \psi * y}$; $p_{Xi} = \left(1 - p_Y\right) * \frac{x_i}{y}$, *i* от 1 до $n-1$.

The same expressions can be applied to materials of the type: $Y_y^{+\psi} X1_a^{-\alpha} X2_b^{-\beta} \ldots Xn-1_x^{-\chi}$

For alloys, weights are set manually.

(1) Create a matrix where each line represents the values of one attribute for each ion that forms the material.
(2) Pass to the function material the obtained *p*, *n*, matrix with the values of the attributes.

(3) Apply the model to predict the critical temperature for the resulting material.
(4) Add the prediction to the database from point 1. In this paper, the names of the ions contained in the material, their weight, and the calculated critical temperature were added as features.

3.3 Results of simulation of superconductive materials

The obtained critical temperature determination algorithms were applied for the following materials:

(1) 15,953 possible compounds with $n=2$. Let us compare the calculated and experimental critical temperatures for some well-known superconductors with $n=2$ (Table 7). And then we give some simulated compounds and their calculated critical temperatures (Table 8). Three models were used, trained on sc_min, sc_mean, sc_max data, and the one_plus database was used as positive ions.
(2) Compounds with $n=3$, with the ratio of ions in the material 1:1:2, 1:2:3, 1:3:4, 2,291,190 possible compounds; the model trained on sc_mean data was used, and the database was used as positive ions one_chast_plus. Some of these are shown in Table 9.
(3) Using the one_chast_plus database obtained in the work as positive ions, only oxygen as a negative ion, for a composition with a ratio of 1:1:1:3, 1:1:2:4 a model based on sc_mean data was applied to 1,498,640 possible materials (Table 10).
(4) A model based on sc_min data was applied to some known (Table 11) and possible (Tables 12, 13) superconducting alloys. In this case, the attributes associated with valency, electron affinity, and the number of chemical elements were removed from the model. Quality metrics for the model: r_2-score = 0.912, mean absolute error = 5.52, mean square error = 97.55.

Table 7 Comparison of the calculated and experimental critical temperatures for some superconductors with $n=2$

Material	T_c, K sc_min	T_c, K sc_mean	T_c, K sc_max	Experimental T_c, °K [15]
MgB_2	21.2	34.10	35.66	33–40
FeSe	6.27	13.55	33.78	8
$NbSe_2$	4.6	5.77	6.58	7
TaS_2	3.10	3.89	4.55	3.25
NbN	11.5	11.71	12.05	16
Mo_3Si	9.52	7.86	7.70	7.2

Table 8 Some simulated compounds with $n=2$

sc_min		sc_mean		sc_max	
Material	T_c, K	Material	T_c, K	Material	T_c, K
AgF_2	61.83	$CuCl_2$	60.73	AgF_2	61.36
CuF_2	61.41	AgF_2	56.23	CuF_2	58.13
$CuCl_2$	56.36	CuF_2	55	$CuCl_2$	57.3
BrO_2	55.6	BrO_2	52	BrO_2	52.06
Br_2O	50.74	Br_2O	48.25	KrF_2	48.26
KrF_2	45.59	KrF_2	43.9	Br_2O	47.13
IF_7	44.66	IF_7	43.89	BeF_2	45.13
IF_3	43.9	BeF_2	43.13	IF_7	40.94
BeF_2	41.24	IF_3	42	IF_3	38.77
AgBr	33.5	AgBr	38.29	MgB_2	35.66
VBr_3	31.72	MgB_2	34.16	AgBr	35.61
$GeBr_2$	31.33	$GaBr_2$	30.5	FeSe	33.78
VCl_3	31.28	$LaBr_3$	29.75	AgI	31.67
$GaBr_2$	31.13	VCl_4	29.6	ZnSe	27.41
$GeBr_4$	31.11	AgI	29.25	$CdBr_2$	25.81
$ZrCl_4$	30.90	$ScBr_3$	29.23	MnSe	25.19
$GaBr_3$	30.75	$ZrBr_4$	29.15	LiF	24.96
$ZrBr_4$	30.74	VCl_3	29.08	TiS_2	24.2
$SbBr_3$	30.71	$SiCl_2$	29	VCl_3	24.12

4 Discussion

In this paper, the random forest method was chosen as a regression model. This is due to the fact that a simpler linear regression method allows to describe well only linear dependencies, and according to [13], a strong linear dependence of the features on the critical temperature was not found, which makes this method not applicable. Neural networks were not used as a model since the quality and implementation of this method strongly depends on the size of the data [17, 18], which in this paper is not enough to implement this method.

The feature selection algorithm, which is used in the work, largely repeats the ADD-DEL procedure [19], except that at each iteration the selected feature is added or

Table 9 Some simulated compounds with $n=3$ type $X1_a^{+\alpha}X2_b^{+\beta}Y_y^{-\psi}$

Material 1:1:2	T_c, K	Material 1:2:3	T_c, K	Material 1:3:4	T_c, K
$Cu_2Cr_2O_5$	92	$SiCu_2O_4$	94	$SiCu_3O_5$	95
$CuCO_3$	83	CCu_2O_4	93	CCu_3O_5	94
$Cu_2Sb_2O_5$	83	$CuCr_2O_4$	83	$CrCu_3O_6$	85
$CuSrO_2$	83	$V_2Cu_4O_9$	82	$V_2Cu_6O_{11}$	82
$CuSiO_3$	83	$CuSr_2O_3$	82	$Sb_2Cu_6O_9$	79
CuSnO3	81	$Cr_2Cu_4O_7$	78	$Al_2Cu_6O_9$	79
$Cu_2B_2O_5$	77	$MnCu_2O_4$	78	$Ga_2Cu_6O_9$	79

Table 10 Some simulated compounds with an ion ratio of 1:1:1:3, 1:1:2:4

Material 1:1:1:3	T_c, K	Material 1:1:2:4	T_c, K
$CdCsAgO_2$	106	$Ag_2Cd_2Cs_4O_5$	102
$ZnRbAgO_2$	106	$AgAuBa_2O_4$	102
$ZnCsAgO_2$	105	$Ag_2Mg_2Rb_4O_5$	102
$MgRbAgO_2$	104	$Ag_2Be_2Rb_4O_5$	102
$CdRbAgO_2$	103	$Ag_2Ca_2Rb_4O_5$	101
$AuBaAgO_3$	103	$Ag_2Zn_2Cs_4O_5$	101
$BeRbAgO_2$	102	$AgRbBe_2O_3$	101
$CaRbAgO_2$	102	$AgZnRb_4O_5$	99
$Au_2Ba_2Cu_2O_7$	100	$Ag_2Cd_2Rb_4O_5$	99
$Au_2Cs_2Ag_2O_5$	94	$Ag_2Ba_2Au_4O_9$	98
$Au_2Rb_2Ag_2O_5$	93	$Cu_2Au_2Ba_4O_9$	96

Table 11 Comparison of calculated and experimental critical temperature for some known alloys

Alloy	Estimated T_c, K	Experimental T_c, K [15, 16]
NbTi	4.6	9.6
PbIn	4.84	7
PbBi	8.18	8.3
Nb_3Sn	15.3	18
Nb_3Al	14.4	16.4
V_3Ga	14.6	14.2–14.6
$Pb_{0.8}Tl_{0.2}$	6.38	6.8
$Pb_{0.8}Bi_{0.2}$	8.19	7.95
$Nb_{0.8}Zr_{0.2}$	9.09	11
$Pb_{0.4}Tl_{0.6}$	5.2	4.6

deleted, not to achieve the greatest improvement of the model among all the features at this iteration, and the condition for improvement or deterioration is checked models. This approach is different in that it requires a larger number of iterations to find the optimal set of features, but this algorithm is simpler and faster to implement each step of the iteration.

The method "feature_importances_" for selecting features of a random forest shows that for the three models built on the basis of data sc_min, sc_mean, sc_max, the first five features in importance are the same, including thermal conductivity, density and valency. The importance of thermal conductivity determines the importance of the concentration of particles that can transfer heat. Assuming that these particles are electrons, we can therefore assume that there is a dependence on the concentration of superconducting particles. The concentration of superconducting particles is related to the London penetration depth, which means the result indirectly confirms the results of works [20, 21]. The importance of density in determining the critical temperature can be related to the role of phonon oscillations in the formation of the superconducting state [22–24]. The importance of valency reflects the effect of valency of copper in cuprate superconducting compounds [25].

The data were divided by the number of chemical elements in order to obtain models of higher quality and to consider the possibility of breaking the data into classes of low-temperature and high-temperature superconductors, since this feature significantly correlates with the critical temperature [13]. The obtained random forest models on all data are not inferior in quality to works [10–12], and the models obtained on the basis of dividing the data by the number of chemical elements have a higher quality. Based on the results of the dividing, it cannot be supposed that the obtained ratios by the number of chemical elements reflect the division of data into classes, but it can be concluded that in all the best models there is a quantity of chemical elements equal to two.

The limitations of using the random forest model to predict new superconducting materials are related to the following. When choosing a random forest as a model, we are forced to limit the interval for predicting the critical temperature to the temperature values in the initial data due to the specifics of constructing this method [26]. This will not allow predicting the room-temperature superconductor, but will allow much more accurately than using linear regression to predict the critical temperature of new superconducting materials, the temperature of which will not exceed the maximum in the initial data (150 K).

The resulting models based on a random forest are of high quality. After the removal of outliers, the quality of the models deteriorated; therefore, models were used to determine the critical temperature of new materials before the outlier removal. As a model for predicting the critical temperature of new superconducting materials except alloys, we used the random forest model, which is built on data sc_mean. It is this model that is used to simplify the writing of algorithms because the selection of features did not significantly increase the quality of the models. Also, models with data dividing by the number of chemical elements were not used due to the fact that the material can show the properties of materials with a different number of chemical elements that are not included in any of these models. Model sc_mean was chosen due to the fact that it has the best quality compared to models based on sc_min, sc_max. For alloys, a model based on data sc_min was used, in view of the expectation of a low critical temperature of these materials. The model based on data sc_max, in view of the availability of data on measuring the critical temperature at high pressures, is not

Table 12 Some simulated alloys with compositions 1:1, 1:2, 1:3, 1:4, 2:3, 1:6

Alloy 1:1	T_c, K	Alloy 1:2	T_c, K	Alloy 1:3	T_c, K
ScAl	22	$AgSc_2$	28	UAg_3	25
LiNa	20	UAg_2	24	$PuAg_3$	24
LiMg	19	$PuAg_2$	23.8	$AgPu_3$	23.2
PuAg	18.8	$FrMg_2$	21.6	KLi_3	22.6
UAg	18.4	$CsCa_2$	21.5	$CsCa_3$	21.8
MoNa	18.1	$FrCa_2$	21.5	$FrMg_3$	21.2
TiAg	18	$RbTi_2$	21.2	$RbCa_3$	21.1
CrLi	18	$ScAg_2$	21.2	$CsTi_3$	20.7
MgCu	17.9	$RbCa_2$	21.2	$FrRb_3$	20.6
BeLi	17.5	$AgTi_2$	19.3	$RbSc_3$	20.5
Alloy 1:4	**T_c, K**	**Alloy 2:3**	**T_c, K**	**Alloy 1:6**	**T_c, K**
UAg_4	25.5	Ag_2Sc_3	23.9	$PuAg_6$	25
$PuAg_4$	25.1	U_2Ag_3	22.6	UAg_6	24.5
$AgPu_4$	24.1	Pu_2Ag_3	22.4	$AgPu_6$	23.6
$NpAg_4$	23.8	Al_2Be_3	21.7	$NpAg_6$	23
KLi_4	21.9	Sc_2Ag_3	21.4	KLi_6	21.8
$CsCa_4$	21.8	Rb_2Ca_3	20.9	$FrMg_6$	21.7
$FrCa_4$	21.7	Fr_2Ca_3	20.9	$CsCa_6$	21.5
$NbAl_4$	21.6	Rb_2Ti_3	20.7	$PoBe_6$	21.4
$FrMg_4$	21.3	K_2Ti_3	20.7	$NbAl_6$	21.3
$RbTe_4$	20.3	Rb_2Sc_3	20.6	$TeAl_6$	21.2

Table 13 Some simulated alloys with compositions 1:1:1, 1:1:2

Alloy 1:1:1	T_c, K	Alloy 1:1:2	T_c, K
AlSrFr	47	$SrFrAl_2$	48
AlScFr	46	$AlFrSr_2$	46
AlBaFr	38	$CsFrBe_2$	45
BeCaFr	37	$AlSrFr_2$	44
BeRbFr	37	$AlScFr_2$	44

recommended for determining the transition temperature under normal conditions.

Note that the obtained assessments of the quality of the random forest model show that this model is not ideal and allows predicting the critical temperature with a certain error; therefore, using this model to predict new superconducting materials, it should be expected that the model works with an error not less than the error obtained model on the test sample. The error of the model indicates that this model with more accuracy makes sense to apply to the prediction of new high-temperature superconductors.

It should be noted that the model does not answer the question of whether the material is superconducting or not, but predicts the critical temperature if the material is superconducting. How the model behaves if applied to a material that is not a superconductor is also a non-trivial research question due to the specifics of constructing a random forest. But in any case, using this model, it is possible to find new, more technologically advanced superconducting materials in comparison with the ones used or possible compositions in which stoichiometry can be slightly changed for the material to manifest superconducting properties.

We also note that the constructed models do not take into account the crystallography of the considered substances, which can be important when considering, for example, various modifications of carbon. The models are not suitable for the application of determining the critical temperature of the same compositions with different crystallographic orientations.

Among the possible superconducting binary compounds with a high critical temperature are AgF_2. No superconductivity has been detected in this material at the moment, but there is a paper [27] in which the authors discovered superconductivity in the Ag-Be-F system with a similar critical temperature obtained. Unfortunately, the authors were unable to establish which phase is superconducting. In [28], the similarity of this material with superconducting cuprates was discussed.

Among the simulated binary compounds, $CuCl_2$ has a high critical temperature. For a long time, researchers discussed the possibility of high-temperature superconductivity of copper chloride [29, 30], but no superconductivity has been detected in copper chlorides.

Also, among the binary compounds, many materials with the highest critical temperatures were obtained, which under normal conditions are in a liquid state, such materials as BrO_2, Br_2O, IF_7, IF_3. Many of the obtained double possible superconducting compounds are crystalline hydrates, and the search for superconductivity in the obtained materials can be related to the influence of the number of water molecules on the temperature dependence of resistance. Among the possible triple superconducting compounds with a high critical temperature, the most frequent systems are: Cu–Cr–O, Cu–Si–O, Cu–C–O, Cu–Sb–O, Cu–Sr–O, Cu–V–O. Among possible compounds consisting of four ions, materials containing silver have a high critical temperature. Among the possible superconducting alloys with the highest critical temperature, alloys with heavy radioactive metals are often found.

5 Conclusion

The critical temperature prediction models are constructed on the widest database containing almost all known superconducting materials. Based on the selection of features, the most important physical parameters of the models are identified. The feature_importance_ method shows that the most important features for the three datasets sc_min, sc_mean, sc_max for predicting the critical temperature in descending order are: range thermal conductivity, weighted geometric mean value of thermal conductivity, standard density deviation value, weighted geometric mean valency value, weighted mean valency value. For the first time, models were constructed by dividing data by the number of chemical elements. According to the constructed models, we can conclude that in all models of the best quality there are materials with the number of chemical elements equal to two, and no data can be divided into low-temperature and high-temperature materials. The results obtained are important for constructing a general theory of superconductivity and for predicting new superconducting materials.

The constructed models are used to model new possible superconducting materials. The resulting materials exceed the known used superconducting materials at a critical temperature and can be a substitute for them. Future studies suggest synthesizing the obtained materials and investigating their superconducting properties. The results of this work can help to discover new classes of superconducting compounds and new more technologically advanced superconductors.

The limitations of the results of the work are related to the used datasets, the limitations of the random forest model, and the quality of the constructed models. With more extensive data on the measured critical temperatures and higher quality models, a more accurate simulation result of new possible superconducting materials can be achieved. In further studies, it is recommended to consider the use of linear regression models and the consideration of nonlinear dependences for predicting room-temperature superconductors.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Availability of data and material The initial data required to reproduce these findings are available to download from [https://github.com/matasovav/DATA_SC], [13]. The processed data required to reproduce these findings are available to download from [https://github.com/matasovav/Prediction-SC].

Code availability The code required to reproduce these findings is available to download from [https://github.com/matasovav/Prediction-SC].

References

1. Cooper LN (1956) Bound electron pairs in a degenerate fermi gas. Phys Rev 104:1189
2. Bardeen J, Cooper LN, Schrieffer JR (1957) Theory of superconductivity. Phys Rev 108:1175
3. Kulić ML (2000) Interplay of electron-phonon interaction and strong correlations: the possible way to high-temperature superconductivity. Phys Rep 338(1–2):1–264
4. Mishchenko AS (2009) Electron–phonon interaction in high-temperature superconductors. UFN 179(12):1259–1280, 2009; Phys Usp 52(12):1193–1212, 2009
5. Vasiliev BV (2011) Superconductivity as a consequence of an ordering of the electron gas zero-point oscillations. Physica C 471:277–284
6. Vasiliev BV (2012) Superconductivity and condensation of ordered zero-point oscillations. Physica C 483:233–246
7. Vasiliev BV (2013) Superconductivity, superfluidity and zero-point oscillations. In: Recent advances in superconductivity research. Nova Publisher, New York, pp 249–280
8. Matthias BT (1955) Empirical relation between superconductivity and the number of electrons per atom. Phys Rev 97:74–76
9. Owolabi TO, Akande KO, Olatunji SO (2015) Estimation of superconducting transition temperature T_C for superconductors of the doped MgB_2 system from the crystal lattice parameters using support vector regression. J Supercond Nov Magn 28:75–81
10. Owolabi T, Akande A, Olatunji S (2014) Prediction of superconducting transition temperatures for fe-based superconductors using support vector machine 35:12–26
11. Stanev V, Oses C, Kusne AG et al (2018) Machine learning modeling of superconducting critical temperature. NPJ Comput Mater 4:29

12. Hamidieh K (2018) A data-driven statistical model for predicting the critical temperature of a superconductor. Comput Mater Sci 154:346–354
13. Anton Matasov, Varvara Krasavina (2020) Visualization of superconducting materials. Preprint at https://doi.org/10.13140/RG.2.2.27686.47682
14. Breiman (2001) Random forests. Mach Learn 45(1):5–32
15. Kleiner R, Buckel W (2016) Superconductivity an introduction. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
16. Poole CP (2000) Handbook of superconductivity. Academic Press, New York
17. Romero N, Gutoski M, Hattori L, Lopes H (2017) The effect of data augmentation on the performance of convolutional neural networks, conference: Brazilian Society of Computational Intelligence. Niterói, Rio de Janeiro, pp 1–12
18. Crowther PS, Cox RJ (2005) A method for optimal division of datasets for use in neural networks. In: Khosla R, Howlett RJ, Jain LC (eds) Knowledge-based intelligent information and engineering systems. KES 2005. Lecture Notes in Computer Science, vol 3684. Springer, Berlin, Heidelberg
19. Verma NK, Salour A (2020) Feature selection. In: Intelligent condition based monitoring. Studies in systems, decision and control, vol 256. Springer, Singapore
20. Matasov AV (2019) Characteristic lengths and Plasmon superconductivity mechanism of some high-temperature superconductors. In: 2019 International youth conference on radio electronics, electrical and power engineering (REEPE), Moscow, Russia, pp 1–4
21. Matasov AV (2019) In: AIP conference on proceedings 2163, 020005-1–020005-5
22. Jha PK, Gupta SD, Gupta SK (2012) Puzzling phonon dispersion curves and vibrational mode instability in superconducting $MgCNi_3$. AIP Adv 2:022120
23. Dolgov OV, Mazin II, Golubov AA, Savrasov SI, Maksimov EG (2005) Critical temperature and enhanced isotope effect in the presence of paramagnons in phonon-mediated superconductors. Phys Rev Lett 95:257003
24. Shimahara H (2003) Isotope effect in superconductors with coexisting interactions of phonon and nonphonon mechanisms. J Phys Soc Jpn 72:1851–1854
25. Tallon JL (2005) Oxygen in high-T_c cuprate superconductors. In: Narlikar AV (ed) Frontiers in superconducting materials. Springer, Berlin
26. Biau G (2012) Analysis of a random forests model. J Mach Learn Res (JMLR) 13(1):1063–1095
27. Grochala W, Porchc A, Edwards PP (2004) Meissner-Ochsenfeld superconducting anomalies in the Be–Ag–F system. Solid State Commun 130:137–142
28. Gawraczyński J et al (2019) Silver route to cuprate analogs. PNAS 116(5):1495–1500
29. Saerbeck T, Pereiro J, Wampler J, Stanley J, Wingert J, Shpyrko OG, Schuller IK (2013) Ferromagnetism in partially oxidized CuCl. J Magn Magn Mater 346:161–165
30. Chu CW, Early S, Geballe TH, Huang CY (1978) Study of cuprous chloride under pressure. J Less Common Met 62:463–467