IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Towards the advanced predictive modelling in epidemiology

To cite this article: C Brester *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **537** 062002

View the article online for updates and enhancements.

This content was downloaded from IP address 219.131.242.170 on 13/10/2019 at 03:40

IOP Conf. Series: Materials Science and Engineering **537** (2019) 062002 doi:10.1088/1757-899X/537/6/062002

Towards the advanced predictive modelling in epidemiology

C Brester[1,3,4], T P Tuomainen[2], A Voutilainen[2], J Kauhanen[2], E Semenkin[3] and M Kolehmainen[1]

[1] Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, Finland
[2] Institute of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland
[3] Institute of Computer Science and Telecommunications, Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia

[4] E-mail: christina.brester@gmail.com

Abstract. Data-driven prediction systems used in epidemiological studies are still unsatisfactory from a practical point of view. Different pitfalls should be considered while transferring technologies from research to practice. The proposed k-Nearest Neighbors approach is designed to make disease-related predictions in a more holistic manner: we detect cases of novelty among unobserved subjects to identify situations when model predictions are not reasonably valid. Moreover, it copes with overlapping classes, finds new examples which cannot be labelled with the high confidence and reveals healthy subjects in the training data who might be at risk. Additionally, variable selection is built-in to select relevant predictors. The approach was applied to predict cardiovascular diseases based on the data collected within an ongoing follow-up study undertaken in Eastern Finland. According to the experimental results, our proposal allows increasing the accuracy of predictions made.

1. Introduction

Predictive modeling as an essential part of preventive medicine has recently evolved into the main technological ingredient of a novel up-and-coming field called precision public health [1]. This field involves advanced data-driven methods and applies them to prevent diseases, understand risks better and promote health [2, 3]. Due to constant expanding of data storage capacity and development of highly productive hardware tools as well as intellectual learning algorithms, it has become possible to move predictive models to a new level of much higher efficiency and reliability [4]. However, there are a number of issues [5] which should be addressed while training predictive models and in this paper, we are raising some of them.

As opposed to traditional statistical approaches in bioinformatics which are mostly based on averaging and comparing with other observed subjects, precision public health requires from the models applied to be more subtle and detect specific subgroups or individuals to treat them adequately. This relates to the sample representativeness which is always limited with participants of the particular study [6]. An external validation aims to estimate the biasness of the data used and the reliability of generalizations made and their extrapolations to unobserved subjects. The model trained should identify cases of novelty which cannot be processed confidently based on the patterns revealed from the sample. This is the first issue which we are covering in this paper.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

In more detail, there is a clear difference between two types of wrong predictions: false positive and false negative. The second one has much more serious consequences and the number of these mistakes should be minimum. The use of ongoing follow-up study data for training predictive models entails additional risks of mislabelling training examples. Disease events and hospitalizations give us information about sick subjects, whereas there are many health problems which do not have any symptoms (silent stroke and silent myocardial infarction) and stay undiagnosed before the thorough clinical screening [7, 8]. Therefore, if follow-up examinations are conducted very rarely, there is a risk to give a label 'healthy' mistakenly to those who are not healthy anymore just because this information is not presented in the up-to-date records of events and hospitalizations. Moreover, it is hard to say definitely whether healthy subjects will remain healthy for a long time in the future. To avoid wrong 'healthy' labels in the training data which may cause false negative predictions, training examples should be processed carefully. If there are some overlapping regions of sick and healthy subjects in the training data, it should give the reason to treat these healthy subjects as a risk group. This is the second issue which is considered in our study.

Besides, to train an effective model, predictor variables associated with a particular disease should be included in the vector of inputs. The lack of relevant variables leads to the deterioration of the model predictive ability. It is more reasonable to start a learning process with an extended variable set and select relevant variables during it than to pre-select a smaller subset based on the existing knowledge and miss some informative data [9]. In our study, this point is also highlighted to some extent.

More specifically, this work is primarily focused on handling overlapping classes in the cardiovascular predictive modeling. Popular techniques applied to solve the problem of overlapping classes include a k-Nearest Neighbours (kNN) approach [10], a fuzzy set representation [11], and Support Vector Data Description (SVDD) [12]. The method proposed in this paper is based on the k-Nearest Neighbours approach, which has been complemented with a novelty detection and a feature selection technique. We applied our method to one of the most extensive study populations in the field of epidemiology, the Kuopio Ischemic Heart Disease [13] cohort, and proved that it could increase sensitivity of predictive modelling as well as detect the risk group of subjects mistakenly labelled as 'healthy'.

2. Methods

In the original k-Nearest Neighbours approach developed for handling overlapping classes [10], the possibility to detect outliers which look like distant isolated points of the training data is built in the approach. However, many predictive models are robust to outliers in the training data if the number of outliers is reasonable (such as Random Forest [14]). In the epidemiological predictive modeling, the detection of outliers in the test data, which represent cases of novelty, is specifically important because it allows preventing the model from unjustified extrapolations and wrong predictions. Therefore, in our method proposed, unobserved test subjects could be labelled as: *sick* or *healthy* with the high confidence if they belong to non-overlapping regions; *at risk*, which corresponds to subjects from overlapping regions; or *novelty* if the case subjects demonstrate their specificity and these phenomena could not be explained based on the training data.

The approach includes three main steps:

Step 1. Determine sick and healthy training examples which overlap and do not overlap in the space of predictor variables.

Let $\overline{S}$ and $\overline{H}$ denote sick and healthy subjects of the training data which belong to the overlapping regions, S^* and H^* denote sick and healthy subjects which are from the non-overlapping regions.

Step 2. Train two predictive models using two different training sets:

$TrainSet_1 = \overline{S} \cup S^* \cup H^*$ and $TrainSet_2 = \overline{H} \cup H^* \cup S^*$, where all training examples are labelled with -1 or 1, indicating healthy and sick, respectively.

Step 3. Assign labels for the test data $TestSet$ based on the following scheme:

- Check whether $TestSet_i$ is a case of novelty, $i=\overline{1,N_{test}}$, where N_{test} is the number of test examples: if yes, set its label to *novelty*; if no, proceed with the next step.
- Apply both trained models to $TestSet_i$. If the first model returns -1 (*healthy*) or the second model returns 1 (*sick*), this prediction is given with the high confidence and should be accepted; otherwise set its label to *at risk.*

In Step 1, to define overlapping and non-overlapping regions, k-Nearest Neighbours should be found for each training subject: if the number of nearest neighbours which belong to the other class is larger than a threshold *K_boundary*, then a training example is considered to be from the overlapping region. k-Nearest Neighbours are determined using the Euclidian distance. The number of nearest neighbours is denoted by *K_nearest*.

In Step 3, to check if a test subject is a case of novelty, reverse k-Nearest Neighbours from the training data should be found for each test subject: if the number of reverse k-Nearest Neighbours is lower than a threshold *K_noise*, a test example is admitted to be a case of novelty. Reverse k-Nearest Neighbours of a test example $TestSet_i$ are subjects from the training data which have $TestSet_i$ within their k-Nearest Neighbours.

As a predictive model, we apply Random Forest [15], which is an ensemble of decision trees trained on different sub-samples on the training data. Averaging over ensemble predictions tends to enhance the model performance and tackle over-fitting. The following settings are defined [16]:

- The number of trees in the forest is 250;
- The maximum depth of the tree is 10;
- The function to assess the quality of a split is the Gini impurity;
- Bootstrapping (random sampling with replacement) is True;
- The number of features considered while looking for the best split is $\sqrt{M}$, where M is the number of predictors;
- The minimum number of samples at a leaf node is 1;
- The minimum number of samples needed to split an internal node is 2.

Additionally, variable selection has been incorporated into Step 1. The idea behind this extra step is to minimize overlapping of sick and healthy training subjects by selecting relevant predictors. In this study, we apply Random Search [17] to demonstrate possible benefits of variable selection.

Random Search is implemented as follows:

Step 1. Start with a binary vector X^* of the length M and fill it with 1, which means that all variables are included at the beginning of the search. Evaluate the number of overlapping subjects in the training data $F(X^*)$.

Step 2. Repeat *N* times Step 2.1.

Step 2.1. Based on X^*, generate *K* candidate solutions X_i, $i=\overline{1,K}$ changing each coordinate of X^* to the opposite value $0\to1$ or $1\to0$ with the probability p_m: 1 means that the corresponding variable is selected, whereas 0 means that it is not. From the set of X_i, $i=\overline{1,K}$, choose the best candidate X_{min}, which provides the minimum number of overlapping training subjects $F(X_{min})$. If $F(X_{min})< F(X^*)$, then $X^*= X_{min}$.

After variable selection models are trained on the set of selected predictors.

3. Dataset description

The epidemiological data KIHD used in this study has been collected during the ongoing project initiated in 1984 to investigate risk factors of cardiovascular diseases (CVDs) and some other disorders in the population sampled in the city of Kuopio and its surrounding communities in Eastern Finland [13]. KIHD is one of the most extensively characterized epidemiological study populations in the world, with thousands of biomedical, psychosocial, behavioural, and clinical variables.

In this paper, we use an excerpt of KIHD which contains results of examinations of 60-81-year-old men in 2006–2008. The predictor variables have been produced from this information and utilized to predict CVDs from the examination till 2015. CVD diagnoses from this period, recorded by the national Hospital Discharge Register maintained by the National Institute for Health and Welfare and the national Causes of Death Register maintained by Statistics Finland, have been considered to generate an output variable with two possible values 'healthy' or 'sick'. Healthy subjects do not have any CVD-related diagnosis in their records from 2006–2008 till 2015, whereas sick subjects may have incidents such as stroke, coronary heart disease (CHD), acute myocardial infarction (AMI).

Some pre-processing was applied, which led us to the dataset with 775 subjects and 81 predictors:

1) We removed variables containing more than 30% of missing values, as a result, the number of variables was reduced from 86 to 81. Subjects represented in the dataset with rows having more than 10% of gaps were also excluded: the sample size decreased from 1241 to 1229.

2) Subjects who had any CVD-related diagnosis in the examination records in 2006-2008 were excluded: the sample size reduced to 775.

After these steps, we obtained the dataset with 417 healthy and 358 sick subjects, which was relatively balanced.

4. Results and discussion

The 10-fold cross-validation procedure with stratification was performed to estimate the Random Forest performance (accuracy, sensitivity, specificity) in the experiments. To standardize the range of variables, variance scaling was applied. First, we trained this model on the full set of input variables, the proposed kNN-based approach was not applied. As a result, we obtained the following confusion matrix:

Table 1. Confusion matrix obtained by the Random Forest model.

	Actual Sick	Actual Healthy
Predicted Sick	197	120
Predicted Healthy	161	297

The basic measures were calculated based on table 1: accuracy = 63.74%, sensitivity = 55.03%, specificity = 71.22%. This result was taken as a baseline. The number of false negative predictions is quite high: 161 subjects who should be under medical supervision to prevent incidents of CVDs were mistakenly labelled as healthy. The amount of these errors should be as minimum as possible. Meanwhile, false positive predictions mean overdiagnosing and possible additional clinical tests with no actual reason. However, the model may also predict diseases which will occur in the nearest future but this information has not been in the records yet.

Next, we applied the proposed kNN-based approach to make predictions for the overlapping groups of sick and healthy subjects more carefully. The following settings were chosen: $K\_noise = 2$, $K\_boundary = \{3, 5, 7, 9\}$, $K\_nearest = \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}$. In figure 1, we demonstrate how these settings affect the ratio of confidently labelled subjects to those who are in the risk group or cases of novelty.

Generally, we note that increasing *K_nearest*, we decrease the amount of confidently labelled subjects and increase the number of subjects in the risk group (overlapping region), while *K_noise* and *K_boundary* are constant. This is because the more nearest neighbours are taken into account, the more subjects from the other class might be detected nearby, which increases chances of training subjects to be treated as from overlapping regions.

At the same time, the increase of *K_nearest* causes the decrease in the number of subjects categorized as cases of novelty. Larger values of *K_nearest* correspond to the higher number of reverse nearest neighbours which have a considered test subject within their k-Nearest Neighbours.

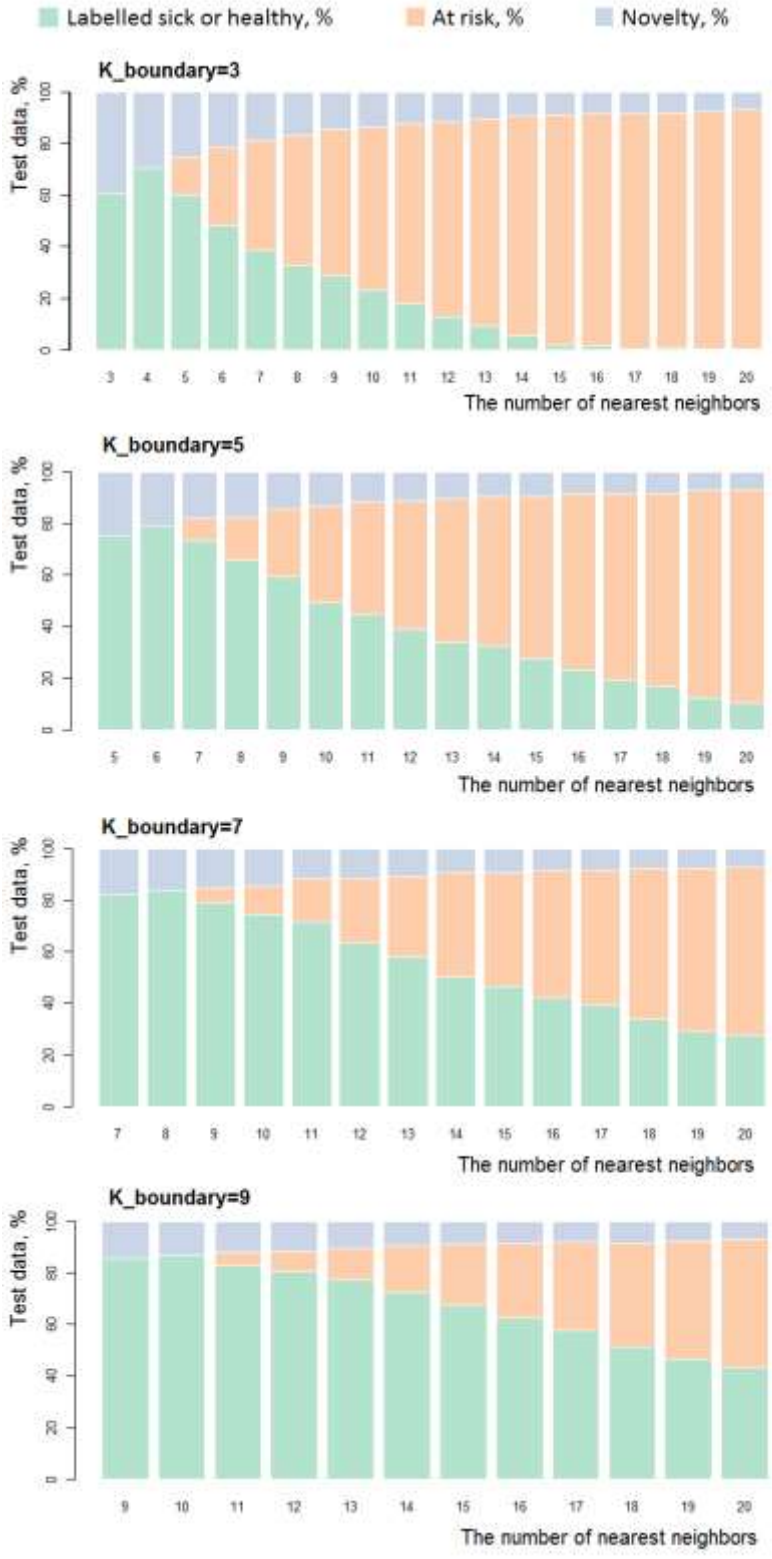

Figure 1. Percent of confidently labelled subjects for different settings.

As we increase *K_boundary*, while *K_nearest* and *K_noise* are constant, we increase the number of test subjects labelled. This happens because we soften our threshold *K_boundary* and fewer training subjects are considered to be from overlapping regions.

Then, we chose two moderate levels of *K_boundary*, which were 5 and 7, to investigate how sensitivity and specificity changed with the growth of *K_nearest*. Test subjects from the risk group or cases of novelty were processed as *required a medical check-up* because our model could not make confident predictions for them. Therefore, healthy test subjects who were considered to be at risk or cases of novelty referred to false positive predictions, whereas sick test subjects who were required a medical check-up (at risk or cases of novelty) related to true positive predictions. In figures 2 and 3, we illustrate how sensitivity and specificity vary with the increase of *K_nearest*. Blue and red horizontal dashed lines refer to the baseline. Asterisk labels represent sensitivity (the upper plot) and specificity values (the lower plot).

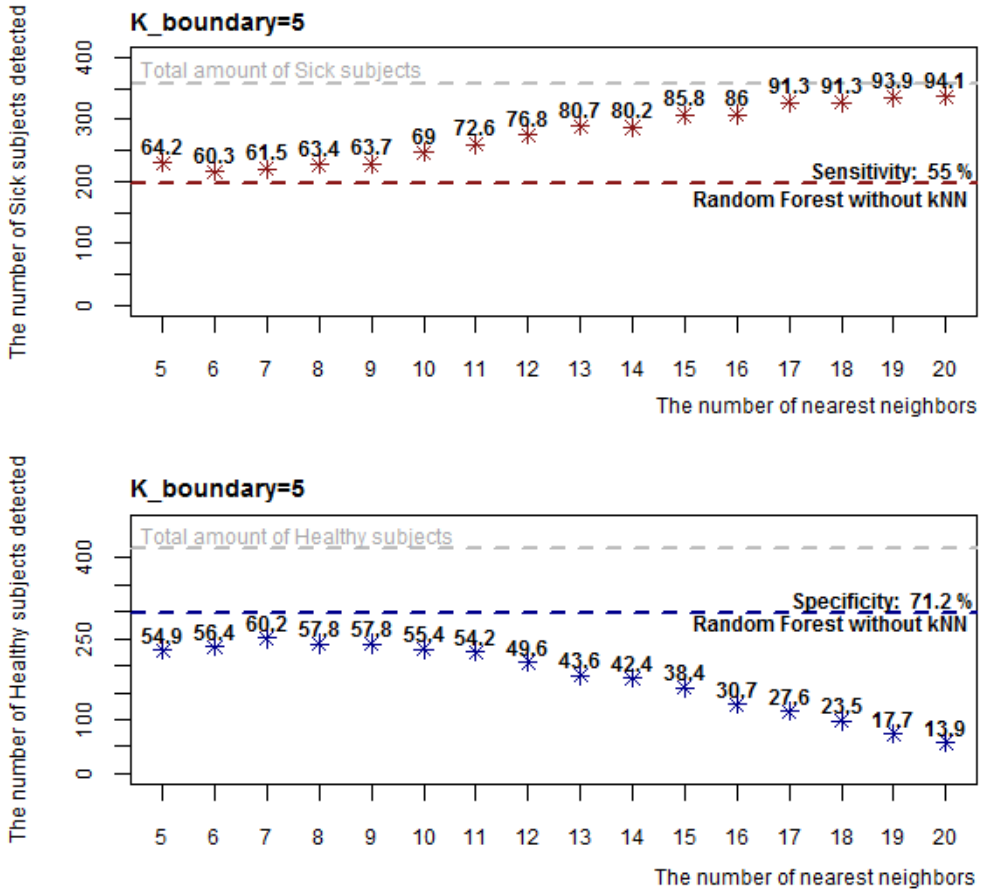

Figure 2. Analysis of sensitivity and specificity for different levels of *K_nearest* and *K_boundary* = 5.

The increase of *K_nearest* leads to the growth of sensitivity and the reduction in specificity. On the one hand, the reduced specificity might be treated as overdiagnosing and the cost of the increased sensitivity. On the other hand, false positive predictions may recognize incidents of silent diseases with no symptoms or CVDs from the nearest future, which have not been in the records yet. If we compare how sensitivity and specificity change with the increase of *K_nearest* for two different levels of *K_boundary*, we may note that for *K_boundary* = 5 they increase or decrease more abruptly, whereas in case of *K_boundary* = 7 changing is more smooth.

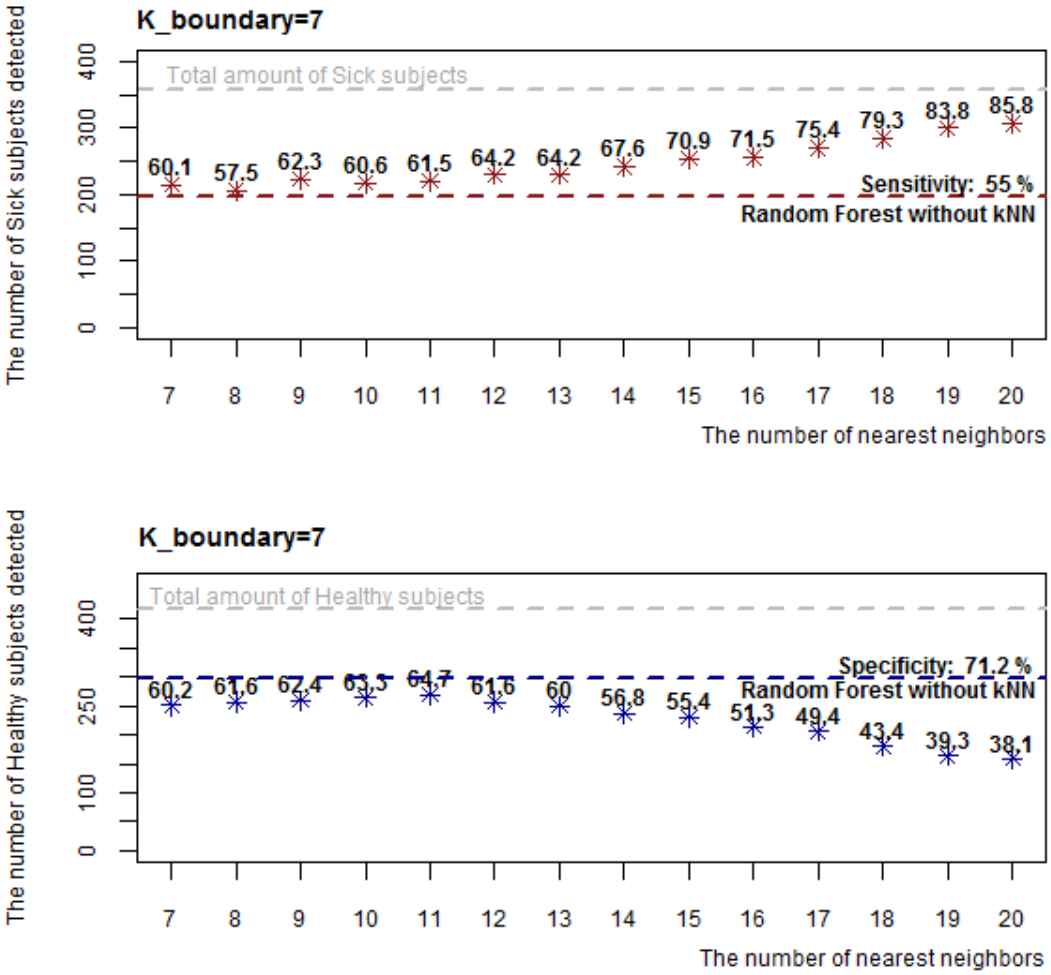

Figure 3. Analysis of sensitivity and specificity for different levels of *K_nearest* and *K_boundary* = 7.

We also investigated how the application of the kNN-based approach affected the accuracy of predictions. To estimate the accuracy, we took into account predictions made with the high confidence and analysed how it varies for the different number of nearest neighbors *K_nearest*. Figure 4 proves that when we categorize subjects not only as *sick* and *healthy* but also use the additional categories *at risk* and *novelty*, it allows us to increase the accuracy of predictions. For example, with the following settings *K_nearest* = 11 and *K_boundary* = 5 we could make predictions for 45% of the test subjects, increase the accuracy from 63.7% to 69.9% (9.7% of the relative improvement) and achieve 72.6% sensitivity and 55.4% specificity. If *K_nearest* = 12 and *K_boundary* = 5, we made predictions for 39% of the test subjects, obtained 71.1% accuracy (11.6% of the relative improvement) and achieved 76.8% sensitivity and 54.2% specificity. Thus, the kNN-based approach enabled us to enhance the model sensitivity and diminish the number of false negative predictions; increase the accuracy by choosing subjects for whom the model could make predictions with the high confidence; find healthy subjects who might be at risk and detect cases of novelty which were quite specific and could not be treated by the model trained.

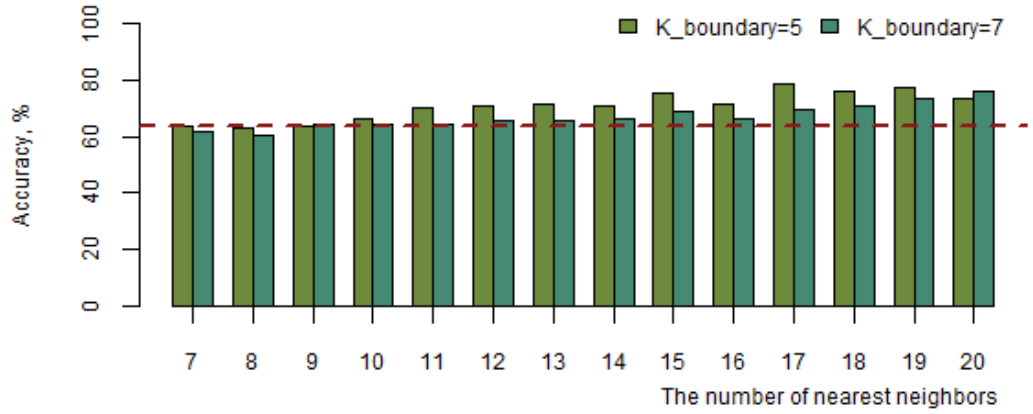

Figure 4. Accuracy of predictions made with the high confidence.

In the additional experiment, we incorporated variable selection into the proposed approach. This step aims at selecting the relevant predictors and, consequently, reducing the costs of clinical tests needed to collect this data. The following settings of Random Search were chosen: $p_m = 0.1$, $N = 30$, $K = 20$. Table 2 contains the results of this experiment: the same metrics obtained without variable selection are given for the comparison.

Table 2. Application of the kNN-based approach with variable selection.

	Feature Selection					All Features				
K_nearest	9	10	11	12	13	9	10	11	12	13
K_boundary	5	5	5	5	5	5	5	5	5	5
Labelled, %	62.7	53.8	49.7	42.8	38.7	59.2	49.4	44.7	38.8	33.7
At risk, %	22.5	32.9	38.5	44.9	50.2	26.2	37.3	43.9	50.1	55.9
Novelty, %	14.8	13.3	11.9	12.3	11.1	14.6	13.3	11.5	11.1	10.5
Sensitivity, %	67.3	70.7	72.6	74.3	77.1	63.7	69.0	72.6	76.8	80.7
Specificity, %	55.2	53.0	51.1	45.6	43.4	57.8	55.4	54.2	49.6	43.7
Accuracy, %	63.0	66.4	67.0	66.6	68.3	63.8	66.3	69.9	71.1	71.3
The number of selected features	39	41	40	42	39	81	81	81	81	81

The results obtained demonstrate a possibility to decrease the number of predictors at least by a factor of two and maintain approximately the same level of the model performance. If we compare two cases with and without variable selection which provide us with the same amount of labelled test subjects (highlighted with gray), we note that all the metrics (accuracy, sensitivity and specificity) are similar, whereas the number of predictors used much lower. This experiment proves one more advantage of the proposed approach. The application of more advanced search strategies, instead of the presented greedy search, may lead to even better results.

5. Conclusion

Modern concepts of predictive modeling in epidemiology require advanced and subtle methods of data utilizing. Models applied should be sensitive and flexible to distinguish the situations when knowledge discovered from observed training subjects could be extrapolated to unobserved ones. Since inaccurate predictions may lead to dramatic consequences, only prognoses made with the high confidence should be accepted.

In this paper, we have proposed a kNN-based approach that enables us to reveal three categories within unobserved subjects, which are *reasonably labelled*, *at risk* and *cases of novelty*. If a new

example looks like a distant isolated point and clearly differs from the training data, it relates to cases of novelty. When unobserved subjects are from the regions where healthy and sick training examples overlap heavily, they are processed as a risk group. Moreover, the detection of overlapping regions may help to find healthy subjects from the training data who are at risk. All the other subjects are labelled as sick or healthy with the high confidence.

This approach allowed us to increase the model sensitivity and accuracy of predictions made. Another advantage of our proposal is a built-in variable selection step. Finally, the presented categorization into three groups is intuitively clear as well as the reasons behind it, therefore, the results obtained are easily interpreted.

Acknowledgements
The reported study was funded by Russian Foundation for Basic Research, Government of Krasnoyarsk Territory, Krasnoyarsk Regional Fund of Science, to the research project: 18-41-242011 «Multi-objective design of predictive models with compact interpretable strictures in epidemiology».

References

[1] Desmond-Hellmann S 2016 Progress lies in precision *Science* **353(6301)** 731 doi: 10.1126/science.aai7598
[2] Dowell S F, Blazes D and Desmond-Hellmann S 2016 Four steps to precision public health *Nat News* **540(7632)** 189 doi:10.1038/540189a
[3] Weeramanthri T S, Dawkins H J S, Baynam G, Bellgard M, Gudes O and Semmens J B 2018 Editorial: Precision public health *Front. Public Health* **6(121)** doi: 10.3389/fpubh.2018.00121
[4] Beam A L and Kohane I S 2016 Translating artificial intelligence into clinical care *JAMA* **316(22)** 2368-9 doi:10.1001/jama.2016.17217
[5] Shah N D, Steyerberg E Wand Kent D M 2018 Big data and predictive analytics: Recalibrating expectations *JAMA* **320(1)** pp 27–8 doi:10.1001/jama.2018.5602
[6] Bernard A 2017 Clinical prediction models: A fashion or a necessity in medicine? *Journal of Thoracic Disease* **9(10)** 3456–57 http://doi.org/10.21037/jtd.2017.09.42
[7] Aghdam M R F, Vodovnik A and Sund B S 2016 Sudden death associated with silent myocardial infarction in a 35-year-old man: A case report *J Med Case Rep* **10** 46 doi: 10.1186/s13256-016-0823-9
[8] Benjamin E J, Blaha M J, *et al* 2017 Heart disease and stroke statistics-2017 A report from the American Heart Association *Circulation* **135(10)** e146-e603 doi:10.1161/CIR.0000000000000485
[9] Brester C, Kauhanen J, Tuomainen TP, Voutilainen S, Rönkkö M, Ronkainen K, Semenkin E and Kolehmainen M 2018 Evolutionary methods for variable selection in the epidemiological modeling of cardiovascular diseases *BioData Mining* **11(18)** 10.1186/s13040-018-0180-x
[10] Tang Y and Gao J 2007 Improved classification for problem involving overlapping patterns *IEICE Transactions on Information and Systems* **90(11)** 1787–95
[11] Visa S and Ralescu A 2003 Learning imbalanced and overlapping classes using fuzzy sets *Proceedings of the ICML* **3** 97–104
[12] Xiong X, Wu J and Liu L 2010 Classification with class overlapping: A systematic study *The 2010 International Conference on E-Business Intelligence* 491–7
[13] Salonen J T 1988 Is there a continuing need for longitudinal epidemiologic research? The Kuopio Ischaemic Heart Disease Risk Factor Study *Ann Clin Res* **20(1-2)** 46–50
[14] Díaz-Uriarte R, Alvarez de Andrés S 2006 Gene selection and classification of microarray data using random forest *BMC Bioinformatics* **7(3)** doi: 10.1186/1471-2105-7-3
[15] Breiman L 2001 Random Forests *Machine Learning* **45(1)** 5-32
[16] Pedregosa *et al* 2011 Scikit-learn: Machine Learning in Python *JMLR* **12** 2825-30

[17] Cormen T H, Leiserson C E, Rivest R L and Stein C 2009 *Introduction to Algorithms* (MIT Press and McGraw-Hill) p 1292