Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2020) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on Computer Science and Computational Intelligence 2020

Comparing Various Machine Learning Algorithms for Cardiovascular Disease Prediction Model

Britney[a], Michelle Belinda[a], Sunny Li Larosa[a], Elliana Lukmanto[a], Paulin[a], Bharuno Mahesworo[b,*], Haryono Soeparno[b,c]

[a]*Faculty of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[b]*Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia 11480*
[c]*Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

Abstract

Cardiovascular disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the world. With the number of deaths caused by cardiovascular disease twice as much as the second killer, cancer, it is clear that early detection system is needed to avoid premature death. Right now, many machine learning algorithms is used to help predict cardiovascular disease. However, the accuracy of those algorithms are still bellow the experts level. In this paper, we explored previous studies that used various machine learning algorithm to predict cardiovascular disease. The result of this paper can help us or other researcher who will develop better prediction model in the future.

Peer-review Statement: Peer-review under responsibility of the scientific committee of the 4th International Conference on Computer Science and Computational Intelligence 2020.

Keywords: Machine learning; Prediction; Classification; Cardiovascular.

1. Introduction

Since programming and computational algorithms became more related to our everyday life, programming languages have become more commonly studied in the early academic stage. These algorithms have made our life more convenient in many ways, such as better search engine, better product recommendations in e-commerce, and many others. In Computer Science, algorithms are sequences or steps used for calculation or to solve a problem written sequentially. Thus, programming algorithms are sequences or steps to solve computer programming problems. Nowadays, algorithm has evolve and becoming more data driven.

One of the most well-studied algorithms is a predictive algorithm. The algorithm or we also called machine learning, works by predicting the value of variables based on its previous values. It very often than it also have other

* Bharuno Mahesworo. Tel.: +62-822-2646-6536.
E-mail address: bharuno.mahesworo@binus.edu

Peer-review Statement: Peer-review under responsibility of the scientific committee of the 4th International Conference on Computer Science and Computational Intelligence 2020.

indicator variables. The use of these machine learning is not only for convenience in life but also have played important roles in our society, including predicting the future weather or predicting stock market prices. In the medical field, machine learning also can be used to predi ct the risk of getting certain diseases. It works by comparing the health condition of the person with significant factors that differentiate healthy and sick person. For example to predict the risk of getting a cardiovascular disease.

With 17.65 million death worldwide in 2016, double the deaths caused by cancer, cardiovascular disease is the main cause of death and the nemesis of mankind. Even though having a healthy lifestyle can reduce the risk of cardiovascular disease significantly, hearth attack often comes suddenly and cause death if not managed properly. Therefore, a routine check is needed when people have passed a certain age. Electrocardiogram or ECG is a common method to evaluate heart performance or detects cardiac abnormalities. ECG measures the electrical activity generated by the muscles of the heart cavities as it contracts. The ECG test result can be used to differentiate healthy and patient with cardiovascular disease.

A well-developed algorithm is expected to be able to classify healthy and patient with cardiovascular disease by only using the ECG test result. ECG test is cheap, widely available, fast and very convenient both for medical staff and patient. It is often used as an early screening method. However, some ECG marks can be also hard to spot even by experienced doctors. Thus, a reliable system without human error is needed to help medical staff interpret the ECG test result. This paper is a pilot study for our future research, developing a reliable system to predict cardiovascular diseases using ECG result and other relevant indicators. In this paper, we compare various machine learning for predicting cardiovascular disease. We also did a literature review for those machine learning to find the critical factors why some machine learning performed better compared to others. The result of this paper is a review of which machine learning is the most compatible for cardiovascular disease prediction using ECG result based on previous studies results. This study also creates critical opinions about to improve it in the future, according to the features of the machine learning.

2. Literature Review

2.1. Previous Study

This paper reviews several previous efforts on machine learning for predicting cardiovascular disease. The review includes how the machine learning is used, the data sets used, and the accuracy of those algorithms.

Swati and Priyadarshi [1] proposed a system that uses ECG result and other diagnostic results such as blood pressure and blood sugar level, as the data set, which then gets pre-processed and transformed. After that, the data mining algorithm, K-Nearest Neighbor (KNN) and Naïve Bayes are applied to the transformed dataset, producing the predicted heart disease output. The output is then tested for accuracy. From the results, the proposed system's accuracy for KNN and Naive Bayes algorithm is 76% and 84%. Thus, they concluded that Naïve Bayes was performed better than KNN.

Shylaja and Muralidharan [2] proposed a prediction system that uses a hybrid of Suppor Vector Machine (SVM) and Artificial Neural Network (ANN). The system trained using data set from Cleveland Clinic Foundation Heart disease dataset from the University of California, Irvine (UCI). The hybrid algorithm is applied to the pre-processed dataset. The accuracy was measured and compared with five other algorithms that also trained using the same dataset. The authors of this study stated that most of the previous study used 14 attributes from 76 attributes that are available from the Cleveland dataset. Meanwhile, this system uses only 11 from those 76 attributes. The hybrid algorithm works by separating the data into normal and cardiovascular data set using SVM then using the ANN to process the heart disease data set through different stages. The accuracy of the hybrid SVM-ANN is 88.54%. While the other five algorithms, RIPPER, Decision Support, Naïve Bayes, SVM, and ANN, are 81.08%, 79.05%, 82.97%, 85.30%, and 86.12% respectively. The results indicate that the proposed system has better accuracy compared to other algorithms.

Jabbar, Deekshatulu, and Chandra [3] proposed a method that uses the Random Forest algorithm for the prediction system. The data sets are obtained from corporate hospitals in Hyderabad (11 variables) and Statlog (heart) dataset (14 variables). To choose which attributes that influence the prediction, it uses chi-square and Genetic Algorithm (GA). The proposed algorithm works by loading the data set and ranking the features based on chi-square and GA value on descending order. Then, select the feature(s) with the highest value and apply the Random Forest algorithm to the rest of the data. While Random Forest without GA only has an accuracy of 80%, with a total of 100 trees used. The

result shows that the accuracy of Random Forest with GA is 82.96%. On the Hyderabad data set, the accuracy of the decision tree algorithm is 98.66%, while the proposed method gave 100% accuracy.

Mukherjee and Sharma[4] proposed a model that uses multi class ANN as the algorithm to predict the risk probability of cardiovascular disease. The dataset used in this study was obtained from Cleavelan UCI dataset data repository and from 76 attributes, 14 was chosen to be the attributes that will be trained. The proposed model consists of 10 input layer nodes, 10 hidden layer nodes, and 2 output layer nodes. The result of the experiment shows that the accuracy of the proposed model is 97%. Compared to the other algorithms, such as Naïve Bayes, KNN, and SVM, only have an accuracy of 73%, 70%, and 80% respectively. Based on comparing the accuracy of the model with other algorithms, the proposed method successfully surpass other algorithms performance.

Varun, Mounika, Sahoo, and Eswaran[5] proposed a predictive model for cardiovascular disease using Logistic Regression as the predictive algorithm. The model trained using a dataset that consists of 14 variables. The system consists of two main phases. The first phase is regularized cost function and the second phase is regularized gradient descent. The cost function is used for reducing errors from both the predicted and actual label while the gradient descent is used for calculating the coefficient until a minimum value for the class label is obtained. After the system is tested using several testing data sets, the result shows that the Logistic Regression gets an accuracy value of 87% and when compared with Naïve Bayes and Random Forest algorithm, both of them only get an accuracy value of 83.7% and 80%. Therefore, it can be said that the proposed system using Logistic Regression can predict heart disease more accurately compared to the other two algorithms.

Gomathi and Shanmugapriyaa[6] proposed a cardiovascular disease prediction model using data mining classification that uses three techniques: the Naïve Bayes, ANN, and the J48 Decision Tree Algorithms, to process data set consists of 210 records and uses Waikato Environment for Knowledge Analysis (WEKA) as data mining tool. It has four applications explorer, experimenter, knowledge flow, and simple CLI. The result shows some effective techniques that can be used for heart diseases classification and the accuracy are as follows. Naive Bayes has 79.9% accuracy and takes time 0.01 seconds to build the model, J48 has accuracy 77.0% and takes time 0.01 seconds to build the model, and ANN has 76.5% accuracy and takes time 1.55 seconds to build the model. The result shows that Naive Bayes is more accurate compared to others.

Shafique, Majeed, Qaiser, and Mustofa[7] proposed a cardiovascular disease predictive model that uses three algorithms: J48 Decision Tree, ANN, and Naïve Bayes. The experiments used cardiovascular disease dataset from UCI and filled the missing values using "ReplaceMissingValue" filter from WEKA tool, a variable selection data reduction technique. This technique reduces the size of the dataset by removing irrelevant or redundant attributes. The experiments show that the Naive Bayes algorithm has the highest accuracy, with the accuracy of 82.914%, while ANN got an accuracy of 79.89% and J48 has the lowest accuracy, 77.219%.

Hammad and Monkaresi[8] proposed a cardiovascular disease predictive model that uses four algorithms that are applied for diagnosing, including J48 Decision Tree, KNN, Naive Bayes, and SVM, which were implemented on Cleavelan UCI dataset and tested by using 10 fold cross-validation with the WEKA tool. The results shows that the accuracy of the used algorithm, J48 Decision Tree, KNN, Naïve Bayes, and SVM, are 78.54%, 80.85%, 87.45%, 83.82%. It is clear that the Naïve Bayes has the best performance.

Jain, Ahirwar, and Pandey[9] proposed a review on intuitive prediction of cardiovascular disease using various classifiers and algorithms, such as Decision Tree, Naïve Bayes, and SVM, used for expectation of cardiovascular illness. Based on the result of comparison between heart disease prediction system using data mining techniques, Support Vector Machine (SVM) technique has the highest accuracy with 84.44% and specificity 89.5%, followed by Naïve Bayes which shows high level performance after SVM technique with 83.66% accuracy and 86.41% specificity.

Patel, Upadhyay and Dr. S. Patel[10] proposed a cardiovascular disease prediction system using various algorithms such as J48 Decission Tree, Logistic Model Tree (LMT) and Random Forest, to evaluate the best algorithm for cardiovascular disease diagnosis. Based on the result, J48 Decission Tree achieved the highest sensitivity and accuracy, but LMT achieved higher specificity than J48 Decission Tree and Random Forest algorithm. They concluded that J48 Decission Treeis the best classifier for heart disease prediction because it have better accuracy and least total time to build the model with 56.76% accuracy and total time to build the model is 0.04 seconds, followed by LMT algorithm with 55.77% accuracy and total time to build the model 0.39 seconds.

The summary of used machine learning algorith and its accuracies from previous study can be seen in Table 1.

Table 1. Previous Studies Summary.

Author	Year	Algorithms	Accuracy
S. Swati and A. Priyadarshi	2015	KNN	76%
		Naïve Bayes	84%
U. Shafique, F. Majeed, H. Qaiser, and I. U. Mustafa	2015	Naïve Bayes	82.91%
		J48 Decision Tree	77.22%
		ANN	79.89%
J. Patel, T. Upadhyay, and S. Patel	2015	J48 Decision Tree	56.76%
		LMT	55.77%
M. A. Jabbar, B. L. Deekshatulu, and P. Chandra	2016	Decision Trees (Hyderabad)	98.66%
		Random Forest (Hyderabad)	100%
		Random Forest (without GA)	80%
		Random Forest (with GA)	82.96%
K.Gomathi and Dr. Shanmugapriyaa	2016	Naïve Bayes	79.90%
		J48 Decision Tree	77%
		ANN	76.50%
D. Hammad and H. Monkaresi	2017	Naïve Bayes	87.45%
		J48 Decision Tree	78.54%
		KNN	80.85%
		SVM	83.82%
S. A. Varun, G. Mounika, P. K. Sahoo, and K. Eswaran	2019	Logistic Regression	87%
		Naïve Bayes	83.70%
		Random Forest	80%
S. Shylaja and R. Muralidharan	2019	Hybrid SVM-ANN	88.54%
		RIPPER	81.08%
		Decision Support	79.05%
		Naïve Bayes	82.97%
		SVM	85.30%
		ANN	86.12%
S. Mukherjee and A. Sharma	2019	Multiclass ANN	97%
		Naïve Bayes	73%
		KNN	70%
		SVM	80%
A. Jain, M. Ahirwar, and R. Pandey	2019	SVM	84.44%
		Naïve Bayes	83.66%

2.2. Machine Learning Overview

Previous section has mentioned some of the machine learning that can be used for cardiovascular disease predictive model, such as Neural Network, Decision Tree, Naïve Bayes, KNN , SVM, and other. Most of the algorithms used are part of the classification or prediction model. Before discussing the performance of the algorithms from previous work, how these algorithms work, its strength and weakness, and its specialization need to be understood. The following are brief explanation of the used machine learning algorithm to predict cardiovascular disease from previous studies.

2.2.1. Artificial Neural Network (ANN)

ANN is one of the classification data mining techniques and works similarly to a human brain. The algorithm works as an interconnected network made up of tiny processors called "neurons"[11]. These neurons take input data sets and process them to produce suitable output. Each connection has a weight attached to it. This model is often used as regression and classification tools. With enough data, the accuracy of this model can be as good as experts in the related sector and without a human error.

2.2.2. Decision Tree

Decision Tree is one of the classification data mining techniques and one of the most popular algorithms for creating classifiers. Decision trees build classification or regression models in the form of tree structures[12]. This breaks the data set into smaller and smaller subsets while at the same time the related decision tree is developed in

stages. The result is a tree with decision nodes and leaf nodes. A decision node can have two or more branches. Leaf nodes are classifications or decisions.

2.2.3. Naïve Bayes

Naïve Bayes algorithm is based on the Bayes theorem with strong, naïve, independence assumptions[13]. This algorithm believes that the presence or absence of an attribute from a class is not related to the presence or absence of other attributes in the same class if the class' variable is given. The advantage of Naïve Bayes algorithm is that it makes the computation process easier and for large amounts of data, it increases the speed and also the accuracy. Another advantage is it allows adding new data at runtime.

2.2.4. K-Nearest Neighbour (KNN)

Another classification data mining technique is the KNN algorithm. This algorithm searches for the similarity between data as an individual and then groups them together. It calculates the distance of an individual data to every centroid to find which one is the nearest. This algorithm can also calculate values continuously for target data[14]. The average target value of the nearest neighbour is then used to get the predicted value of the new case.

2.2.5. Support Vector Machine (SVM)

SVM, which was introduced in 1995 by Vapnik, is a supervised learning model that can work for both classification and regression technique[15]. This algorithm works by dividing data into classes so that data that belongs to one class will be on one side and the rest on the other side. It then uses precision to generalize errors. Basic SVM divides data only into 2 classes whereas multiclass SVM executes the basic SVM multiple times on the training data to get multiple classes.

2.2.6. Logstic Model Tree (LMT)

LMT is a classification model that is a combination of Logistic Regression and Decision Tree. LMT consists of a decision tree structure with a logistic regression function at the leaves[16]. LMT uses cost-complexity pruning. This algorithm is slower compared to other algorithms because it regressed all classifiers in the same leave, which caused a lot of computing process.

2.2.7. Random Forest

Random Forest algorithm was proposed by Tin Kam Ho of Bell labs in 1995[17]. This algorithm combines a random selection of features to construct a decision tree with controlled variations. Every decision tree is made by randomly selecting data from the subset data. Random Forest algorithm can improve accuracy when most of the data are missing, it has methods for adjusting error in an unequal dataset.

3. Material and Methodology

This paper is a part of a bigger research project to create a cardiovascular disease predictive system and take part in the exploration phase. The existing models from various studies were gathered and compared to find which machine learning algorithm is the most promising for future development. When comparing these algorithms, we used two indicators, popularity and reliability. The popularity is represented by the frequency of those algorithms used in the previous studies. The popularity of a certain algorithm is an important consideration when developing a system. An algorithm with good popularity among researches indicates that the algorithm is well developed. Also, most of the time, popular algorithm have many tools for the researchers to use it.

The second indicator is reliability, which is represented by the average accuracy of the algorithm. This indicator is obviously needed when considering developing a system. This indicator can help us to find the less popular algorithms that could fit in predicting the cardiovascular disease problem. Besides these two indicators, further read about the mentioned algorithm is also needed to explored these algorithms specialization, weakness and strength, which can help us to create critical opinions about the discussed algorithm. The workflow of this study follows Figure 2.

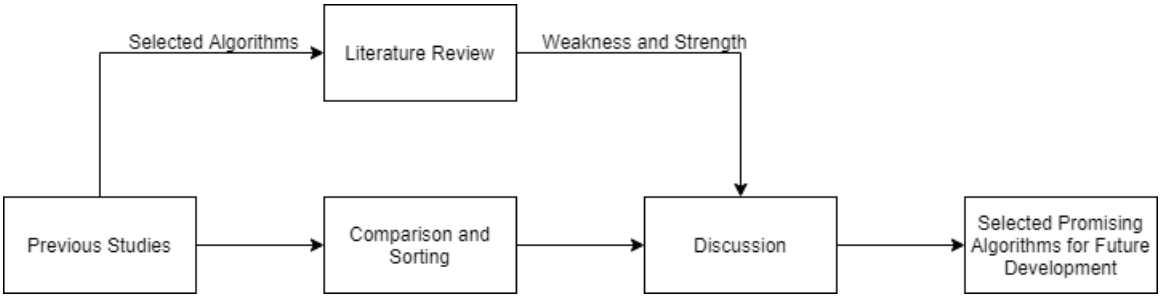

Fig. 1. Study workflow

4. Results and Discussion

Based on the reviews that have been mentioned above, we created a sorting mechanism to rank the used algorithm by looking at two indicators, which are the frequency of algorithm used on the reviewed previous studies and the average of the accuracy. The result of this sorting mechanism can be seen in Table 2.

Table 2. Sorting Result.

Algorithms	Frequency used	Accuracy
Naïve Bayes	8	82.20%
Decision Tree	5	77.64%
Random Forest	4	85.74%
ANN	4	84.88%
SVM	4	83.39%
KNN	3	75.62%
Hybrid SVM-ANN	1	88.54%
Logistic Regression	1	87.00%
RIPPER	1	81.08%
Decision Support	1	79.05%
LMT	1	55.77%

Based on our indicators, the LMT algorithm is ranked last in the list because it was used only once and has the lowest accuracy, which is only 55.77%. Then followed by Decision Support , RIPPER, Logistic Regression, and Hybrid SVM-ANN which all also used only once and has an accuracy of 79.05%, 81.08%, 87%, 88.54%. Afterwards, KNN was used three times and has an average accuracy of 75.61%. The Random Forest algorithm, ANN, and SVM was used four times and have average accuracy of 85.74%, 84.88% and 83.39% respectively. Then, the Decision Tree algorithm was used five times with an average accuracy of 77.64%. Latly, the most used algorithm in previous studies is Naïve Bayes, which is used 8 of 10 times, and has an average accuracy of 82.20%.

By looking at the frequency of the algorithms used in previous studies, we can see if the algorithm is popular among researchers or not. The more popular the algorithm, the more mature the algorithm is. We also averaging the accuracy of the algorithms among the studies to measure if the algorithm can give accurate results. According to these indicators, Naïve Bayes algorithm came as the best algorithm for predicting cardiovascular disease. The accuracy of Naïve Bayes from the reviews is 82.20% while the overall average from the 11 used algorithms is 81.21%.

In a technical perspective, the dataset trained to these models consist of many different data types (nominal, ordinal, ratio or interval; singular or series), the Naïve Bayes algorithm has advantages because it can easily process large amounts of data and increases its accuracy. At the same time, Naïve Bayes algorithm trained faster compared to other algorithms which made it favourable for researchers with limited infrastructure. Also, Naïve Bayes algorithm has a characteristic which due to their simplicity in allowing all of the variables contained in the training dataset to contribute towards the final decision equally and independently from the other variables. This simplicity equates to computational efficiency, which makes Naïve Bayes techniques attractive and suitable for many domains. Naïve Bayes works very well and has a higher level of accuracy than other classifier models [14].

However, because Naïve Bayes works independently among variables, this could reduce the level of accuracy if there is a possibility that these attributes are interrelated. But this might be solved if we want to increase the accuracy by enhancing the standard of Naïve Bayes, for example by hybridizing it with another algorithm. Collaboration with other techniques has been done by other research works as well, like SVM-ANN.

It is also worth considering the ANN algorithm for future development. ANN is a powerful technique for representing complex relationships between inputs and outputs. It works based on human neural behaviour. Several studies have been conducted to calculate the risk score of certain diseases using the ANN algorithm and succeeded in providing high accuracy results. However, it is also difficulty in understanding the reasoning behind the decision-making process of the ANN [14].

5. Conclusion

The fact that cardiovascular disease caused most deaths with double the amount of second killer, cancer, shows how important to create early detection for this disease. Currently, there are many data mining algorithms that are used to develop a model for heart disease prediction. However, the accuracies of these models are still bellow experts level. From the review that has been done, it can be concluded that Naïve Bayes is the most popular among researcher with promising results for the prediction. This algorithm is effective and efficient in term of training time and accuracy which make it promising for future development. Nevertheless, other algorithms also work well, depending on the data set used and its veracity. This paper reviews and discusses several previous studies that have been done by researchers to see how they use various algorithms to predict cardiovascular disease. We planned to use Naïve Bayes algorithm for future cardiovascular disease prediction development. We may also combine this algorithm with other algorithms since it is proven that the combined algorithms could produce a better result.

References

1. Swati, S., Priyadarshi, A.. Decision Support System on Prediction of Heart Disease Using Data Mining Techniques. *International Journal of Engineering Research and General Science* 2015;**3**(2):1453–1458. URL www.ijergs.org.
2. S.Shylaja, , Muralidharan, R.. Hybrid SVM-ANN Classifier is used for Heart Disease Prediction System. *International Journal of Engineering Development and Research* 2019;**7**(August):365–372.
3. Jabbar, M.A., Deekshatulu, B.L., Chandra, P.. Intelligent heart disease prediction system using random forest and evolutionary approach. *Journal of Network and Innovative Computing* 2016;**4**(1):175–184.
4. Mukherjee, S., Sharma, A.. Intelligent heart disease prediction using neural network. *International Journal of Recent Technology and Engineering* 2019;**7**(5):402–405.
5. Varun, S., Mounika, G., Sahoo, P., Eswaran, K.. Efficient System for Heart Disease Prediction by applying Logistic Regression. *International Journal of Computer Science and Technology (IJCST)* 2019;**10**(1):13–16.
6. Kamaraj, K.G., Priyaa, D.S.. Heart Disease Prediction Using Data Mining Classification. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* 2016;**4**(II):60–63.
7. Shafique, Umair and Majeed, Fiaz and Qaiser, Haseeb and Mustafa, I.U.. Data Mining in Healthcare for Heart Diseases. *International Journal of Innovation and Applied Studies* 2016;**10**(January):1312.
8. Hammad, D.S., Monkaresi, H.. Using Data Mining Techniques to Enhance Heart Disease Diagnosis Using Data Mining Techniques to Enhance Heart Disease Diagnosis. In: *Conference: Using Data Mining Techniques to Enhance Heart Disease Diagnosis*; September. Tehran, Iran; 2017, p. 0–6.
9. Jain, A., Ahirwar, M., Pandey, R., Jain, A., Ahirwar, M., Pandey, R., et al. A Review on Intutive Prediction of Heart Disease Using Data Mining Techniques To cite this version : HAL Id : hal-02265617 International Journal of Computer Sciences and Engineering Open Access A Review on Intutive Prediction of Heart Disease Using Data M. *International Journal of Computer Sciences and Engineering* 2019;**7**:109–113.
10. Patel, J., Upadhyay, T., Patel, S.. Heart Disease prediction using Machine learning and Data Mining Technique. In: *Conference: Journal - International Journal of Computer Science & Communication (IJCSC)*. 2015, p. 129–137.
11. B., Y.. *Artificial Neural Network*. New-Delhi: Asoke K. Ghosh, Prentice-Hall of India Private Limited; 1999. ISBN 81-203-1253-8.
12. Quinlan, J.R.. Induction of decision trees. *Machine Learning* 1986;**1**(1):81–106. doi:\bibinfo{doi}{10.1007/bf00116251}.
13. I., R.. An empirical study of the naive Bayes classifier. *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence* 2001;**3**(22):41–46.
14. Peterson, L.E.. K-nearest neighbor. *Scholarpedia* 2009;**4**(2):1883. doi:\bibinfo{doi}{10.4249/scholarpedia.1883}. Revision #137311.
15. Noble, W.S.. What is a support vector machine? *Nature Biotechnology* 2006;**24**(12):1565–1567. doi:\bibinfo{doi}{10.1038/nbt1206-1565}. URL https://doi.org/10.1038/nbt1206-1565.
16. Sumner, M., Frank, E., Hall, M.. Speeding up Logistic Model Tree induction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2005;**3721 LNAI**:675–683. doi:\bibinfo{doi}{10.1007/11564126_72}.

17. Breiman, L.. Random Forests. *Machine Learning* 2001;**45**(1):5–32. doi:\bibinfo{doi}{10.1023/A:1010933404324}. URL https://doi.org/10.1023/A:1010933404324.