In general, an ensemble constructing process also includes some other steps:

1. Selection of a method by which to bring diversity into the base models;
2. Selection of a method for model combination;
3. Selection of a type of the base model to be used.

There are two main methods for model combination: averaging and voting. The first one is mostly used for the numeric output combinations, but the voting method is used for the nominal output combination [2].

The most popular ensemble method is represented by the majority voting ensemble. The main idea of this method is that each base classifier votes for a specific class, and the class that collects the majority of votes is predicted by the ensemble. Majority voting method applies to the ensemble fusion method group. The methods of it combine all the outputs of the base classifiers, while the ensemble selection methods try to choose better classifiers among the set of the available base learners [3].

Ensemble modelling can be used for the different purposes – as a tool for the analysis, classification, prediction etc. The ensemble can be constructed using results of different available experiments and computations. In case of constructing ensemble for a concrete system, an ensemble of models can be used as a tool for the system behaviour study and detailed analysis.

III. Overview of Machine Learning Methods

There are four machine learning methods briefly described in this article that are used for the model training to design a heterogeneous model ensemble.

A. Genetic Algorithm

Genetic algorithm (GA) presents one of two basic approaches in computer science that copies evolutionary mechanisms. The GAs use two genetic operators – crossover and mutation to produce new solution candidates. The crossover produces offspring by combining part of the parent from existing generation. The mutation helps prevent the premature convergence by randomly sampling new points on the search place.

Extension of the GA, such as GA with offspring selection, is used as one of the methods for model training in the present research. Offspring selection was proposed by M. Affenzeller and S. Wagner [4] as an additional step to the standard selection procedure of GA performed after the crossover and mutation.

The main idea of the offspring selection is that a certain ratio of the next generation has to be filled with children that outperform their parents. During the procedure of selection, the fitness value of the produced offspring is compared with the fitness values of its own parents in order to decide whether or not the evenly produced offspring is accepted as a member of the next generation [4].

Procedure of GA with offspring selection consists of three sequential steps:

1. Selection of two parent chromosomes;
2. Child generation by using the crossover and mutation operators;
3. Child is accepted as a member of the next generation if it outperforms his own parents or its chromosome is unique in the actual generation [5].

The termination criterion of GA with offspring selection is defined by a maximum number of generations that can be reached or there is no possibility to generate further on a sufficient number of children outperforming their parents.

B. Support Vector Machine

Support vector machine (SVM) is a learning system that uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from the optimisation theory that implements a learning bias derived from the statistical learning theory [6].

A basic idea of the support vector machines is to find an optimal hyperplane for the linearly separable patterns [7].

The SVM provides a useful technique for data classification tasks. It performs classification by constructing an *N*-dimensional hyperplane that optimally separates the data into the specific categories [8].

In the paper, the SVM is selected as one of the methods for model training, because it is suitable for binary classification tasks and has already shown good results in the medical diagnostics, optical character recognition, electric load forecasting and other fields [9].

C. Artificial Neural Networks

Artificial neural network (ANN) can be described as an extremely simplified model of the brain cells that cooperate with each other to perform the desired function. ANN can be used for different tasks, such as classification, noise reduction and prediction.

An artificial neuron [10] is a computational model inspired by natural neurons. Natural neurons receive signals through synapses located on the dendrites or the membrane of the neuron. When the signals received are strong enough (surpass a certain threshold), the neuron is activated and emits a signal through the axon. This signal might be sent to another synapse, and might activate other neurons.

The complexity of real neurons is highly abstracted when modelling artificial neurons. These neurons basically consist of the inputs (like synapses), which are multiplied by the weights (strength of the respective signals), and then computed by a mathematical function, which determines the activation of the neuron. Another function computes the output of the artificial neuron (sometimes depending on a certain threshold) [10].

One of the main advantages of the ANN is the opportunity to retrieve hidden information that allows solving complex problems. But in case of using neural networks for any type task, the main rule that should be obeyed is to prevent overtraining of a neural network. An overtrained neural network becomes unable to detect unique patterns in the original data. Another advantage of the neural networks is the ability to generalise and produce both linear and non-linear outputs [11].

D. Random Forests

The method called random forests (RF) is based on the decision trees and can be described as ensemble classifiers, which use many decision tree models [12]. The method is based on the aggregation of a large number of decision trees. The methodology of the RF is used to address two main classes of the problems: to construct a prediction rule for a supervised learning problem and to assess and rank variables with respect to their ability to predict the response.

There are several types of RFs that are characterised by the way each individual tree is constructed, the procedure used to generate the modified datasets, on which each individual tree is constructed, and the way the predictions of each individual tree are aggregated to produce a unique consensus prediction [12].

For each tree training nearly 2/3 of the selected dataset is used, but the remaining data are used to estimate a prediction error and importance of input variable. The class assignment is made by a number of votes from all trees; and for the regression, the average of the results is used.

The RFs have several advantages: there is no need for pruning trees; the accuracy and the variable importance are generated automatically; and it is easy to set parameters in the RF. In addition, the overfitting is not a problem for the RF, and they are not very sensitive to outliers in the training data. Apart from the mentioned advantages, the RFs have limitations that arise in case of the regression analysis. The first one is that regression cannot predict beyond a range of training data, and the second limitation is that its extreme values are often not predicted accurately – underestimating highs and overestimating lows [13].

The RF becomes a major analysis tool in different fields, especially in bioinformatics. Different investigations show that usually results of random forests are quite good.

In the context of the present research, random forests are used as one of the machine-learning algorithms for model training.

IV. Model Training

A. Problem Statement

Timely diagnostics of the cancer can help reduce a number of people deaths from the cancer disease. Ensemble modelling can be used for cancer diagnostics procedure by virtue of the analysis of the results of the medical examination. The cancer diagnostics procedure can be considered a two-class classification task where the main goal is to correct classification of each given instance depending on the values of attributes, e.g. Alpha-fetoprotein, cancer antigen 125, C-reactive protein, bilirubin etc.

In the present research, datasets of breast cancer, melanoma and respiratory system cancer have been used. Each dataset contains measured blood values and tumour markers of patients.

Topicality of this problem is defined as follows. There are different types of breast cancer, e.g. ductal carcinoma, lobular carcinoma and invasive breast cancer. The most common type is ductal carcinoma. The breast cancer occurs in both men and women, although the male breast cancer is rare. The average number of new cases of breast cancer is 124.6 per 100,000 women per year. The average number of deaths is 22.6 per 100,000 women per year [14], [15].

Melanoma is a form of cancer that begins in the melanocytes – cells that make the pigment melanin. It may begin in a mole and it will be skin melanoma, but can also begin in other pigmented tissues, such as in the eye or in the intestines. The majority of the melanomas are black or brown, but they can also be skin-coloured, pink, red, purple, blue or white. There are four basic types of melanoma. Three of them occupy only the top layers of the skin and sometimes become invasive. The fourth type of melanoma is invasive from the beginning and this type is more serious because it may spread to other areas of the body. The average number of new cases of melanoma of the skin is 21.3 per 100,000 men and women per year. The average number of deaths is 2.7 per 100,000 men and women per year [15], [16].

Respiratory system cancer is classified as cancer that affects any part of the respiratory system, which includes the lungs, bronchus and pleura. One of the major causes of respiratory cancer is cigarette smoking, a large contributor to the high incidence of lung cancer. Other respiratory cancers, such as mesothelioma, can be caused by occupational exposure to asbestos. Lung cancer is a common cancer of the respiratory system and also the most commonly occurring cancer. The average number of new cases of lung and bronchus cancer is 60.1 per 100,000 men and women per year. The average number of deaths is 49.5 per 100,000 men and women per year [15].

Diagnosing any kind of cancer in its earliest stage allows decreasing a number of deaths and starting treatment in a timely manner. Ensemble modelling is considered to be one of the possible tools for the early diagnostics of cancer.

B. Settings of Algorithms

Most classification problems can be solved using the neural network with one hidden layer. Results of the cross-validation have shown that to train models it is sufficient to use a neural network with one hidden layer for one part of experiments. According to these results, the 2nd hidden layer is used for training only in some experiments. Apart from a number of hidden layers, the following ANN parameters are used, i.e., decay and a number of nodes in each hidden layer. The first parameter is used for the training phase of the neural network. It determines the strength of the regularisation and is set to a value between 0.001 (weak regularisation) to 100 (very strong regularisation).

There are three parameters that can be changed for random forests: the number of trees, M and R. As suggested by HeuristicLab, the number of trees is the number in the range between 50 and 100. The parameter M corresponds to the ratio of features that will be used in the construction of individual trees and it should be in the range (0; 1]. The parameter R is the ratio of the training set that will be used in the construction of individual trees. It should be adjusted depending on the noise level in the dataset in the range from 0.66 (low noise) to

0.05 (high noise). This parameter should be adjusted to achieve a good generalisation error.

The parameters for the SVM represented in the HeuristicLab are: cost, degree, gamma, kernel type, nu, SVM type. There are two possible values of the used support vector machine type – NU-SVM and C-SVM. In case of using C-SVM, it is necessary to define the value of cost parameter or C. In case of using NU-SVM, the value of NU parameter should be defined. The kernel type parameter determines the kernel function to use for the support vector machine. Possible values of it are: linear, polynomial, sigmoid and RBF. The degree parameter is used if a user has defined that the kernel type is the polynomial kernel function. The parameter gamma is used as an appropriate parameter in the kernel function and can be used as default of the one defined by a user.

Genetic algorithms have observably more parameters available for use than previously described algorithms. These parameters are the analyser, comparison factor lower and upper bound, comparison factor modifier, crossover, elites, maximum evaluated solutions, maximum generations, maximum selection pressure, mutation probability, mutator, offspring selection before mutation, population size, selected parents, selector and success ratio.

As GA has quite a lot of different parameters that can vary during experiments, several experiments are performed in order to determine what parameters are more important than others. In this case, changing parameters might produce very different results. An appropriate GA configuration provides convergence to the optimal result in limited time; while worse settings of its parameters might cause its long run required for finding a good solution. Moreover, results of changing multiple parameters are not predictable because in practice they are not completely independent of each other and might have effects on others [17].

C. Model Training

Within model training, different settings of algorithms are used to obtain more reliable models for the ensemble design. Training is performed using the software HeuristicLab [4]. For the heterogeneous ensemble design, four different methods, which have been previously described, are used. Each of three datasets is divided into two subsets: training set (70 % of data) and test set (30 % of data). One part of experiments is performed using tumour markers, and the other one is performed using blood parameters without tumour markers.

Training of base models with ANN is performed using one or two hidden layers according to the cross-validation results and a number of nodes equal to 5, 10 or 20 in each hidden layer. A decay parameter is set in the range from 0.1 to 5.

While training models using RF, each forest contains 25, 50 or 100 trees, M and R parameters are set in the range from 0.1 to 1.

Cross-validation of SVM algorithm has shown that the type of SVM that should be used for model training is N-support vector machine. For all three kinds of the cancer, RBF and polynomial kernel function are used where the parameter degree for the polynomial function is in the range from 2 to 5.

The following parameters of the GA with offspring selection during model training are changed: comparison factor lower bound, mutation probability, population size and selector. The lower bound of a comparison factor is set to 0 or to 1 that is used to determine whether a child should outperform both parents or not. Mutation probability values are set to 0.15, 0.2, 0.25, and 0.3. A population size is set in the range from 100 to 1000 with step 100. Proportional and gender specific selectors are used.

Best trained models are selected for each ensemble. Finally, 8 homogenous and 2 heterogeneous ensembles are designed for each kind of cancer – one half of them for cancer using tumour marker values as input variables and another half only using blood parameter values for model training. In the framework of the present research, homogeneous ensembles are used for the comparison analysis between them and heterogeneous ensemble [18].

V. ANALYSIS

Accuracy of the ensemble for each type of cancer with and without tumour markers is represented graphically using a histogram. Each vertical bar represents one of the ensembles designed in the present research; the first one represents accuracy of the heterogeneous ensemble, but the next four bars represent the accuracy of each homogeneous ensemble.

Figure 2 shows results of the ensemble design for breast cancer. The heterogeneous ensemble shows the highest accuracy of all designed ensembles that proves necessity of using heterogeneous ensemble for the same type of tasks.

In comparison with all homogeneous ensembles, the accuracy of the heterogeneous ensemble is at least 2 % higher and it is equal to 91.32 % applying tumour markers for model training and 87.85 % without tumour markers that is a high result and it proves a possibility of successful ensemble using for this type of cancer.

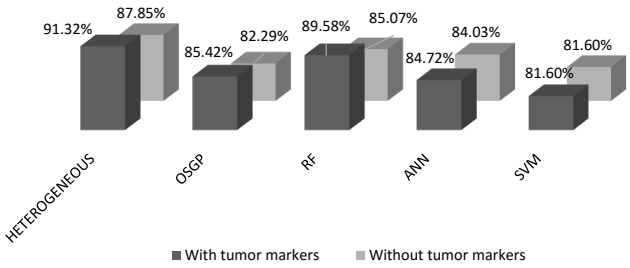

Fig. 2. Comparison of ensembles for breast cancer.

The obtained results for melanoma are quite good, although they are slightly worse than results of breast cancer. The best ensemble is heterogeneous ensemble and its accuracy is equal to 87.85 % (with tumour markers) and 85.48 % (without tumour markers). As for other types of cancer, heterogeneous ensemble shows the best results among all ensembles. In comparison with homogeneous ensembles designed in the present research, the test accuracy is at least 2 % higher. Figure 3 shows test accuracy of each ensemble.

Although heterogeneous ensembles show better results than homogeneous ensembles, it is important to note that all obtained results are satisfactory and can be used for the solving of the given problems.

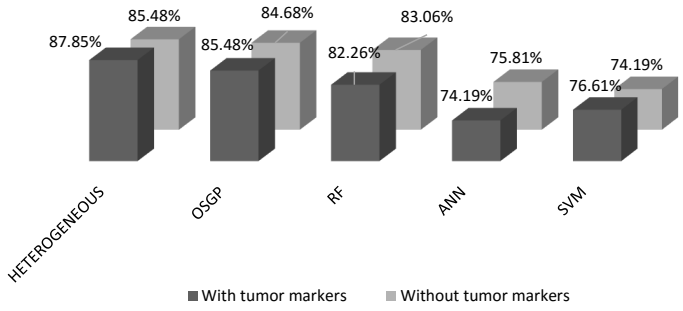

Fig. 3. Comparison of ensembles for melanoma.

Results of ensemble design for respiratory system cancer are not as good as results for breast cancer and melanoma, but can be considered acceptable. Figure 4 shows results of the ensemble design for the respiratory system cancer.

The accuracy of heterogeneous ensemble and ensemble that consists of the models trained by GA with offspring selection is the same and it is equal to 79.66 % in case of using tumour markers for the cancer prediction. The accuracy of all designed ensembles is not especially high and it ranges between 72 % and 80 % that is worse than results of melanoma and breast cancer. The accuracy of the best ensemble for the respiratory system cancer without tumour markers is 78.81 %. Heterogeneous ensemble and homogeneous ensemble that consists of models trained by ANN show the same results.

The explanation of worse results for this type of cancer can be scanty amount of available data for training.

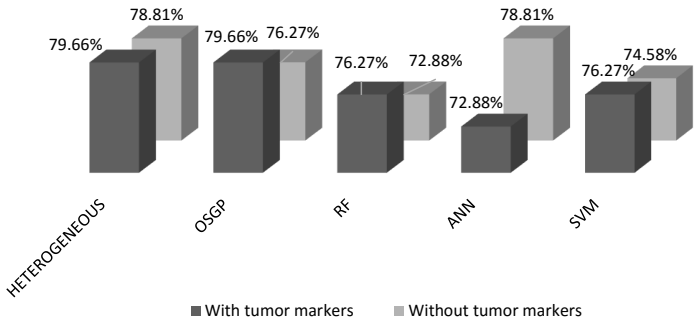

Fig. 4. Comparison of ensembles for respiratory system cancer.

The results obtained show that ensemble modelling for cancer prediction task is an appropriate approach, especially due to the fact that accuracy of ensembles that consist of models developed without using tumour markers is quite high. As it is known, tumour markers can be available only as a result of quite expensive examinations of patients. Ensembles that provide reliable predictions for cancer diagnostics without using tumour markers are extra important and their usage is capable of reducing considerably the costs of cancer diagnostics [18].

VI. Conclusion

This paper demonstrates the use of the ensemble for solving the classification task in the medical data mining area. The ensemble accuracy is higher than the accuracy of a single model, as well as an ensemble of models allows evaluating confidence of the prediction results. It can be done by using the confidence measure for each instance of the dataset. Due to the confidence measure application, a probability of estimating an error in the prediction process is appreciably higher than in case of using a single model. The ensemble usage allows solving different issues that can occur in case of using a single model. These issues are related to appropriate model selection, choice of the correct local minimum and impossibility of expanding the search space.

In the framework of the present research, ensemble modelling is considered to be a tool that can improve a procedure of cancer diagnostics. Problem of cancer diagnostics represents a two-class classification task, where each instance of the dataset should be classified as cancer positive or cancer negative.

The determination of appropriate settings of machine learning algorithms is based on the investigations known in literature about their influence on the results and cross-validation applied to determine more suitable settings of the neural networks, random forests and support vector machine. As the process of model training is stochastic, the same configuration of the algorithm was run ten times to obtain as more reliable models as possible.

Although previous studies of the ensemble modelling usage for such kind of problem are concentrated on the homogeneous ensemble design, the present research focuses on the appliance of the heterogeneous ensemble. The obtained results of the heterogeneous ensemble are compared with homogeneous ensembles. This comparison allows proving the assumption that predictions of the heterogeneous ensembles are more accurate and also more confident than the same predictions of the ensembles that consist only of one type models.

The accuracy of the designed ensembles is sufficient to conclude that ensemble modelling is a suitable tool for the given problem. Results show that the accuracy of the heterogeneous ensembles is higher or the same as the accuracy of homogeneous ensembles that allows recommending heterogeneous ensembles to get reliable predictions. In case of respiratory cancer, results are worse in comparison with the results of the ensembles for breast cancer and melanoma, although they are satisfactory and designed ensembles can also be used.

The obtained results show that ensemble modelling usage is more appropriate than single hypothesis usage. In addition to higher accuracy of predictions, the ensemble also provides higher values of the prediction confidence measure. Ensemble

provides a possibility of analysing the prediction with the help of the confidence measure. If the value of the confidence measure is insufficiently high, there is a possibility of checking the prediction additionally. It is especially helpful in case of cancer diagnostics, where the main task is to detect patients who are cancer positive. Further research in this direction can intend to investigate the application of other methods to improve quality of predictions and to obtain more reliable results.

REFERENCES

[1] Hennicker, R., Klarl, A., "Foundations for Ensemble Modelling – The Helena Approach," *Lecture Notes in Computer Science: Specification, Algebra, and Software*, 2014, vol. 8373, pp. 359–381.

[2] Zhou, Z.-H. *Ensemble Methods: Foundations and algorithms*. Boca Raton: Chapman and Hall/CRC, 2012. 236 p.

[3] Re, M., Valentini, G. A*dvances in Machine Learning and Data Mining for Astronomy*. Boca Raton: Chapman & Hall/CRC, 2012, 744 p.

[4] Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications. M. Affenzeller, S. Wagner, S. Winkler. Boca Raton: Chapman & Hall/CRC, 2009. 379 p.

[5] Affenzeller, M., Wagner, S. "Offspring Selection: A New Self-Adaptive Selection Scheme for Genetic Algorithms," in *Proc. of the Int. Conf. in Coimbra*, Portugal, 2005, pp. 218–221. http://dx.doi.org/10.1007/3-211-27389-1_52

[6] Cristianini, N., Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, 2000. 198 p. http://dx.doi.org/10.1017/CBO9780511801389

[7] Burges, C.J.C. "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, 1998, vol. 2, pp. 121–167. http://dx.doi.org/10.1023/A:1009715923555

[8] *SVM – Support Vector Machines* [Online]. Accessed 12th November 2015. Available: https://www.dtreg.com/solution/view/20

[9] Auria, L., Moro, R.A. "Support Vector Machines as a Technique for Solvency Analysis," *Discussion papers of German Institute for Economic Research*, 2008, 18 p.

[10] Cheung, V., Cannons, K. *An Introduction to Neural Networks* [Online]. Available: http://www2.econ.iastate.edu/tesfatsi/neuralnetworks.cheungcannonnotes.pdf. Accessed on: Nov. 10, 2015.

[11] Rogers, C. *The Advantages of Artificial Neural Networks* [Online]. Available: http://www.ehow.com/info_8148024_advantages-artificial-neural-networks.html. Accessed on: Nov. 10, 2015.

[12] Steinberg, D., Golovnya, M., Cardell, N.S. *A Brief Overview to Random Forests* [Online]. Available: http://nymetro.chapter.informs.org/prac_cor_pubs/RandomForest_SteinbergD.pdf. Accessed on: Nov. 10, 2015.

[13] A.L. Boulesteix, S. Janitza, J. Kruppa. "Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics," *WIREs Data Mining &KnowledgeDiscovery*, vol. 129, 2012, pp. 1–31.

[14] *Breast cancer description* [Online]. Available: http://www.cancer.gov/cancertopics/types/breast. Accessed on: Nov. 10, 2015.

[15] *Cancer Statistics* [Online]. Available: http://www.cancer.gov/statistics/. Accessed on: Nov. 10, 2015.

[16] *Description of Melanoma* [Online]. Available: http://www.skincancer.org/skin-cancer-information/melanoma. Accessed on: Nov. 10, 2015.

[17] Sarmady, S. "An Investigation on Genetic Algorithm Parameters, SiamakSarmady," School of Computer Science, Universiti Sains Malaysia. 2007, 10 p.

[18] Petrakova, A. *Uz mašīnapmācības metodēm balstīta heterogēnu modeļu ansambļa izveide*. Master thesis. Riga Technical University, Riga, 2014, 130 p.

Aleksandra Petrakova is a Doctoral Student at the Institute of Information Technology of Riga Technical University (Latvia). She obtained her Master's Degree in Information Technology from the above-mentioned institution in 2014. She works as MS Dynamics AX Functional Consultant at the Latvian information technology company ERP PRO. Her main research fields are ensemble modelling and its appliance together with simulation. She participated in Arena Student Simulation Competition Project "Gold Mining System" and in Research Cooperation project in cooperation with the Upper Austria University of Applied Sciences, School of Informatics, Communication and Media (Department of Software Engineering) and RTU (Department of Modelling Simulation).
E-mail: Aleksandra.Petrakova@edu.rtu.lv

Michael Affenzeller is Prof. (FH) Priv.-Doz. *Dipl.-Ing*. Dr., Head of Heuristic and Evolutionary Algorithms Laboratory, Upper Austria University of Applied Sciences. His professional interests include heuristic algorithms, evolutionary algorithms, algorithm theory and development, production planning and logistics optimisation, nonlinear system identification; structure identification, regression and time series, heuristic optimisation techniques in bioinformatics. He is the author of more than 220 publications.
E-mail: michael.affenzeller@heuristiclab.com

Galina Merkurjeva is *Dr. habil. sc. ing*., Full Professor at the Department of Modelling and Simulation of Riga Technical University. Her professional interests include methodology of discrete-event simulation, simulation metamodelling, simulation-based optimisation, decision support systems, logistics, production planning and control, supply chain management and simulation-based training. She is the author of more than 170 publications.
E-mail: Galina.Merkurjeva@rtu.lv