Justus-Liebig-Universität Gießen
Chair of Statistics and Econometrics
Prof. Dr. Peter Winker

Master Seminar
Summer Term 2020

Topic:
„Evaluating the accuracy of cash flow forecasting methods“

Tutored by:
Dr. Frauke Schleer van Gellecom and Jenny Bethäuser

Submitted by:
Daniel Kinkel
Diezstraße 7
35390 Gießen
Tel.: 0177 5402264
E-Mail: Daniel.Kinkel@wirtschaft.uni-giessen.de
Enrolment number: 6005391
Study program: Business Administration (MSc)
Semester: 2

Table of contents

1. Introduction

The forecasting of future cash flows (CF) has valuable benefits in a variety of areas. For instance, in one of the first studies regarding this topic, Bowen, Burgstahler and Daley (1986: 714) identified distress prediction, risk assessment with respect to the size and timing of business loans, predicting credit ratings, valuing closely-held companies and provision of incremental information to security markets as possible areas of application for medium to long-term CF estimates.

Another application especially for the forecasting of daily CFs is the area of cash flow management (da Costa Moraes, Nagano and Sobreiro 2015: 12). In literature different models on how to manage the cash balance have been developed. The first one dating back to 1952 was proposed by Baumol who modeled the problem as a tradeoff between opportunity cost of short-term investments and transfer cost similar. However, this model depends on the assumption of a fixed and predictable demand. Miller and Orr (1966) rejected this assumption and introduced a model that tried to solve the optimization problem on the basis of random, normally distributed cash flows. This stochastic approach dominated research for some time until Gormley and Meade (2007) first developed a model using daily cash flow forecasts as the main input (da Costa et al. 2015: 12-18). For this purpose, they conducted an autoregressive model that is used to forecast a variable based on its past observations. This however, is not the only method that can be used to obtain daily CF forecasts. Stone and Wood (1977) for example used a regression model with an additive specification in order to reflect the day-of-month and day-of-week effects with another dummy variable for holiday effects. Further regression models were used in the studies of Miller and Stone (1985) and Stone and Miller (1987). In contrast to the first study these studies also considered multiplicative specifications to account for the possibility that the day-of-month and day-of-week effects might not be structurally independent (Stone and Miller 1985: 338).

Both mentioned approaches belong to the linear models, which depend on the assumption of normality. However, as has been shown in Salas-Molina, Rodriguez-Aguilar, Serra and Martin (2016), this assumption rarely holds even after data transformation. Therefore, the authors argue that non-linear models like random forest models and neural networks might be a suitable alternative that could provide better results (Salas-Molina et al. 2016: 3). Following this line of thought, Salas-Molina, Martin, Rodriguez-Aquilar, Srra and Arcos (2017) show at the example of two real datasets that a random forest and a radial basis function network might improve the accuracy of daily

CF forecasts, even though the results vary for the different datasets. Furthermore, the study of Dadteev, Shchukin and Nemeshaev (2018) on daily cash flow forecasting in commercial banks provides an additional case in which an artificial neural network could in some cases increase the predictive accuracy of the forecasts. In this paper a multilayer perceptron (MLP) was trained and compared to both an autoregressive and a regression model.

Achieving higher predictive accuracy has strong implications for the utility of CF forecasts in cash management (Salas-Molina et al. 2017: 414). Concrete reasons for this are that good daily cash flow forecasts are able to improve profitability of short-term investments and reduce the cost of short-term borrowing and idle cash balances (Stone and Miller 1987: 45). The potential for cost savings is especially high in the presence of high cash flow volatility and riskier cash management policies (Salas-Molina et al. 2017: 413).

In order to evaluate their predictive accuracy an autoregressive model, a regression model and a random forest models will be created for each dataset [Section 3]. Subsequently with the help of time series cross validation their forecasts will be analyzed regarding the accuracy for different prediction horizons [Section 4]. Before that the next chapter will focus on the underlying dataset.

2. Data description and transformation

The two selected datasets are part of the datasets from the study of Salas-Molina et al. (2016) that contained daily cash flows from 54 small and medium sized companies in Spain. Additionally, to the daily cash flows each dataset provides information about the date, the day of the month and the day of the week at which the cash flow occurred and a binary variable for holidays. In the further course the datasets of company 3 and 8 are referred to as “dataset1” and “dataset2”. The statistical properties of their cash-flows are summarized in Table 1.

Table 1: Data sets statistical summary. Mean, standard deviation, minimum, maximum in thousands of €

	Length	Mean	Std	Min	Max	Skewness	Kurtosis	Null %
dataset1	1,247	-0.56	35.85	-663.15	671.04	-1.14	200.74	25.02
dataset2	1,017	-0.01	2.81	-24.21	11.31	-3.36	23.07	22.42

The last column describes that a large proportion of the cash flows in both datasets is equal to zero. This also leads to the observation of high positive excess kurtosis and arises the question if the daily cash flows are normally distributed. In order to assess this, the Shapiro-Walk test (Royston 1982) is performed in the statistics computing software *R*. Both tests result in p-values way below 0.05, which is why the null hypothesis of normally distributed cash flows has to be rejected. As stated by Salas-Molina et al. (2016: 6) this might be problematic for the prediction quality of linear models if this results in non-normally distributed residuals. Therefore, outlier correction and a Box and Cox (1964) transformation are considered as possible solutions for this issue (Salas-Molina et al. 2016: 6). The outlier correction is processed by identifying and removing cash flow values that are greater than 5 times the standard deviation. While this might result in an information loss, Gormley and Meade (2007) argue that cash managers are generally aware of extremely large daily cash flows (Gormley and Meade 2007: 6). Since the objective of daily cash flow forecasting is not to forecast already known cash flows, it seems reasonable to remove these outliers, as they would have a significant influence on the model identification and thus the estimation of the unknown cash flows. Additionally, the two parameter Box-Cox transformation that is also suitable for negative cash flows is used to approximate the cash flows to a Gaussian distribution:

$$y^{(\lambda)} = \begin{cases} \dfrac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda} & \text{if} \quad \lambda_1 \neq 0 \\ \log(y + \lambda_2) & \text{if} \quad \lambda_1 = 0 \end{cases} \tag{1}$$

Here the *R* package "geoR" of Ribeiro Jr, Diggle, Schlather, Bivand and Ripley (2020) is utilized to identify the values for the two lambdas. A function is implemented that replaces the cash flows in the dataset based on the obtained parameters. After the transformation another Shapiro-Walk test is conducted. While the test statistics increase noticeably the test still does not support the conclusion that the daily cash flows of the datasets are normally distributed. As normality could not be obtained for the datasets, it is reasonable to assume that non-linear models should yield better results. This hypothesis will be tested starting by describing and applying different models in the next chapter.

3. Forecasting Methods and Implementation

3.1 Autoregressive Model

The application of autoregressive models in cash flow forecasting is based on the idea that future cash flows can be forecasted based on the values of the immediately previous cash flows. A more refined approach is the autoregressive integrated moving average (ARIMA) model introduced by Box and Jenkins (1976) that also incorporates moving averages of the residuals as well as the option of differencing which is especially helpful if the time series is non-stationary. Thus the later, more flexible approach will be used.

$$\phi_p(B)\ (1 - B)^d\ Y_t = \theta_q\ (B)\ a_t \qquad (2)$$

The orders of p, d, q indicate the number of autoregressive terms, the number of differencing and the number of lagged forecast errors. B is the Box-Jenkins backshift operator. The autoregressive and moving average parameters that are estimated are denoted as ϕ and θ respectivly.

In order to determine the optimal parameters (p, d, q) the auto.arima-function in *R* is applied to both datasets. Since this function uses an information criterion to evaluate the optimal parameter selection this procedure of model selection is less prone to overfitting which could have a negative impact on the forecast quality. The function returns an ARIMA(1, 0, 0) model including a non-zero mean for dataset1 and an ARIMA(2, 1, 2) without a mean for dataset2, which are the ones that will be used for the forecasting.

3.2 Regression Model

As has been demonstrated by Miller and Stone (1977) a regression model is able to use the presence of day-of-week, day-of-month and holiday effects in the daily cash flow data. This can be done by setting up a regression model with multiple dummy variables for the different characteristics:

$$y_t = \beta_1 HOL_t + \beta_2 DOW_t + \beta_3 DOM_t + \epsilon \qquad (3)$$

Here y_t is the cash flow at day t. HOL is a dummy variable that takes the value 1 on holidays and 0 on every other day. DOW is set of seven dummy variables for each

week of the day and accordingly DOW a set of 31 dummy variables for each day of the month. The regression coefficients are denoted β_i and ϵ represents the residual.

To account for the fact that the effects might not be independent additional interaction parameters are included resulting in an another regression model with a multiplicative specification:

$$y_t = \beta_1 \mathrm{HOL}_t + \beta_2 \mathrm{DOW}_t + \beta_3 \mathrm{DOM}_t + \beta_4 \mathrm{HOL}_t * \mathrm{DOW}_t + \beta_5 \mathrm{HOL}_t * \mathrm{DOM}_t + \beta_6 \mathrm{DOW}_t * \mathrm{DOM}_t + \epsilon \quad (4)$$

The added terms $\mathrm{HOL}_t * \mathrm{DOW}_t$, $\mathrm{HOL}_t * \mathrm{DOM}_t$ and $\mathrm{DOW}_t * \mathrm{DOM}_t$ are the three possible interaction variables.[1]

For both datasets the two models are separately implemented in *R* by using the *lm*-function. Subsequently a stepwise regression with backward elimination is performed as an orientation if each variable is actually necessary. The algorithm of the *step*-function archives this by starting from the full model and testing the deletion of each variable and deleting the variable that best improves the AIC information criterion. This process is repeated until no further variable can be left out without significantly deteriorating the model fit. For dataset1 this procedure returns the full models from (4). The fact that no variable could be removed by the procedure indicates that the different effects are indeed not structurally independent as has been hypothesized by Stone and Miller (1985). Consequently, the full model will be used for the accuracy evaluation in later part of the paper. For dataset 2 the backwards elimination variation of stepwise regression removes the interaction term of β_6 DOW_t*DOM_t. Double checking this assessment by looking at the t-test in the original regression shows a value of -0.052 for this term which is remarkably low. Hence the model from equation (4) with the exception of the last interaction variable will be used for dataset2.

3.3 Random forest model

Random forests as it was proposed by Breiman (2001) are a machine learning method that can be both applied to classification and regression problems. The underlying algorithm trains an ensemble of decision trees, that together form a decision forest. For the purpose of introducing randomness in the sample and lowering the correlation of the trees, a version of bootstrap aggregation (bagging) is used (Hastie, Tibshirani and

[1] Also note that their inclusion changes the interpretation of the original regression coefficients.

Friedman et al. 2009: 587). First the bagging technique randomly selects a random sample of the training set and fits the trees to these samples. Additionally, the idea of node randomization from Dietrich (2000) is incorporated. This technique ensures that while the trees are trained at each node the algorithm randomly chooses a selection of the available features and picks the best criterion for a split among them. In the case of a regression, as it will be used in this paper, the resulting decision trees are simple averaged in order to build a predictive model (Hastie et al. 2009: 587).

For application of random forests in *R* the package *randomForest* was developed by Liaw and Wiener (2002). It allows the input of several parameters that determine the training and consequently the predictions of the model. Different models are prepared with variations in the number of trees that are grown (ntree) as well as different numbers of variables that are considered for each split (mtry). In order to determine the ideal parameters a 10-fold cross valuation method is applied. For this purpose, the data set is first divided into 10 parts. Afterwards 9 parts will be used to train the model and one part will be used to test the accuracy of the predictions each time until every data set has served as a test sample. This is done with the help of the Kuhn's (2008) *caret*-package and is repeated 10 times. The results indicate that mtry = 2 is the ideal parameter for both datasets which is reasonable as it is the only parameter selection that allows the intended node randomization. For both dataset1 and dataset2 ntree will be set to 750 since the for this parameter the lowest root-mean-square error (RMSE) and the highest R^2 are obtained.

4. Evaluation of forecast accuracy and discussion

In order to determine the accuracy of the three conducted forecasting models, a time series cross-valuation as proposed by Hyndman and Athanasopoulos (2019) was performed. In contrast to the k-fold cross-valuation that was used earlier this procedure allows for the assessment of different prediction horizons. This is archived by splitting the dataset into a series of test sets with one cash flow each. The training sets consist of all prior cash flow observations (as well the additional regressors) while in the case of multi-step forecasts, observations immediately preceding the test set are skipped. For this analysis forecast horizons from one to ten days are considered. The resulting forecast errors can be used in a variety of ways to measure the accuracy of the forecasting models. Here both the commonly used Mean Absolute Error (MAE) and the Root mean squared error (RMSE) are calculated. For easier comparison these results

are divided by the results of a naïve estimator that simply uses a mean forecast to predict the future cash flows. Scaling by this benchmark also allows for an additional way of interpreting the results since scaled errors below 1 need to be obtained for the forecasting model in order to have any predictive power.

The results for dataset1 are summarized in figure 1 and 2. As figure 1 shows, the hurdle of performing better than the naive estimator is not trivial.

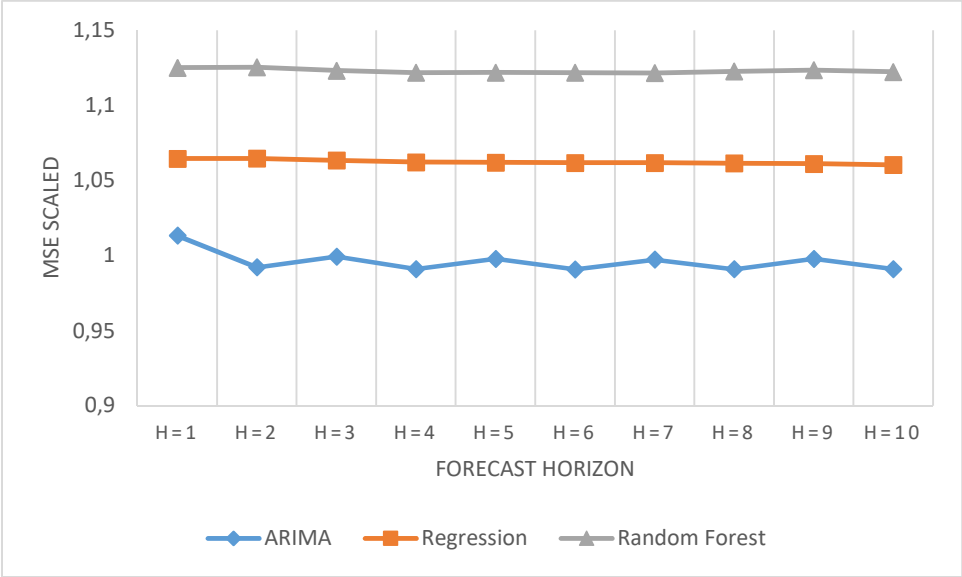

Fig. 1 Comparison of mean squared errors of different forecasting models divided by the MAE of a naïve estimator for dataset1 in dependence on the forecast horizon.

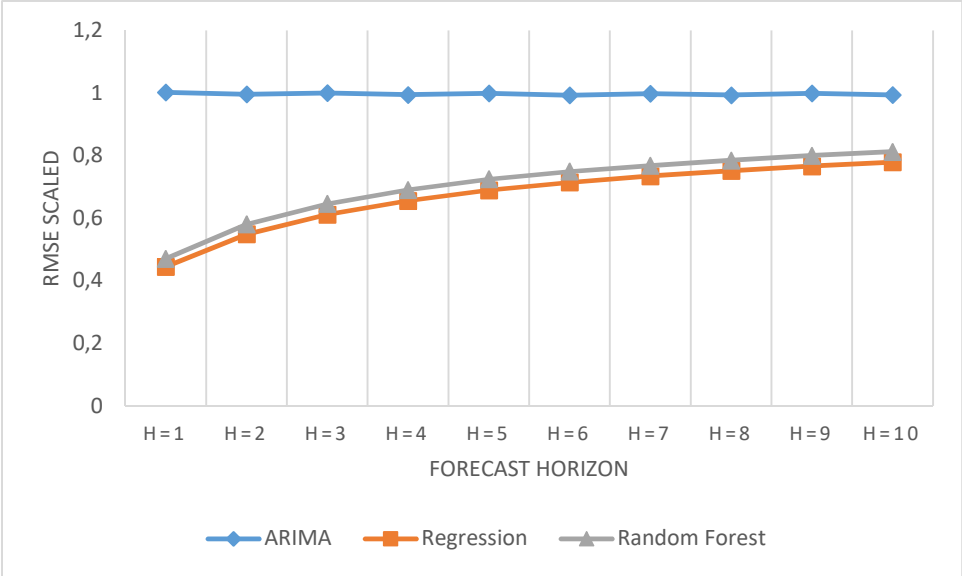

Fig. 2 Comparison of root mean squared errors of different forecasting models divided by the RMSE of a naïve estimator for dataset1 in dependence on the forecast horizon.

In fact, both the regression and the random forest model exhibit notably higher values for the MAE than a mere mean forecast. This however has to be interpreted in the context of the statistical properties of the dataset as presented in table 1. These include

a mean of -0.57 and a very high proportion of cash flows with the value of zero (25.02%). Still based on this metric alone none of the three considered models is able to outperform the benchmark. This is for the ARIMA model also the case when the RMSE is consideration as the scaled error is consistently close to one. A different picture presents itself for the other two models that possess a similar accuracy that especially for short forecasting horizons indicates a much higher accuracy than the benchmark and the ARIMA model.

This reverse relationship in the results for the different metrics results from the way in which these error measures are computed. While both express average forecasting errors that do not differentiate between negative and positive values, in the case of RMSE the errors are squared before they are averaged. As a result, RMSE gives more weight to greater deviations than the MAE metric that weights all deviations equally. It is reasonable to assume that the lower RMSE results for regression and random forest model stem from a better mapping of the major cash flow movements. This however ties into the question what the main focus of the cash flow forecasting is. It was argued in section 2 that cash managers might generally be aware of the major cash flows. If this means that most of the cash flow variation is known already and the forecasting model is applied to predict the remaining less volatile part the model that has achieved the lowest MAE might be a good fit.[2] In the event however, that this assumption does not hold or not sufficiently so, the RMSE might be the better metric for selecting a forecasting model. This is because larger deviations in cash flow forecasts can be especially problematic for cash management as they might for example result in liquidity squeezes, which would give reason to the extra penalization of large forecasting errors.

The results for dataset2 are presented in figure 3 and 4. Again the MAE of the random forest model is significantly above the MAEs of the other forecast models and even the naïve estimator. In contrast the regression and ARIMA model contain slightly lower MAEs than the benchmark of the naïve estimator in this metric. While the ARIMA model shows a similar RMSE as the mean estimator, the regression model is also able to outperform the mean estimator in this metric. As for dataset1 the regression model shows especially high accuracy in this metric for low forecasting horizons. The

[2] While for dataset1 this would be the ARIMA model, using the naïve estimator would work just as well based on this result.

phenomena of increasing scaled RMSE can also be observed for the random forest model whose values range from 0.64 to 0.96. While the predictive accuracy of the random forest model is quite good when forecasting cash flows for the next day, its predictive advantage over the naïve estimator and the ARIMA model is nearly gone when forecasting ten days ahead. The fact that for both datasets the random forest and the regression model can more effectively avoid high forecast errors, might be due to the fact that they also have more additional data inputs.

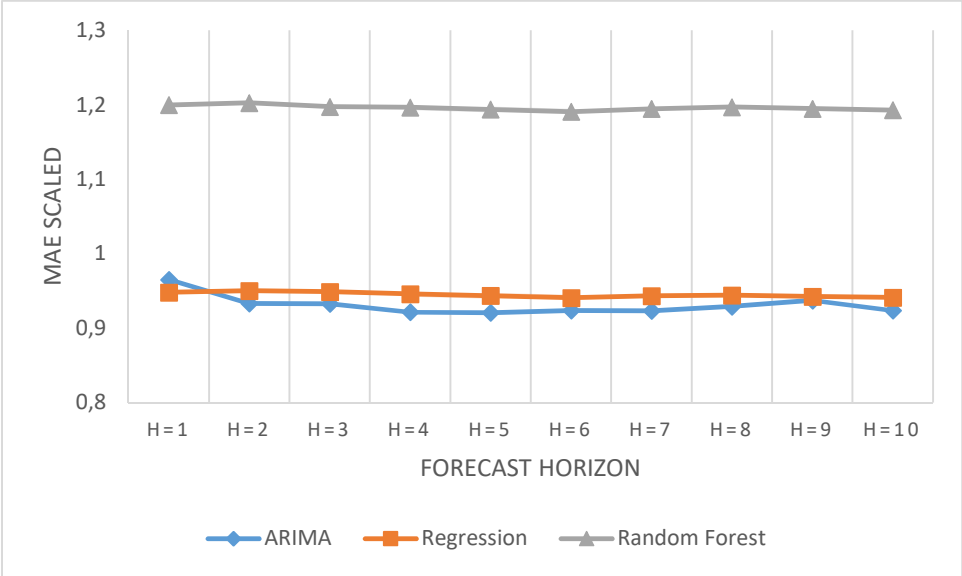

Fig. 3 Comparison of mean squared errors of different forecasting models divided by the MAE of a naïve estimator for dataset2 in dependence on the forecast horizon.

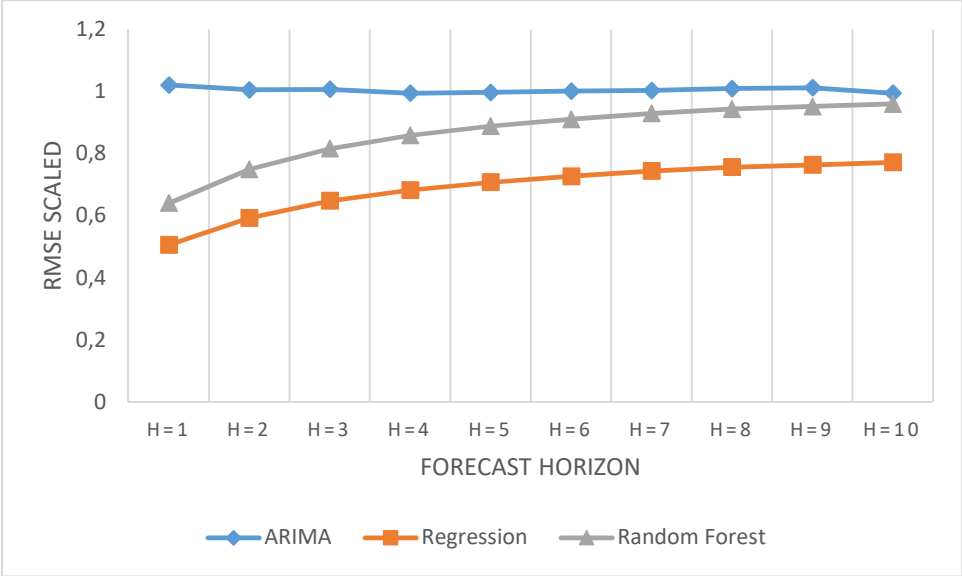

Fig. 4 Comparison of root mean squared errors of different forecasting models divided by the RMSE of a naïve estimator for dataset2 in dependence on the forecast horizon.

While the ARIMA model is exclusively based on the data of previous daily cash flows, the regression and random forest model are apparently able to make use of the additional information regarding holidays, day of the week and day of the months.
However, it is notable that not in a single comparison the random forest model showed the highest accuracy. This is rather surprising as machine learning approaches have shown promising results for forecasting applications and it was argued that non-linear approaches could have an advantage in this field.
This result could be due to a number of reasons. First it should be noted that the number of datasets in this study is quite low. Extending the methodology to more data sets might lead to a completely different result. Additionally, the number of observations might not be large enough for the random forest model to archive optimal results, especially considering that the time series cross validation starts with only a fraction of the available data. Further possible reasons could also lie in the parameters selection or the specifics of the implementation in *R*.

5. Concluding remarks

The previous analysis has by applying an ARIMA, a regression and a random forest model on two daily cash flow datasets shown that the assessment of the predictive accuracy depends on the metric that is considered. On one side the ARIMA forecasts contained lower mean absolute errors than the random forest model and partly the regression model. On the other side regression and random forest model showed particularly for short forecasting horizons better result when the RMSE metric was considered.
Which metric is more applicable for practical use depends on the company and the main focus of the cash flow forecasting. However, since the forecasting is done in the context of cash management which has direct implications for the liquidity of a company, it is reasonable to assume that the higher weighting of larger deviations in the RMSE metric is generally more appropriate.
This study is solely an exemplary application of the methodology with no claim to gain generalizable results. However, this opens up a promising direction for potential future work. To achieve more reliable results, the methodology would have to be applied to a larger number of data sets. In addition, it would be useful to extend the number of forecasting models included. Possible methods here would be radial basis function networks (RBF), multilayer perceptrons (MLP) or extreme learning machines (ELM).

List of References

Affidavit