Idea paper

Comparison of Random Forest Classifier and MLP Classifier

By Aleksander Vanberg Eriksen

18.11.2022

Scientific methods and theory

Masters in applied computer science

Østfold University College

Table of contents

1. Introduction

Following visible successes on a wide range of predictive tasks, machine learning techniques are attracting substantial interest from medical researchers and clinicians [1]. These techniques can be used to analyze data characterizing patients, through their categorization, to the creation of advanced predictive models determining the probability of a patient suffering from a specific disease [2]. Artificial Intelligence (AI) improves diagnostics and increases safety thanks to algorithms that can interpret data from various sources and devices in real-time [2].

This study will compare two machine learning techniques called multi-layer perceptron classifier and random forest classifier and will be tested on a medical dataset about heart failure.

2. Research Questions

The research question that is raised is based on gaining information about how it works, and the difference between the two types.

RQ1: How does a neural network work when solving classification problems?

- RQ1a: How does MLP classifier work?
- RQ1b: How does Random Forest Classifier work?

RQ2: What's the difference in performance when solving a classification problem using MLP classifier and Random Forest Classifier?

3. Literature Review

3.1. What are MLP classifiers, and how do they work?

Multilayer perceptron's are powerful classifiers that contains several free parameters [3, 4]. Even though they may provide superior performance, they need Cross validation which can be time-consuming and biased [3, 4]. Slow convergence and lack of guarantee of global minima are further drawbacks [3]. Still, they are commonly used feed forward network as classifiers for pattern classification approaches [5].

The parameters are number of hidden layers, number of hidden nodes, activation function, solver function and learning rate [6,7]. The accuracy of the MLP will then depend on the selection of these parameters which can require much effort in gaining the best architecture [6, 7]. The way that the classifier is set up is that is has three types of layers [8]. The first layer is called the input layer and is where the model receives the data [8]. Next layer is called the hidden layer and consists of neurons that map the data with mathematical functions [8]. The MLP can have more than one hidden layer [8]. The last layer is called the output layer which receives the data computed by the hidden layer and returns a result [8]. Each neuron has a weight attached which its purpose is for strengthening the linkage [8]. The neurons that are in the hidden layer and output layer also has a bias that works like a threshold for the activation of the neuron [8].

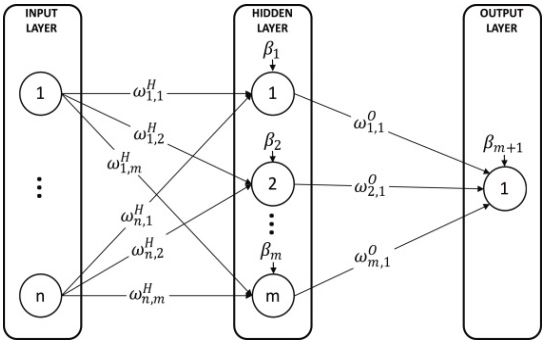

Figure 1: Example of a structure of the MLP classifier

To achieve an acceptable output value, the MLP classifier is trained with a method called Back Propagation [8]. Back Propagation adjusts the weights and biases based on the output value the model gets, until an acceptable error is reached [8]. Another technique you can use to improve the model, is by preventing overtraining [9]. Early stopping is one of these methods, and it works so that the training will stop when the generalization error begins to increase [9].

3.2. What are Random Forest classifiers, and how do they work?

The Random Forest classifiers is of the most popular machine learning techniques [10]. It is created by Breiman and presents many advantages for its application in remote sensing [11]. It performs several tasks which rises the accuracy of training and testing speed [12]. The classifier generates forests [13]. These forests consist of a huge number of unique trees [12,14]. These trees are known as decision trees, and each tree is divided into a class prediction [12]. The class with the most votes become the model's prediction [11, 12, 14].

The design of a decision tree requires the choice of an attribute selection measure and a pruning method [11, 14]. The most frequently used attribute selection is the information Gain Ratio criterion, Gini index and Chi-square [11, 14]. The classifier can be used by splitting a dataset for training, and one for testing [10]. The training set is used to train the classifier, while the testing set is used to simulate the model in terms of prediction [10]. The RF classifier only needs the definition of two parameters for generating a prediction model: the number of trees desired, and the number of prediction variables [11]. When increasing the number of trees, the generalization error always converges, and over-training is not a problem [11]. If you would decrease the number of predicted variables, it would cause each individual tree of the model to be less strong but also have less correlation with each other which makes the model more accurate [11]. Further hyperparameter tuning is used to improve the model [10]. To reduce the possibility of over-fitting the training data, you can restrict the depth of any single trees and have a maximum number of trees. [15]. This is done if the data you are working with is small [15]. The model can either utilize bootstrap aggregating or bagging techniques to learn decision trees [7, 16]. Bootstrap aggregating is a technique used for creation of training data by randomly resampling the original dataset with replacement [11]. The technique offers a strong potential in medical domains [15].

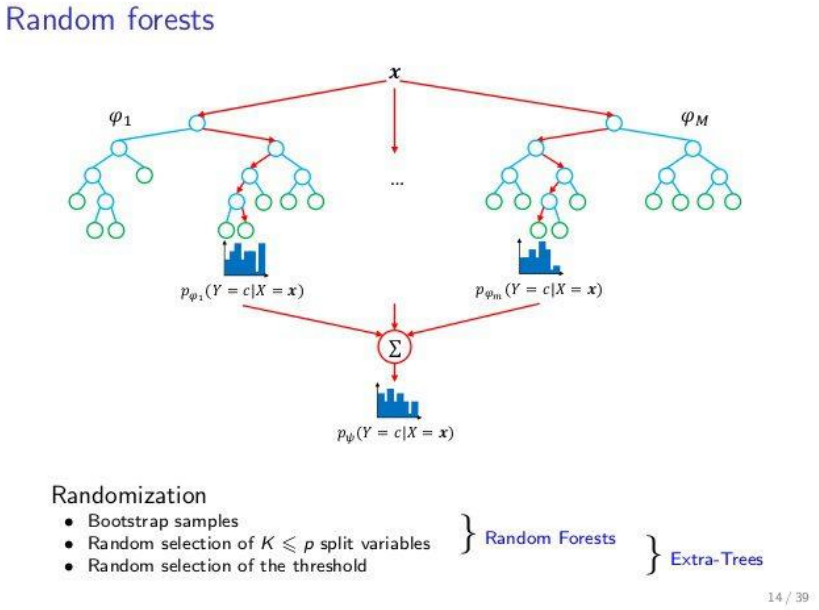

Figure 2: Example of a random forest structure

3.3. What's the difference in performance when solving a classification problem using MLP classifier and Random Forest Classifier?

A study by A. Nasim, D. C. Nchekwube, F. Munir and Y. S. Kim about Arrhythmia Classification with Optimum Features Using Single-Lead Electrocardiogram used an optimization called MOPOS on MLP, k-nearest neighbor, support vector machine, Random Forest, and extra decision tree. This resulted in that the MLP consistently preformed the best with significantly reduced number of features for the targeted 15-class classification problem. The MLP classifier used optimizer Adam with a learning rate of 0.0001, activation function ReLU, Input size of 253, Output size of 15, and hidden layer sizes of 220, 180, 120 and 60. The random forest model consisted of 25 trees with an Optimum split criterion of entropy [17].

Rachael Hagan, Charles J. Gillan and Fiona Mallett conducted a study about comparing machine learning methods for classification of cardiovascular disease. This study used two public datasets, UCI arrythmia dataset and Kaggle cardiovascular dataset, with significantly different characteristics to assess the potential differences in the uncertainty of the methods. For this study, the random forest classifier uses random subset of data and adds random selection of the available features. It had Gini criterion, 100 estimators and a max depth of 20 for the UCI dataset and Gini criterion, 100 estimators and max depth 10 for the Kaggle dataset. The MLP classifier model consisted of a tenfold cross validation for hyperparameter search. The model had an input layer of 15 nodes with 15 output nodes, a hidden layer with 15 input nodes and 50 output nodes, a drop out of 50 input nodes and 50 output nodes, another hidden layer with 50 input nodes and 25 output nodes, and an output layer of 25 input nodes and 2 output nodes. They got the best accuracy by using batch size 32 with 400 epochs, and Nadam optimizer function. The MLP classifier resulted in an accuracy of 0.74 for the UCI arrythmia dataset and 0.705 for the Kaggle cardiovascular dataset. The random forest classifier resulted in an accuracy of 0.95 for the UCI arrythmia dataset and an accuracy of 0.74 for the Kaggle cardiovascular dataset. This resulted in that the random forest classifier preformed the best [18].

Ashfaq Ahmad Najar & S. Manohar Naik conducted a study that was about detecting DDoS attack using MLP and Random Forest Algorithms. The random forest model was used for binary classification. The model checked if either a packet was normal or an attack packet, and got an accuracy of 0.9913 on train data, 0.9913 on validation data and an accuracy of 0.97 on test data with an F1 score of 0.9661. The MLP model was used to detect what type of attack it was, and got an accuracy of .9796 on train, 0.9853 on validation data and 0.74 on the test data [19].

A study by A. J. Lado et al were comparing Neural Network and Random Forest Classifier on Dragon Fruit Disease. This study was an image classification problem and contained 41 images of healthy and sick fruit and leaf which was divided into four classes. The models used a 10-fold cross validation, and both obtained an accuracy in range of 70 to 82 % [7].

4. Methodology

To further research and answer our problem statement, our methodology will consist of using a dataset about heart failure from Kaggle in a python program [20]. This dataset consists of information about certain patients, and if they have a heart disease or not. It does not contain data that can point towards the identity of the patients. The dataset is a combination of five other dataset and claims to be the largest heart disease dataset for research purposes available. The data comes from Cleveland, Hungarian, Switzerland, Long Beach VA and Stalog. Those who created the individual datasets are Andras Janosi from Hungarian Institute of Cardiology, William Steinbrunn from University Hospital, Zurich, Matthias Pfisterer from University Hospital, Basel and Robert Detrano from V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

4.1. Analyzing the dataset

The dataset is a csv file, so the first technique that is done is passing the dataset into a data frame *df* using a library called pandas. By creating the data frame *df*, some functions will be available which are used to further analyzing the dataset. The first that is looked at is what type of data does the dataset contain. Function *df.info()* shows us the name of the columns in the dataset and what type of value they consist of.

Categorical data	Numerical data
Sex	Age
Chest Pain Type	Resting BP
RestingECG	Cholesterol
ExerciseAngina	FastingBS
ST_Slope	MaxHR
	Oldpeak
	HeartDisease

Table 3: Describing what type of data we are dealing with

The dataset contains 11 columns with different values for each column. The columns that contain non-numerical values are under Categorical data in Table 3.

4.2. Preprocessing the data

Categorical data is converted into numerical data using the label encoder function from sklearn.preprocessing library which ensures the model that it can learn from the data [21, 22, 23]. Then a data split is done using train_valid_test_split function from fast_ml.model_development library to have a batch of training data, testing data, and validation data. The training, testing and validation data is split into an X and Y batch, where Y contains the HeartDisease column and X contains the rest. This is to check if the model predicted correct by checking the true value in Y and compare it to the produced prediction. A preprocessing method called Normalizer is then used on the training data to make sure the numerical data don't have much deviation which makes the model have an easier time to learn [23]. The normalizer method is a function from the sklearn.preprocessing library.

4.3. Evaluating the classifiers

This section contains step by step how the approached evaluation of the Random Forest classifier, and the MLP classifier were conducted.

4.3.1. The Random Forest Classifier Model

The random forest classifier model used was imported from sklearn.ensemble and was set up using 200 trees, max depth of 10 and default parameters for the rest. The model first ran on the training data, and then on the testing data. A score were calculated based on the difference between the predicted value and the correct values to see how much accuracy the model has. The model landed on a score of 0.862. This means that the model predicts correct 86.2 % of the times.

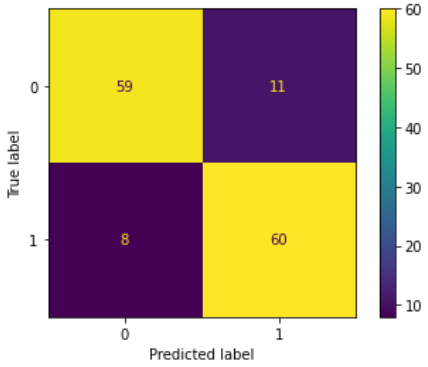

Figure 3: A confusion matrix generated by sklearn.metrics showing the predicted values

Figure 3 shows the result of the model's prediction. The yellow squares show the number of correct predictions, while the purple squares show the wrong predictions.

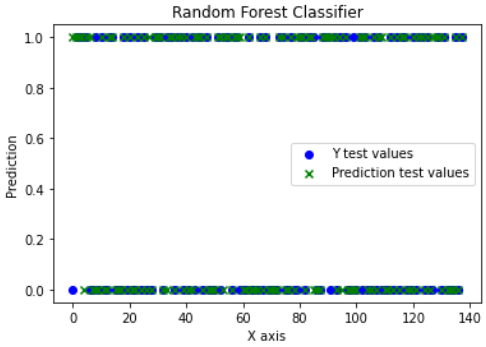

Fig 4: A graph containing the prediction from the random forest classifier compared to the actual values

For the last test, a graph containing the prediction and the true values is shown.

4.3.2. The MLP Classifier Model

The MLP classifier model that was used where imported from sklearn.neural_network and was set up using 200 max iterations, and default parameters for the rest. The model first trained on the training data X and Y, and later were tested on the testing data. The predicted values were then compared to the true values by a score, and an accuracy where set based on that. The model got a score of 0.739, which means that it got an accuracy 73.9 %.

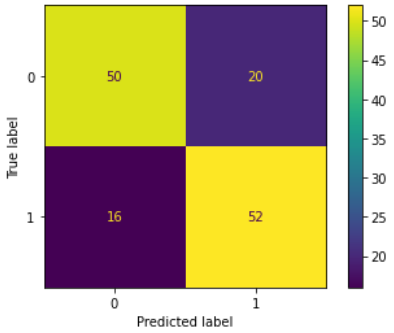

Figure 3: A confusion matrix generated by sklearn.metrics showing the predicted values

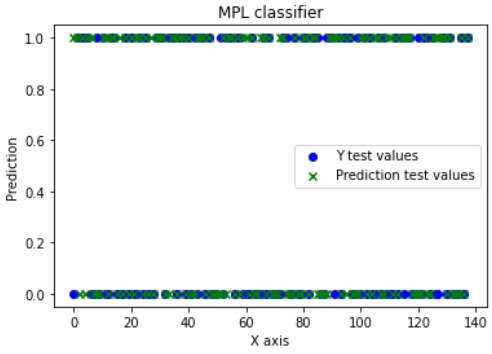

Figure 4: A graph containing the prediction from the MLP classifier compared to the actual values

5. Summary

Based on the findings from the literature review and the methodology used, the Random Forest Classifier might perform better than the MLP Classifier in medical domains [15]. The Random Forest Classifier uses a lot of decision trees to go through different approaches to achieve the most optimal model by utilizing the bagging method, while the MLP Classifier is heavily based on hyperparameter tuning [6, 7, 15].

6. References

[1] Sidey-Gibbons, J., Sidey-Gibbons, C. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* **19**, 64 (2019).https://doi.org/10.1186/s12874-019-06814

[2] Beata Butryn, Iwona Chomiak-Orsa, Krzysztof Hauke, Maciej Pondel, Agnieszka Siennicka, Application of Machine Learning in medical data analysis illustrated with an example of association rules, Procedia Computer Science, Volume 192, 2021, Pages 3134-3143, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2021.09.086.

[3] T. Windeatt, "Accuracy/Diversity and Ensemble MLP Classifier Design," in IEEE Transactions on Neural Networks, vol. 17, no. 5, pp. 1194-1211, Sept. 2006, https://doi.org/10.1109/TNN.2006.875979.

[4] Windeatt, T. (2008). Ensemble MLP Classifier Design. In: Jain, L.C., Sato-Ilic, M., Virvou, M., Tsihrintzis, G.A., Balas, V.E., Abeynayake, C. (eds) Computational Intelligence Paradigms. Studies in Computational Intelligence, vol 137. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79474-5_6

[5] X. Zhai, A. A. S. Ali, A. Amira and F. Bensaali, "MLP Neural Network Based Gas Classification System on Zynq SoC," in IEEE Access, vol. 4, pp. 8138-8146, 2016, https://doi.org/10.1109/ACCESS.2016.2619181.

[6] M. Roy, D. Routaray, S. Ghosh and A. Ghosh, "Ensemble of Multilayer Perceptrons for Change Detection in Remotely Sensed Images," in IEEE Geoscience and Remote Sensing Letters, vol. 11, no. 1, pp. 49-53, Jan. 2014, https://doi.org/10.1109/LGRS.2013.2245855.

[7] A. J. Lado et al., "Comparison of Neural Network and Random Forest Classifier Performance on Dragon Fruit Disease," 2021 International Electronics Symposium (IES), 2021, pp. 287-291, https://doi.org/10.1109/IES53407.2021.9593992

[8] Matías Gabriel Rojas, Ana Carolina Olivera, Pablo Javier Vidal, Optimising Multilayer Perceptron weights and biases through a Cellular Genetic Algorithm for medical data classification,Array,Volume 14,2022,100173,ISSN 2590-0056, https://doi.org/10.1016/j.array.2022.100173.

[9] Medeiros, C.M.S., Barreto, G.A. A novel weight pruning method for MLP classifiers based on the MAXCORE principle. Neural Comput & Applic 22, 71–84 (2013). https://doi.org/10.1007/s00521-011-0748-6

[10] N. A. Maung Maung, B. Y. Lwi and S. Thida, "An Enhanced RSS Fingerprinting-based Wireless Indoor Positioning using Random Forest Classifier," 2020 International Conference on Advanced Information Technologies (ICAIT), 2020, pp. 59-63, https://doi.org/10.1109/ICAIT51105.2020.9261776.

[11] V.F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, J.P. Rigol-Sanchez, An assessment of the effectiveness of a random forest classifier for land-cover classification, ISPRS Journal of Photogrammetry and Remote Sensing,Volume 67,2012,Pages 93-104,ISSN 0924-2716, https://doi.org/10.1016/j.isprsjprs.2011.11.002. (https://www.sciencedirect.com/science/article/pii/S0924271611001304)

[12] V. Geetha, A. Punitha, M. Abarna, M. Akshaya, S. Illakiya and A. P. Janani, "An Effective Crop Prediction Using Random Forest Algorithm," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020, pp. 1-5, https://doi.org/10.1109/ICSCAN49426.2020.9262311.

[13] Damodar Reddy Edla, Kunal Mangalorekar, Gauri Dhavalikar, Shubham Dodia, Classification of EEG data for human mental state analysis using Random Forest Classifier,Procedia Computer Science,Volume 132,2018,Pages 1523-1532,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.116.

[14] M. Pal (2005) Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26:1, 217-222, https://doi.org/10.1080/01431160412331269698

[15] Klement, W., Gilbert, S., Resende, V.F. et al. The validation of chest tube management after lung resection surgery using a random forest classifier. Int J Data Sci Anal 13, 251–263 (2022). https://doi.org/10.1007/s41060-021-00296-8

[16] Khorshidpour, Z., Hashemi, S. & Hamzeh, A. Evaluation of random forest classifier in security domain. Appl Intell 47, 558–569 (2017). https://doi.org/10.1007/s10489-017-0907-2

[17] A. Nasim, D. C. Nchekwube, F. Munir and Y. S. Kim, "An Evolutionary-Neural Mechanism for Arrhythmia Classification With Optimum Features Using Single-Lead Electrocardiogram," in IEEE Access, vol. 10, pp. 99050-99065, 2022, https://doi.org/10.1109/ACCESS.2022.3203586.

[18] Rachael Hagan, Charles J. Gillan, Fiona Mallett, Comparison of machine learning methods for the classification of cardiovascular disease, Informatics in Medicine Unlocked,Volume 24,2021,100606,ISSN 2352-9148, https://doi.org/10.1016/j.imu.2021.100606.

[19] Najar, A.A., Manohar Naik, S. DDoS attack detection using MLP and Random Forest Algorithms. Int. j. inf. tecnol. 14, 2317–2327 (2022). https://doi.org/10.1007/s41870-022-01003-x

[20] fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [14.11.2022] from https://www.kaggle.com/fedesoriano/heart-failure-prediction.

[21] K. P. N. V. Satya Sree, J. Karthik, C. Niharika, P. V. V. S. Srinivas, N. Ravinder and C. Prasad, "Optimized Conversion of Categorical and Numerical Features in Machine Learning Models," *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2021, pp. 294-299, doi: 10.1109/I-SMAC52330.2021.9640967.

[22] A. Gehlot, N. Sidana, D. Jawale, N. Jain, B. P. Singh and B. Singh, "Technical analysis of crop production prediction using Machine Learning and Deep Learning Algorithms," *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, 2022, pp. 1-5, doi: 10.1109/ICSES55317.2022.9914206.

[23] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018, pp. 98-103, doi: 10.1109/ICACCCN.2018.8748777.

Figures

[1] Example of structure of a Multilayer Perceptron (MLP)

https://doi.org/10.1016/j.array.2022.100173

[2] Rishabh Mall, Jan 7, 2019, rand-forest-2.jpg

https://medium.com/@mallrishabh52/random-forest-67afc2ff884f

https://www.kdnuggets.com/wp-content/uploads/rand-forest-2.jpg

[3] Generated confusion_matrix plot by using sklearn.metrics. plot_confusion_matrix

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.plot_confusion_matrix.html

[4] Generated graph using matplotlib.pyplot import

https://matplotlib.org/stable/tutorials/introductory/pyplot.html

Tables

[1] Table 1: Overview of articles found

[2] Table 2: Overview of articles containing research questions

[3] Table 3: Overview of which datapoint in the dataset is of what type

7. Appendix SLR

3. Systematic Review Protocol

To be able to fully answer my research questions, a SLR is done to gather relevant research articles that contain satisfying information. These articles will contain information about how a neural network functions in practice, and if there are any difference between the two distinct one.

3.1. Search strategy

This section contains the strategies I used to obtain the relevant research articles that contains information about the research questions.

3.2. Databases

Utilizing some known databases to find articles and a dataset.

- Kaggle: https://www.kaggle.com/
- Science direct: https://www.sciencedirect.com/
- ACM Digital: https://dl.acm.org/
- IEEE Xplore Digital Library: https://ieeexplore.ieee.org/
- Springer Link: https://link.springer.com/
- Google scholar: https://scholar.google.com/

3.3. Search term

A good search terms is important to have figured out so that it can get articles based on the research questions. My search terms tried to get most out of the research questions without excluding each other.
(Neural network AND Random Forest classifier AND MLP classifier) OR Random Forest Classifier OR MLP Classifier

3.4. Search procedure

All databases used the search terms. This section will give information about what kind of filters I applied when searching using the filter function.
Science direct: I went for only Research articles for the article type, Computer science for subject area.

IEE: I applied filters to only look for journals and keywords learning (artificial intelligence) and neural nets. Sorted by most cited.
Springer Link: I used the filter to only look for English articles containing artificial intelligence and computer science.
ACM Digital: Filters to see only research articles in pdf form.
Google scholar: Sorted by relevance

4. Study selection

When choosing the studies, I looked for
- Studies that used Random Forest classifier
- Studies that used MLP classifier
- Open access
Excluded articles
- No information about classifiers
- Articles that were in other languages than English

4.1. Data extraction

The technique I mostly used to find the most optimal studies were as followed:

- The title of the paper
- Main topic of the paper
- The abstract of the paper
- Passing the study selection requirements

4.2. Execution of the systematic literature review

Database	Articles where I applied the search terms	Articles that passed the data extraction	Articles used
Science Direct	3431	9	4
IEEE Xplore	451	9	7
Springer link	4593	5	5
ACM Digital	200k +	3	0
Google scholar	100k +	1	1
Total		27	17

Table 1. Articles that were found, checked, and used

5. Results

Research questions	Publications
RQ1a	[3] [4] [5] [6] [7] [8] [9]
RQ1b	[7] [10] [11] [12] [13] [14] [15] [16]
RQ2	[7] [17] [18] [19]

Table 2. Overview of Articles that contained research questions.

This section contains an overview of the references which answers the research questions.