Mathematical Sciences and Informatics Journal
Vol. 2, No. 1, May 2021, pp. 21-30
http://www.mijuitmjournal.com
DOI : 10.24191/mij.v2i1.10892

Exploration of Machine Learning Forecasting Methods in M4 Competition

Muhammad Halim Hamdan
Deloitte Consulting Malaysia Sdn. Bhd., Taman Tun Dr. Ismail, Kuala Lumpur, Malaysia
mhamdan@deloitte.com

Shuzlina Abdul-Rahman
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor, Malaysia
shuzlina@fskm.uitm.edu.my

Article Info

Article history:

Received Feb 28, 2021
Revised March 05, 2021
Accepted Mei 05, 2021

Keywords:

Forecasting Methods
Makridakis Competition
Machine Learning
Time series

ABSTRACT

There are so many forecasting algorithms and techniques available. The abilities of Data Mining to obtain and gather data from multiple sources is very useful to researcher, practitioner, business and more. From a long list of forecasting algorithms that have been built throughout the years, it will be exhaustive for someone to go through the list one by one to choose which algorithm to use. With M competition established, there are many more new techniques being innovated each time it is organized. This research aims to compare and contrast the machine learning forecasting techniques that are used in M4 Competition, to get better understanding on each technique and to identify the best technique. Three machine learning techniques from M4 Competition were chosen to be compared in this research. Each technique was replicated, trained and tested accordingly. M4 competition dataset was used in this research, with 100,000 time series data and multiple data frequency, which is enough to replicate the real-world situation. The results indicate that the three techniques have their strength, with RNN+ES technique on top of it. RNN+ES and CNN-TS performed well in relative to Naive2 benchmark, while k-NS model performed the worst. Further research on the datasets and investigation of each model to further improve its capabilities are needed to improve the performance of the model.

Corresponding Author:
Shuzlina Abdul-Rahman,
Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA Shah Alam
Selangor, Malaysia.
email: shuzlina@fskm.uitm.edu.my

1. Introduction

Forecasting competition was first held to compare methods rather than people[1]. Given the communication constraints and available tools at the time, it was not appropriate to conduct a large-scale forecasting competition. The first-ever forecasting competition was probably conducted by David Reid as a component of his PhD[2]. The history of time series forecasting dates as far back as 50 years ago. Time series forecasting competition has been a feature in the International Journal since the journals were founded in the 1980s. This emphasizes how the experimental evaluation of forecasting methods, and the needs to compare the most recent proposed methods against existing "state-of-the-art" methods, play a huge role in propelling researchers to develop and build new methods that can work[2].

New researchers wishing to venture into forecasting are often surprised by how controversial the competitions can get, from the first one conducted by Reid[3] to the present-day's Makridakis Competitions. Wallis mentioned in his paper[4] that the competition held by Reid at the University of Nottingham 50 years ago gave rise to many contentions as the researchers disagreed on the combined methods. The Makridakis Competitions were originally a discussion that evolved into five competitions to date (M Competition, M2, M3, M4, and the most recent, M5 [4]-[7]). There is still a long way to go before a one-size-fits-all technique is produced.

PENERBIT PRESS
UNIVERSITI TEKNOLOGI MARA

Five years after Newbold and Granger[8], Spyros Makridakis and Michele Hibon combined a collection of 111 time series and compared a multitude of forecasting methods. The published paper caused quite a stir and was hotly debated. In response to the comments they received, Makridakis and Hibon established a new competition involving the 1001 series where anyone could submit their forecasting techniques, making it the first true forecasting competition. Multiple metrics were used to determine the most accurate method. Fast forward to 2020, the most recently completed competition by Spyros Makridakis was the M4 competition which comprised of 100,000 time-series.

This paper is structured as follows: Section 2 presents the background on the Makridakis Competitions, Section 3 describes the methodology of the research, while Section 4 discusses the results and findings of this research. Section 5 concludes this research.

2. Research Background

2.1 Makridakis Competition

Competition motivates participants to achieve more as the winner is given a prize. However, there are pros and cons to academic competition.

Makridakis Competition (M Competition) was first kicked off by teams led by a forecasting researcher, Spyros Makridakis. The main intention then was to evaluate and compare the accuracy of different forecasting methods[4]-[6], [9]. From there, Makridakis was inspired to continue the competition to this day, with the most recent being the M5 Competition.

As shown in Table 1, the first Makridakis Competition was held in 1982. A reference to it was made in a forecasting journal as the M Competition[7]. This competition comprised of 1001 time series, 15 forecasting methods, and 9 other similar techniques. The main conclusions of the M Competition[7] were as follows:

i. Complex methods do not necessarily produce better forecast results than simpler methods.
ii. The performance ranking of the methods varies according to the accuracy measurement used.
iii. The accuracy of various combined methods, on average, outperforms the individual method being combined and has the best overall performance.

Table 1. List of M Competition

Informal name	Year of results publication	Number of time series used	Number of methods tested	Other features
M Competition	1982	1,001	15	Not real-time
M2 Competition	1993	29 (23 from collaborating companies, 6 from macroeconomic indicators)	16	Real-time, many collaborating organizations,
M3 Competition	2000	3,003	24	-
M4 Competition	Initial 2018, Final 2020	100,000	All major ML and statistical methods.	-
M5 Competition	2020	Around 10,000 hierarchical time series	All major forecasting methods, including methods from the previous competition	-

The findings were verified and replicated by other researchers through the use of new methods [10],[11]. Newbold and Granger[8] addressed that the idea of using a single competition in an attempt to settle the complex issue was unacceptable. Before the first M Competition was established, makridakis and Hibon[12] managed to publish their study showing that complex and sophisticated statistical methods do not necessarily perform better than simple techniques in the

Journal of the Royal Statistical Society (JRSS). Other researchers criticized their claims, and this motivated them to dispel any doubts arising from the study through the M Competition.

The second competition (M2 Competition) was conducted on a much larger scale just over a decade later. A call to join the competition was published in the International Journal of Forecasting and announced at the International Symposium of Forecasting. Written invitations were also sent out to all known experts. The competition was organized in collaboration with four companies and conducted on a real-time basis. The results were published in a 1993 paper[6] and claimed to be identical to those results in the previous competition (M Competition). The M2 Competition used lesser time series (29 time-series) compared to the M Competition (which used 1001 time series). M2 Competition aimed to replicate real-world forecasting problems better in the following aspects:

i. Allow participants to combine their forecasting methods with personal judgement.
ii. Allow participants to inquire about questions regarding the date to produce better forecasts.

In addition to the published results, many participants wrote short articles regarding the competition and their views on what it demonstrated. Chatfield[13] mentioned in his paper that participants did not have much exposure to collaborating companies that could have allowed them to have a feel of real-world forecasting. Fildes and Makridakis[14] stressed that theoretical statisticians still tended to ignore the implications even after the evidence was produced through these competitions.

The third competition in the series (M3 Competition), organized in 2000, was intended to replicate and develop the features of the one before, achieved by including more methods and researchers, most from the areas of neural network and time series [5]. Table 2 lists the number of time series data for M3 Competition.

Table 2. The number of time series data based on time interval and domain.

The time interval between observations	**Micro**	**Macro**	**Industry**	**Finance**	**Demographic**	**Other**	**Total**
Yearly	146	83	102	58	245	11	645
Quarterly	204	336	83	76	57	0	756
Monthly	4	0	0	29	0	141	174
Other	474	312	334	145	111	52	1,428
Total	828	731	519	308	413	204	3,003

As shown in Table 2, the data comprised a total of 3003 time series, which included Yearly, Quarterly, Monthly and Other. A minimum threshold for each observation was set to ensure that sufficient data was available. Several papers were published with distinctive analyses of the data from the M3 Competition [15],[16]. Hyndman and Koehler[16] also mentioned that the M3 data had been used continuously since 2000 to test new time series forecasting methods. The following were the five measures used to evaluate the performance of different forecasting techniques[5]:

i. Symmetric Mean Absolute Error (SMAE).
ii. Average ranking.
iii. Median Symmetric Absolute Percentage Error (MSAPE).
iv. Percentage better.
v. Median RAE.

Through the years, M Competitions have gathered a sizeable audience and numerous participants from both academics and practitioners to provide evidence of the most suitable forecasting technique. M4 Competition, announced in 2017, started in Jan 2018 and ended in May 2018. The M4 Competition continued to extend and replicate the results from previous competitions by using a more extensive and diverse set of time series to identify the most accurate forecasting methods and techniques to improve forecasting accuracy, as well as to determine the most suitable methods[4]. As shown in Table 3, the competition used 100,000 real-life time series and covered leading forecasting techniques, including traditional methods and those based on Artificial Intelligence, such as Machine Learning. Hyndman and Athanasopoulos[17] congratulated

Makridakis for his massive influence on the field of forecasting with a focus on models that produced good forecasts rather than mathematical characteristics of those models. The findings and conclusions of the M4 competition will be discussed in detail in this research, and the forecasting methods used will be discussed in the next section.

Table 3. The number of time series data based on time interval and domain.

The time interval between observations	Micro	Macro	Industry	Finance	Demographic	Other	Total
Yearly	6,358	3,903	3,716	6,519	1,088	1,236	23,000
Quarterly	6,020	5,315	4,637	5,305	1,858	865	24,000
Monthly	10,975	10,016	10,017	10,987	5,728	277	48,000
Weekly	112	41	6	164	24	12	359
Daily	1,476	127	422	1,559	10	633	4227
Hourly	0	0	0	0	0	414	414
Total	25,121	19,402	18,798	24,534	8,708	3,437	100,000

2.2 Forecasting Methods in M4 Competition

As mentioned in Section 2, the M4 dataset consisting of 100,000 real-life time series were divided into training and test sets. The training set was given at the beginning of the competition; the test set was kept secret and only released after the organizers have used it to evaluate the submissions. It was decided for the training set to have a minimum number of observations, such as 13 for Yearly, 16 for Quarterly, 42 for Monthly, 80 for Weekly, 93 for Daily, and 700 for Hourly. Overall, the M4 dataset is a much longer series compared to the M3 and hence allowed more opportunities for more complex methods.

After the success of the previous competition, various benchmarks, both statistical and Machine Learning (ML), were introduced in the M4 Competition. Since the first M Competition, forecasting has progressed so much it was concluded that “complex or sophisticated statistical methods are not necessarily better than simpler methods”. Over the years, new methods have been proposed, tested and proven to be more accurate than simpler ones[1]. For that, the organizers established ten benchmark methods to be included for two reasons: the first to evaluate the improvement of the M4 submissions over standard approaches, the second to identify the causes of improvements by comparing each offering against different techniques. Examples include: Naïve 2 [18],[19] only captures seasonality, Single Exponential Smoothing (SES) captures the level, Holt uses the linear trend to extrapolate, while Damped, as its name suggests, dampens the linear trend [20],[21].

Table 4 lists out the ten benchmarks used in the M4, plus two more standards for comparison: ETS (exponential smoothing) and Auto-Regressive Integrated Moving Average (ARIMA) [22],[23]. These were added for their broad utilization in forecasting over the years. Two pure ML methods were also included, multi-layer perceptron (MLP) and Recurrent Neural Network (RNN), to highlight the objective of the competition concerning the ML uses in forecasting as well as to boost the participation of ML methods[15]. Both of the benchmarks were basic architectures, trained individually using typical preprocessing to prevent limiting the participants and to encourage more innovative solutions.

2.3 Forecasting Model

i. Hybrid of RNN and ES

The model, replicated from Smyl[24], combines Exponential Smoothing (ES) formula and Recurrent Neural Network (RNN) forecasting engine. The model is a true hybrid algorithm where all parameters, such as the initial ES coefficients, are fitted simultaneously with RNN weight using the same Gradient Descent method. As the dataset is provided without any timestamps, it is deseasonalized using a standard procedure like Seasonality and Trend decomposition using Loess (STL). After deseasonalization is completed, feature extraction is done by using rolling input and output windows of constant size. Windowing, normalization and regressors are done to assist with feeding good data shapes into ES-RNN model.

ii. CNN for time series

This model uses all-time series data in each given data frequency to train the model. By using all data frequencies, the issue of insufficient data can be avoided. It is also unnecessary to explicitly build the mathematical behaviour for each series by using the CNN approach as it learns each behaviour on its own by looking at past data provided. A separate model is used for each data frequency, and cyclic effects are expected on all data frequencies except for yearly. Cyclic effect means Daily or Weekly patterns can exist for Hourly data frequency. Separate models are trained for each data frequency, and the loss function used for the networks is the mean squared error. As the network architecture requires fixed-length inputs, multiple models are trained for each data frequency.

iii. k-NS model

The forecasting model, k-Nearest Similarity model, also known as k-NS, is made based on the preprocessed time series data, where each input and output variables are defined as patterns. The patterns are then modelled using the weight function to define the membership of each learning point. This model uses the idea of pattern similarity-based forecasting, where it assumes that if the input of *ai* and *aii* are similar, then patterns *bi* and *bii* paired with them are also alike. The assumption allows forecasting to be made based on known patterns. Each pattern components are defined using mathematical formulas.

Table 4. Benchmarks and standards for comparison

Methods		**Description**
Statistical benchmarks	Naïve 1	A random walk model, assuming feature values will be the same as that of the last observation.
	Naïve S	Forecasts are equal to the last observation of the same period.
	Naïve 2	Similar to Naïve 1 but the data are seasonally adjusted.
	SES	Exponentially smoothing the data and extrapolating assuming no trend.
	Holt	Exponentially smoothing the data and extrapolating assuming a linear trend.
	Damped	Exponentially smoothing the data and extrapolating assuming a damped trend.
	Theta	As applied to the M3 Competition using two Theta lines, $\vartheta 1 = 0$ and $\vartheta 2 = 2$, with the first one being extrapolated using linear regression and the second one using SES.
	Comb	The simple arithmetic average of SES, Holt and Damped exponential smoothing.
Standard for comparison	MLP	A perceptron of a very basic architecture and parameterization. Some preprocessing like detrending and deseasonalization is applied beforehand to facilitate extrapolation.
	RNN	A recurrent network of a very basic architecture and parameterization. Some preprocessing like detrending and deseasonalization is applied beforehand to facilitate extrapolation.
	ETS	Automatically provides the best exponential smoothing model, indicated through information criteria.
	ARIMA	An automatic selection of possible ARIMA models is performed and the best one is chosen using appropriate selection criteria.

3. Methodology

The methodology of this research is established on the CRISP-DM model, which is based on previous attempts to define knowledge discovery methodologies. The model for data mining gives an insight into a data mining project's lifecycle. The project phases, tasks for each phase, and the outcome are the essence of the model. The lifecycle of CRISP-DM is divided into six main phases

as shown in Figure 1. The arrows illustrate the most significant and common dependencies between phases, and the application for any particular research is dependent on the objective and outcome of each phase to determine the next course of action.

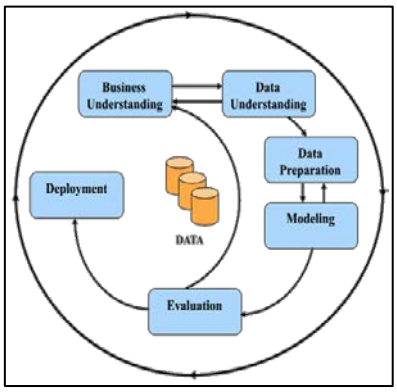


Figure 1. Lifecycle of CRISP-DM

The first phase of this research is Business Understanding. The primary analysis is related to forecasting competition and techniques, as well as the different types of forecasting methods. The next phase, Data Preparation, is crucial before moving on to Modelling. The data consists of 100,000 time series and contains many attributes; the details are generalized, unknown, and come from various sources that are consolidated into one. The M4 dataset is considered complete and ready for use in forecasting, but some work is still needed to prepare the data for the models. Some common issues with data are incomplete, inconsistent, missing, or corrupted values. This kind of data is not in good form and unsuitable for modelling. Quality data is crucial to get a high performing modelling result. Inaccurate data entry or misunderstanding can cause substandard data quality and bias in model performance[25].

The Modelling phase is where the researcher seeks out useful patterns in the data, usually performed repeatedly as getting the model right is not an easy task. Generally, there are many pre-built models available, either pre-installed with the tools or on a public repository. The researcher must be able to work with different models and understand how each model works. However, in this research, the models are adopted from the three submissions in the M4 Competition, namely, a hybrid of RNN and ES, CNN, and k-Nearest Similarity. These three models are addressed in Section 2.2.

The last phase, Evaluation, is the most crucial where the performance and overview of the models are analysed. The following are discussed in this phase: factors leading to the achievement of the model, aspects justifying the accomplishment of the objectives, and drawbacks encountered during this research. Along with that, distinctive results are assessed, and subsequently, additional information, challenges and ideas are established.

In the first activity, the performance of the models is evaluated, and any shortcomings are examined. The evaluation assessment includes measuring the performance of each model using sMAPE, MASE and OWA. In accordance with the M4 Competition, the model that achieved the lowest OWA score is selected as the best model, analysed and compared against other models.

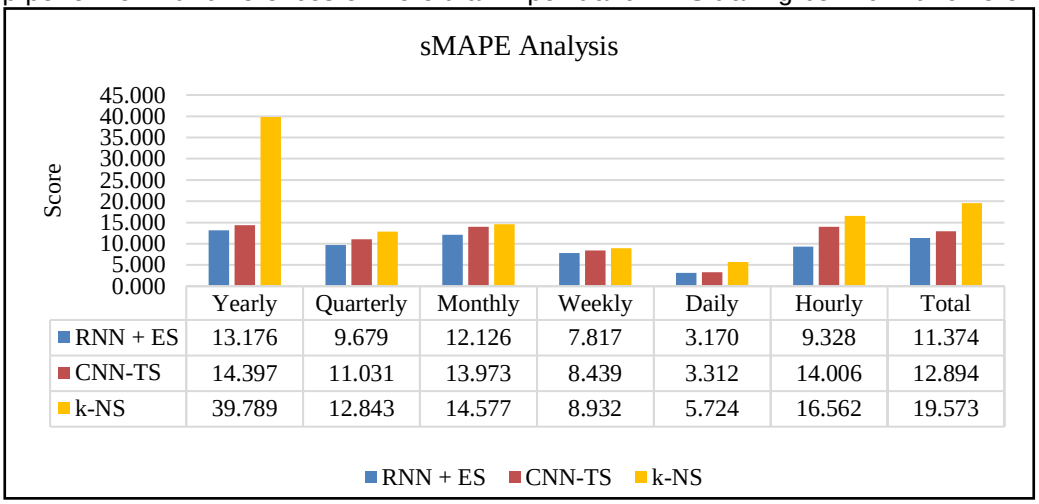
4. Results and discussions

The main point of this section is to analyse the results that correspond with the experiments done on the dataset using selected methods. The discussions will cover the analysis using sMAPE, MASE and OWA metrics[26].

4.1 Overview of sMAPE Analysis

The bar chart in Figure 2 below shows the comparison of the sMAPE analysis for RNN+ES, CNN-TS and k-NS models for all data frequencies. The chart in Figure 2 illustrates the results according to data frequencies and forecasting models. The lowest bar chart is for Daily data frequency, which uses RNN+ES as the method. This model outperforms others with a sMAPE score of 3.170, followed by CNN-TS at 3.312 and k-NS at 5.724. The highest bar chart is for Yearly data frequency, using k-NS as a method with a sMAPE score of 39.789, a significant difference of 25.392

points. RNN+ES and CNN-TS receive scores of 13.176 and 14.397 respectively, significantly lower than k-NS. This is the result of k-NS extensive dependence on pattern similarity, and with only 23,000 data points for Yearly data frequency, the patterns are not distinct enough to forecast future data. For Quarterly data frequency, RNN+ES achieves a sMAPE score of less than 10 points, which at 9.679 is better than other models, where CNN-TS achieves a score of 11.031, followed by k-NS at 12.843. The same order occurs in Monthly, Weekly and Hourly data frequencies, where RNN+ES is the top performer with differences of more than 1 point and k-NS trailing behind with differences of



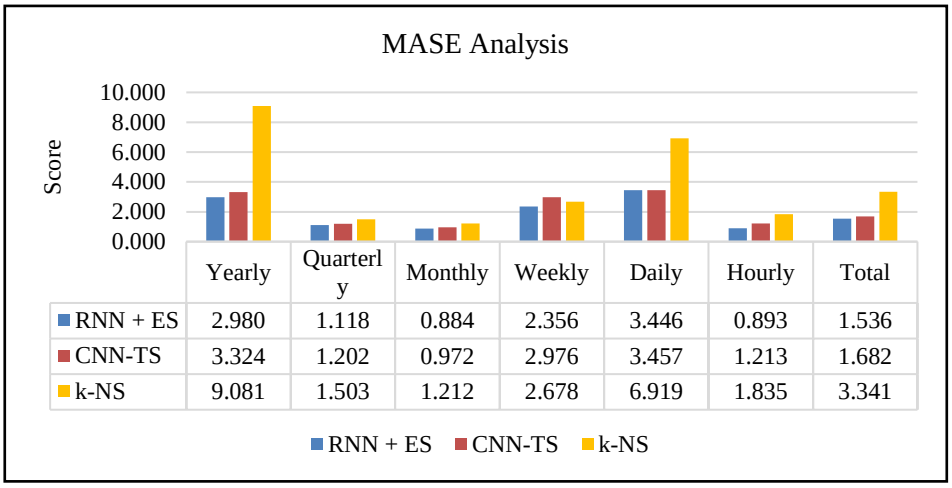
	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly	Total
RNN + ES	13.176	9.679	12.126	7.817	3.170	9.328	11.374
CNN-TS	14.397	11.031	13.973	8.439	3.312	14.006	12.894
k-NS	39.789	12.843	14.577	8.932	5.724	16.562	19.573

more than 0.5 point for every data frequency. The largest difference is 4.678 points in Hourly data frequency. Taking everything into consideration, it can be seen that the RNN+ES model has the best performance with a sMAPE score of 11.374, compared to the other models where the sMAPE score differences exceed 1.5 points.

Figure 2. sMAPE Analysis score on different data frequencies

4.2 Overview of MASE Analysis

The bar chart in Figure 3 shows the comparison of the MASE analysis for RNN+ES, CNN-TS and k-NS models for all data frequencies. As can be seen in this Figure 3, the MASE scores are close for all models in Quarterly, Monthly and Hourly data frequencies. For Quarterly data frequency, RNN+ES performs the best with a MASE score of 1.118, followed by CNN-TS at 1.202 and k-NS at 1.503. For Monthly data frequency, RNN+ES also performs the best with a MASE score of 0.884, followed by CNN-TS at 0.972 and k-NS at 1.212. For Hourly data frequency, RNN+ES performs the best with a MASE score of 0.893, followed by CNN-TS at 1.213 and k-NS at 1.835. The performances for all models are on par with differences of less than 0.6. All three models work well with the given data frequencies. The MASE score for the k-NS model is worse than other models in Yearly and Daily data frequencies, at 9.081 and 6.919 respectively, compared to the best performing model, RNN+ES, at 2.980 and 3.446, with differences of more than 5 and 3 points. Taking everything into consideration, the performances of RNN+ES and CNN-TS are relatively similar across all data frequencies with MASE scores of 1.536 and 1.682 respectively. k-NS has the worst performance with a score of 3.341 against the differences of more than 1.5 points for the other two models.

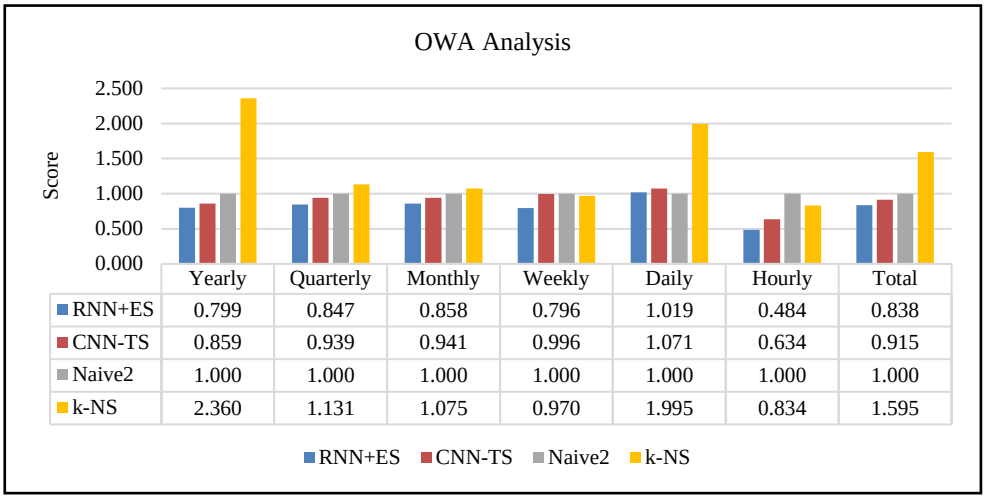


	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly	Total
RNN + ES	2.980	1.118	0.884	2.356	3.446	0.893	1.536
CNN-TS	3.324	1.202	0.972	2.976	3.457	1.213	1.682
k-NS	9.081	1.503	1.212	2.678	6.919	1.835	3.341

Figure 3. MASE Analysis score on different data frequencies

4.3 Overview of OWA Analysis

The bar chart in Figure 4 below shows the comparison of the OWA analysis for RNN+ES, CNN-TS and k-NS models, as well as Naïve 2 (benchmark) for all data frequencies.The chart in Figure 4 illustrates two data frequencies, Yearly and Daily, which achieve notably higher scores than other data frequencies. In Daily data frequency, k-NS has the worst performance with an OWA score of 2.360, which is more than 1 point higher than the other two models. This is due to the k-NS model's failure to capture enough patterns, thus causing the forecasting technique to perform badly. Furthermore, there is insufficient data in Daily data frequency for the model to pick up notable patterns. The same applies to Yearly data frequency, although the model should logically capture enough patterns as Yearly data frequency have significantly higher data points compared to the others. The model fails to produce a good forecast as a result of overfitting. As there are a lot of patterns that can be captured by the model, it learns from the details and noises in the data, thus negatively impacting the results. In Daily data frequency, no model achieves an OWA score of less than 1.0; RNN+ES scores with 1.046, followed by CNN-TS at 1.071 and k-NS at 1.995. Taking everything into consideration, only RNN+ES and CNN-TS perform well relative to the Naïve 2 model. In each data frequency, the RNN+ES model performs slightly better than the other two, resulting in it being the best model overall.



	Yearly	Quarterly	Monthly	Weekly	Daily	Hourly	Total
RNN+ES	0.799	0.847	0.858	0.796	1.019	0.484	0.838
CNN-TS	0.859	0.939	0.941	0.996	1.071	0.634	0.915
Naive2	1.000	1.000	1.000	1.000	1.000	1.000	1.000
k-NS	2.360	1.131	1.075	0.970	1.995	0.834	1.595

Figure 4. OWA Analysis score on different data frequencies

4.4 Discussion on the Analysis

As addressed in the previous section, the models' performances are compared based on three metrics, sMAPE, MASE and OWA. A detailed discussion takes place to determine how one model is better than the others. Each metric's measurement is illustrated and comparisons are made in the bar chart. A thorough study is done for individual models to understand each result better. Each is adopted and completely replicated, and each component is understood. A few key points can be pointed out as a result of this research. The hybrid model of RNN+ESS performs better than the other two models and the benchmark model. From ranking the results of the competition, there are only one hybrid model and less than five pure machine learning models. The combination of ES and RNN as a truly hybrid model, by fitting each parameter concurrently with the RNN weights, greatly improves the forecasting result. The improvement of the Hybrid model over Comb achieves a score close to 10%, showing clear superiority of true hybrid algorithms and the improvement of statistical and ML methods.

The paper published by[4] concludes that a complex method will not necessarily produce a better result compared to a simpler method. It is proven that more complex and sophisticated methods perform better than the simpler ones, such as the Benchmark method. The k-NS model can perform well if there is sufficient data to draw a well-defined pattern. In another instance, the k-NS model tends to underfit and overfit when the data fed into the model is either insufficient or has too much noise. This will produce a bad forecast, hence making it less proficient. From the results of the OWA analysis, the RNN+ES model is chosen as the best performing model as it achieves the lowest score. The model is evaluated by supplying the combined dataset for training and testing. From the results, the model performs as expected and the score is the same.

5. Conclusion

Three models were chosen, replicated, and tested to demonstrate the performance of each model: RNN+ES, CNN-TS and k-NS models. The models were tested on all data frequencies and domains. The performance of each model was interpreted using different metrics such as sMAPE, MASE and OWA.

The main contribution of this research is to assist practitioners and researchers to choose related forecasting methods for their works. Each forecasting method in this research is explained thoroughly and replicated to understand the algorithm functions better. In conclusion, there are notable differences between each chosen algorithm, and they all have their advantages and disadvantages.

Acknowledgements

The authors would like to thank Universiti Teknologi MARA Shah Alam for the support given to complete this project.

References

[1] R. J. Hyndman, "A brief history of forecasting competitions," International Journal of Forecasting, vol. 36, no. 1, pp. 7-14, 2020.

[2] R. Fildes and K. Ord, "Forecasting competitions–their role in improving forecasting practice and research," A companion to economic forecasting, pp. 322-253, 2002.

[3] D. J. Reid, A comparative study of time series prediction techniques on economic data. University of Nottingham, Library Photographic Unit, 1969.

[4] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: Results, findings, conclusion and way forward," International Journal of Forecasting, vol. 34, no. 4, pp. 802-808, 2018.

[5] S. Makridakis and M. Hibon, "The M3-Competition: results, conclusions and implications," International journal of forecasting, vol. 16, no. 4, pp. 451-476, 2000.

[6] S. Makridakis et al., "The M2-competition: A real-time judgmentally based forecasting study," International Journal of Forecasting, vol. 9, no. 1, pp. 5-22, 1993.

[7] S. Makridakis et al., "The accuracy of extrapolation (time series) methods: Results of a forecasting competition," Journal of forecasting, vol. 1, no. 2, pp. 111-153, 1982.

[8] P. Newbold and C. W. Granger, "Experience with forecasting univariate time series and the combination of forecasts," Journal of the Royal Statistical Society: Series A (General), vol. 137, no. 2, pp. 131-146, 1974.

[9] G. Libert, "The M-competition with a fully automatic box–jenkins procedure," Journal of

Forecasting, vol. 3, no. 3, pp. 325-328, 1984.
[10] M. Geurts and J. Kelly, "Forecasting demand for special services," International Journal of Forecasting, vol. 2, no. 3, pp. 90046-4, 1986.
[11] R. Fildes, M. Hibon, S. Makridakis, and N. Meade, "Generalising about univariate forecasting methods: further empirical evidence," International journal of Forecasting, vol. 14, no. 3, pp. 339-358, 1998.
[12] S. Makridakis and M. Hibon, "Accuracy of forecasting: An empirical investigation," Journal of the Royal Statistical Society: Series A (General), vol. 142, no. 2, pp. 97-125, 1979.
[13] C. Chatfield, "A personal view of the M2-Competition," International Journal of Forecasting, vol. 9, no. 1, pp. 23-24, 1993.
[14] R. Fildes and S. Makridakis, "The impact of empirical accuracy studies on time series analysis and forecasting," International Statistical Review/Revue Internationale de Statistique, pp. 289-308, 1995.
[15] A. J. Koning, P. H. Franses, M. Hibon, and H. O. Stekler, "The M3 competition: Statistical tests of the results," International Journal of Forecasting, vol. 21, no. 3, pp. 397-409, 2005.
[16] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International journal of forecasting, vol. 22, no. 4, pp. 679-688, 2006.
[17] R. J. Hyndman and G. Athanasopoulos, Forecasting: principles and practice. OTexts, 2018.
[18] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, Forecasting methods and applications. John wiley & sons, 2008.
[19] S. Wheelwright, S. Makridakis, and R. J. Hyndman, Forecasting: methods and applications. John Wiley & Sons, 1998.
[20] E. S. Gardner Jr, "Exponential smoothing: The state of the art—Part II," International journal of forecasting, vol. 22, no. 4, pp. 637-666, 2006.
[21] E. S. Gardner Jr, "Exponential smoothing: The state of the art," Journal of forecasting, vol. 4, no. 1, pp. 1-28, 1985.
[22] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, "A state space framework for automatic forecasting using exponential smoothing methods," International Journal of forecasting, vol. 18, no. 3, pp. 439-454, 2002.
[23] R. J. Hyndman and Y. Khandakar, Automatic time series for forecasting: the forecast package for R (no. 6/07). Monash University, Department of Econometrics and Business Statistics, 2007.
[24] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," International Journal of Forecasting, vol. 36, no. 1, pp. 75-85, 2020.
[25] I. Taleb, H. T. El Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, "Big data quality: A quality dimensions evaluation," in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016, pp. 759-765: IEEE.
[26] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," International Journal of Forecasting, vol. 36, no. 1, pp. 54-74, 2020.