

DeepAR for Probabilistic Influenza Disease Forecasting


Kai Reffert ¹

Abstract: Epidemic forecasts are of great importance to public health organizations, increasing situational awareness and providing time to set preventive actions in place. Probabilistic forecasts further quantify uncertainty, which is especially important in decision-making and risk management. Since epidemiological forecasting often produces many related time series, I propose that global models, which estimate model-parameters jointly based on every time series, are beneficial for this task. Thus, I investigate the applicability of the global DeepAR model [Sa20] to probabilistic influenza disease prediction in German districts. Furthermore, I compare it to an endemic-epidemic modelling approach and a feedforward neural network to assess the suitability of DeepAR on this task. In summary, I conclude that the DeepAR model is beneficial for the problem and that it outperforms its rivals.

Keywords: DeepAR, Influenza, Probabilistic Time Series Forecasting, Epidemiological Time Series, Global Time Series Forecasting

1 Introduction

For many years, the infectious disease of influenza has immensely endangered humans. The World Health Organisation estimates that annually occurring outbreaks of influenza are responsible for approximately 3 to 5 million cases of severe illness and 290 to 650 thousand deaths [WHO2323]. Overall, the impact reduction of seasonal epidemics and pandemics is of greater importance to public health authorities [Ac12]. Furthermore, Longini et al. [Lo05] and Ferguson et al. [Fe05] show that epidemics may be preventively contained if an early detection system is in place. Thus, one of the key challenges in public health is to continuously monitor and predict the development of influenza in the population, as this presents public health authorities with time to set necessary policies and resources in place [Ju22]. However, while point or single-valued forecasting methods have been among the most common time series forecasting techniques, they lack information about uncertainties of their predictions, which can be a major disadvantage when the forecasts are used in decision-making. Therefore, Bracher et al. [Br21] state that the importance of probabilistic forecasting in epidemiology is continuously growing. Similar to other domains, the age of Big Data introduced soaring amounts of related epidemiological time series, where single time series may contain relevant information to other series. Thus, local models, which estimate their model parameters individually for each time series, are potentially inferior to global models, which estimate their model parameters jointly from all available time series through cross-learning. However, Hewamalage et al. [HBB21] argue that the outperformance conditions of global over local models still remain uncertain.

¹ Universität Mannheim, kai.reffert@students.uni-mannheim.de,  <https://orcid.org/0009-0007-7817-726X>;
Thesis conducted at Karlsruhe Institute of Technology (KIT).

To the best of my knowledge, the DeepAR model was not yet implemented to produce probabilistic epidemiological (or influenza disease) forecasts. Therefore, I assess the short-term prediction quality of DeepAR on the task of forecasting influenza surveillance time series data in German districts by comparing it to two baseline models, a regular feedforward neural network and an endemic-epidemic multivariate spatio-temporal model [HHH05]. The results suggest that DeepAR outperforms the rivaling models by providing sharp and accurate probabilistic forecasts. Thus, I propose DeepAR could prove beneficial to policymakers and public health providers by improving their policy and resource allocation choices in response to potential influenza threats.

The remainder of this paper is organized as follows. In Section 2, I discuss related work. After that, I theoretically describe the applied methodology in Section 3. In Section 4, I present the results and end this work with a concise summary.

2 Related work

There have been several approaches set up for the task of predicting influenza time series, in which influenza surveillance data of different geographical regions is used to predict the development of cases in the (short-term) future. At first, mainly mechanistic and statistical models were implemented, e.g. an SIRC model variation to explore epidemiological effects of influenza A viruses [Ca06] or an autoregressive model fit on influenza surveillance data from the United States [Ac12]. More recently, DL models were also used to forecast influenza surveillance time series. Tapak et al. [Ta19] compared support vector machines, random forests and FNNs on the task of time series modelling and outbreak detection of influenza disease cases in Iran. Yang et al. [Ya20] implemented a long short-term memory (LSTM) model to deal with influenza time series data in Taiwan. Wu et al. [Wu18] developed a deep learning approach by combining a recurrent neural network (RNN) with a convolutional neural network (CNN). They compared their approach with traditional statistical models on regional influenza surveillance time series of the United States and Japan. In fact, Jung et al. [Ju22] incorporated the approach of Wu et al. [Wu18] into their analysis. Furthermore, they additionally introduce a deep learning based self-attention model and compare it with other models on regional influenza data sets from the United States and prefectures in Japan. Similarly, Wu et al. [Wu20] applied a Transformer-based model to forecast influenza time series.

Although the previously mentioned models produce respectable results on influenza time series prediction tasks, all of them consider point instead of probabilistic forecasts. On a similar note, there are probabilistic time series prediction methods applied to other epidemiological time series, for example COVID-19 [Cr22; En23], however the proposed methodologies are often not directly applicable to other use cases, e.g. the prediction of influenza disease. Nevertheless, the work of Soliman et al. [SLG19] is comparable to this work, as they compared probabilistic forecasts of traditional statistical models, a feedforward neural network (FNN) and Bayesian multi-model ensembles on seasonal influenza data of

Dallas County. Their results suggest that the FNN was only outperformed by two ensemble approaches, while outperforming the remaining models.

3 Methods

In this section, I first present the models I considered. Then, I describe the evaluation metrics. Lastly, I shortly go over the composition of the data set.

Models. Because influenza cases of different regions behave similar and are likely correlated, I propose that the DeepAR model, which was introduced by a team of Amazon’s researchers [Sa20], thrives in this task, as its global modelling approach was designed to tackle forecasting problems of thousands or millions of related time series. Furthermore, DeepAR is an autoregressive probabilistic model, which is based upon an RNN backbone to model the next-element distribution $p(y_{i,t}|\theta(h_{i,t}, \Theta))$, where $y_{i,t}$ corresponds to the observation for target time series $i \in [1, \dots, N]$ at time step $t \in [t_0, \dots, T]$. The distribution depends on a function $\theta(h_{i,t}, \Theta)$ that determines the distribution parameters based on model parameters Θ and the output of the underlying recurrent network $h_{i,t}$. Since the objective involves the prediction of positive influenza count data, I chose to model the negative binomial distribution, which is often selected to represent positive count data.

Since we are limited to probabilistic models, many approaches mentioned in Section 2 can not be implemented as comparative methods. Nevertheless, a FNN performed relatively well for Soliman et al. [SLG19], hence I considered a simple FNN² of the GluonTs³ python library. The FNN also utilizes a negative binomial distribution to produce probabilistic forecasts. Besides this, I also decided to compare a traditional statistical probabilistic method, which is based on the endemic-epidemic multivariate spatio-temporal model (hhh4) introduced by Held et al. [HHH05]. However, the original Poisson distribution is dropped in favor of the negative binomial distribution [PH11]. Thus, the infectious disease counts $y_{i,t}$ follow a negative binomial distribution $y_{i,t}|y_{i,t-1} \text{ NegBin}(\mu_{i,t}, \psi)$, where the mean $\mu_{i,t}$ is additively partitioned into endemic and epidemic parts while ψ is a strictly positive overdispersion parameter. The endemic component is responsible for the representation of a default rate of cases, whereas the epidemic component is tasked of modelling sudden target deviations based on previous disease count realizations, the latter is further decomposed into autoregressive and neighborhood effects.

Evaluation metrics. Traditional evaluation metrics, e.g. MAE or MSE, are unsuitable for probabilistic time series forecasts. Instead, the objective of probabilistic forecasting is to maximize the sharpness of its predictive distribution subject to calibration [GBR07]. Sharpness describes the concentration of predictive distributions, i.e. the sharper a forecast, the more informative it is. Calibration refers to the consistency between the predictions and the actual observations, ideally the observations are indistinguishable from random draws of the predictive distribution. Although it is common in probabilistic forecasting to predict and

² https://ts.gluon.ai/stable/api/gluonts/gluonts.mx.model.simple_feedforward.html

³ <https://ts.gluon.ai/>

output full probability distributions, they are difficult to store in full detail unless they are parametric. Furthermore, Bracher et al. [Br21] suggest evaluating probability distribution forecasts reported in an interval or quantile format during short notice epidemic situations. In fact, this format was also adopted by the COVID-19 Forecast Hub [Ra20]. Following this, I also evaluate probability distribution forecasts reported in an interval or quantile format⁴. To assess the calibration of quantile forecasts, I select the empirical coverage metric, which corresponds to the share of observed values that fall below a given prediction quantile, for example the percentage of observations that is lower or equal to the value of the 50% quantile, which is ideally close to 50%. Gneiting et al. [Gn23] further suggest distinguishing between lower and upper coverage for count data. While the upper coverage takes observations into account that are smaller or equal to the predicted quantile value, the lower coverage only accepts observations that are strictly smaller than the predicted quantile value. On a different note, I additionally evaluate sharpness and calibration with a single score by considering the weighted interval score (*WIS*) [Br21], as defined in Equation 1, where y represents a single realized observation while l and u correspond to the lower and upper bound of the predicted interval. *WIS* is a proper scoring rule⁵ if all weights w are nonnegative and unnormalized. The *WIS* incorporates the interval score (*IS*) [GBR07], see Equation 2, for multiple distinct levels of prediction intervals $(1 - \alpha_1) < \dots < (1 - \alpha_K)$, where the predictive median m is interpreted as the $(1 - \alpha_0)$ PI. I selected the weights of the *WIS*, so that it approximates the continuous ranked probability score (*CRPS*) [Br21], which is a popular proper score to evaluate full probability distributions.

$$WIS_{\alpha_{0,k}}(l, u; y) = \frac{1}{K + 0.5} \cdot (w_0 \cdot |y - m| + \sum_{k=1}^K \{w_k \cdot IS_{\alpha}(l, u; y)\}) \quad (1)$$

$$IS_{\alpha}(l, u; y) = (u - l) + \frac{2}{\alpha} \cdot (l - y) \cdot I(y < l) + \frac{2}{\alpha} \cdot (y - u) \cdot I(y > l) \quad (2)$$

Data. The influenza time series $y_{i,t}$ originate from the SurvStat@RKI database⁶, which is administrated by the Robert Koch Institute in Germany. These series represent counts of influenza cases of different German districts, also known as “Landkreise” (LK) and “Stadtkreise” (SK) $i = 1, \dots, 411$, aggregated over weekly time points $t = 1, \dots, 1148$ reported on Sundays. The reporting period starts with the week of the 7th January 2001 and ends with the week of the 1st January 2023. The time series are continuously reported due to the German reporting obligation of direct influenza virus detection. Due to the severe impact of COVID-19 sanctions, which started in the year 2020, on the spread of influenza, I decided to restrict the analysis to weeks before the 30th September 2020. The influenza time series often have a single annual peak at the beginning of the year, while the

⁴ Note that I exclusively consider central prediction intervals (PIs), such that $(1 - \alpha)$ is the nominal coverage rate and the lower and upper bound correspond to the level $(\alpha/2)$ and $(1 - \alpha/2)$ predictive quantiles respectively.

⁵ Proper scoring rules encourage forecasters to express their true beliefs, as there is no enticement for forecasters to report deviating forecasts from their honest opinion about the future [Wi96].

⁶ <https://survstat.rki.de/>

remainder of the year has values close to zero. Hence, I chose the following periods, 1st January 2001–1st October 2016, 2nd October 2016–30th September 2018 and 1st October 2018–27th September 2020 for the train, validation and test set respectively, as this split ensures that two peaks remain for the validation and test set. In addition to the influenza data, the hhh4-baseline model requires exogenous variables, which are also inserted into the other two models, to model the neighborhood and endemic effects. First, a binary neighborhood matrix $W \in \mathbb{R}^{411 \times 411}$ describing the direct connections between German districts. Moreover, a value of 1 corresponds to adjacent districts, whereas a value of 0 represents districts that are not connected. The values on the diagonal of the matrix are also 0, i.e. a district is not regarded as adjacent to itself. I obtained the matrix by determining each district’s neighbours within a shapefile from the year 2011 derived from the “Bundesamt für Kartographie und Geodäsie”⁷. Second, I also incorporated population data from 2011, as this reflects roughly the middle of the observation period, in form of a vector $e_i \in \mathbb{R}^{411}$ representing each German district. The population data was obtained from the “Statistische Bundesamt (Destatis)”⁸ and the “Amt für Statistik Berlin-Brandenburg”⁹.

4 Results

The forecasts for the validation and test period were determined through a rolling window approach, where I first determined forecasts for the next four weeks before I shifted the prediction window by one week into the future and repeated this prediction process for the whole forecasting period. After obtaining the forecasts, I report them in a quantile format with the selected quantile levels 0.025, 0.1, 0.25, 0.5, 0.75, 0.9 and 0.975. While I implemented the DeepAR and the FNN models in python with the GluonTs library, the baseline model was created utilizing the surveillance¹⁰ and hhh4addon¹¹ packages in R. The supplementary code is publicly available at the following GitHub repository¹². Before interpreting test results, I optimized the hyperparameters of the GluonTs models with consecutive grid searches. Table 1 depicts the parameters and validation mean WIS scores of the default models, which were simply initialized with the predetermined hyperparameters of the GluonTs library, with their hyperparameter-tuned variants, which performed the best on the validation split. Interestingly, besides the *use_feat_dynamic_real* parameter, which toggles the access to information about the calendar week, the remaining parameters to include additional information do not differ from their default value, i.e. *False*. Hence, the previously mentioned adjacency matrix and population information are not accessed by the default nor the tuned DeepAR and FNN models.

7 <https://gdz.bkg.bund.de/>

8 <https://www.destatis.de/>

9 <https://statis.statistik-berlin-brandenburg.de/webapi/>

10 <https://surveillance.r-forge.r-project.org/>

11 <https://github.com/jbracher/hhh4addon>

12 https://github.com/Kai-Ref/DeepAR_InfluenzaForecast

Table 1: Hyperparameter-optimization results on the validation set. The respective tuned models with the lowest validation WIS scores where chosen. Note that only paramters differing from the default parameter choices are included.

Models		Hyperparameters				WIS
FNN	hidden_dim	num_batches	context_length	epochs	batch_normalization & -_size	
Default	[40, 40]	50	4	100	False & 32	514
Tuned	[40, 40, 40]	60	104	200	True & 124	381
DeepAR	num_cells	num_layers	context_length	epochs	use_feat_dynamic_real	
Default	40	2	4	100	False	452
Tuned	140	6	2	200	True	353

Table 2: Summary of the average WIS scores of the default and tuned DeepAR and FNN models as well as the hhh4 model on the test data. The lowest scores are highlighted.

Models	Week Ahead Forecast				
	1	2	3	4	Average
Default DeepAR	238.90	338.59	434.09	495.71	376.82
Default FNN	245.67	369.00	512.40	689.82	454.22
hhh4	278.02	328.18	391.24	456.23	363.42
Tuned FNN	282.97	375.63	468.51	536.29	415.85
Tuned DeepAR	236.76	324.91	377.16	411.60	337.61

Table 2 shows the mean WIS of the five models, in which we can pick out the tuned DeepAR model as the best out of the five models, since it has the lowest mean WIS score for each week ahead forecast. The default DeepAR model is outperformed by the baseline model for the 2- to 4-week ahead forecast. Similarly, the default FNN is outperformed by the baseline approach, which suggests that the default models perform well on short-term forecasts but are outperformed by the baseline on longer forecasting horizons. Furthermore, it seems that the tuned models mainly improve the long-term forecasting performance, as the margins of improvement are relatively small for lower week ahead forecasts but increase for longer forecasting horizons. Interestingly, the tuned FNN performs worse than the default FNN on the 1- and 2-week ahead forecasts, suggesting that the tuned FNN sacrifices on short-term forecasting performance to increase its long-term predictive ability. The 3-week ahead performance of the FNN is worse than the 4-week ahead performance of all other models except the tuned FNN. Likewise, the tuned DeepAR and the baseline model perform better on the 4-week ahead predictions than the tuned FNN on the 3-week ahead predictions. Lastly, the tuned DeepAR 4-week ahead forecast also outperforms the default DeepAR’s 3-week ahead forecast.

To evaluate calibration, Figure 1 illustrates the lower and upper coverages of all five models on the test set. Ideally, the coverage plots are closely aligned with the diagonal. Furthermore,

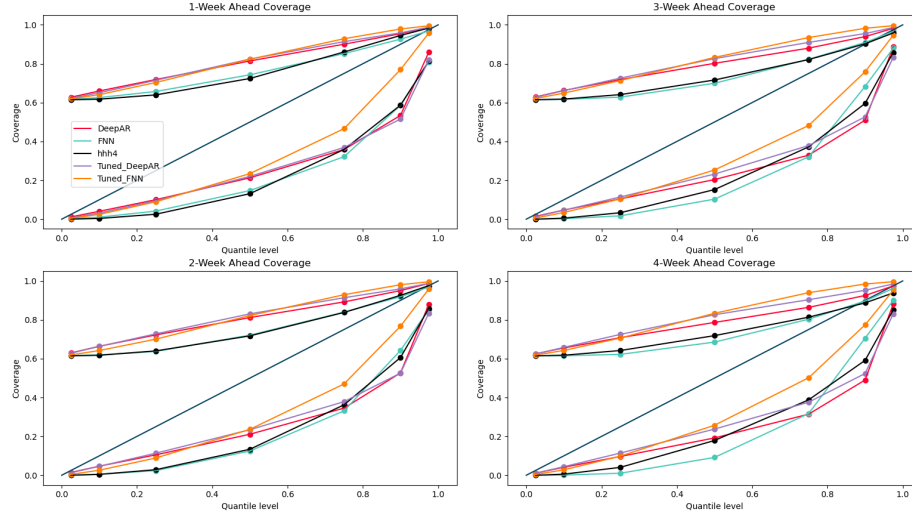


Figure 1: Lower and upper coverage plots of the test period for all five models. The diagonal represents the optimal coverage, whereas the lower and upper plot correspond to the lower and upper coverage respectively.

it is visible that the gap between both coverage plots is relatively large for every model but gets smaller as we move towards higher quantile levels. This large discrepancy likely originates from the high number of zeros between influenza seasons within the test data. Furthermore, due to the incorporated negative binomial output distributions, the model always predicts non-negative values. Thus, a materialized observation of zero can be interpreted as an observation that is always counted by the upper coverage version, as it solely predicts values greater or equal to zero. However, the lower coverage only takes zeros into account, when the predicted quantile is strictly larger than zero. On a different note, it is visible that the tuned FNN performs best for the lower coverage at quantile levels that are higher or equal to 0.5. It is also among the best for lower quantile levels, however the DeepAR models are often slightly closer to the diagonal. Besides that, both tuned models perform better than the default models, however the margin of improvement is larger for the FNN models than for the DeepAR models. Although the baseline and the default FNN model are the worst models for lower quantiles of the lower coverage, they perform better than both DeepAR models on higher quantile levels. In addition, the hhh4 and the default FNN model are the best models with respect to the upper coverage.

Considering another aspect, I visualized the regional 1-week ahead predictions on the test period of all five models in Figure 2. In this Figure I display predictions for the two smallest regions *SK Delmenhorst* and *LK Olpe*, the two largest regions *SK München* and *SK Hamburg* as well as *SK Karlsruhe* and *LK Kaiserslautern*. In general, the predictions of the DeepAR models look narrower than the predictions of the rest, but it is hard to assess which of the

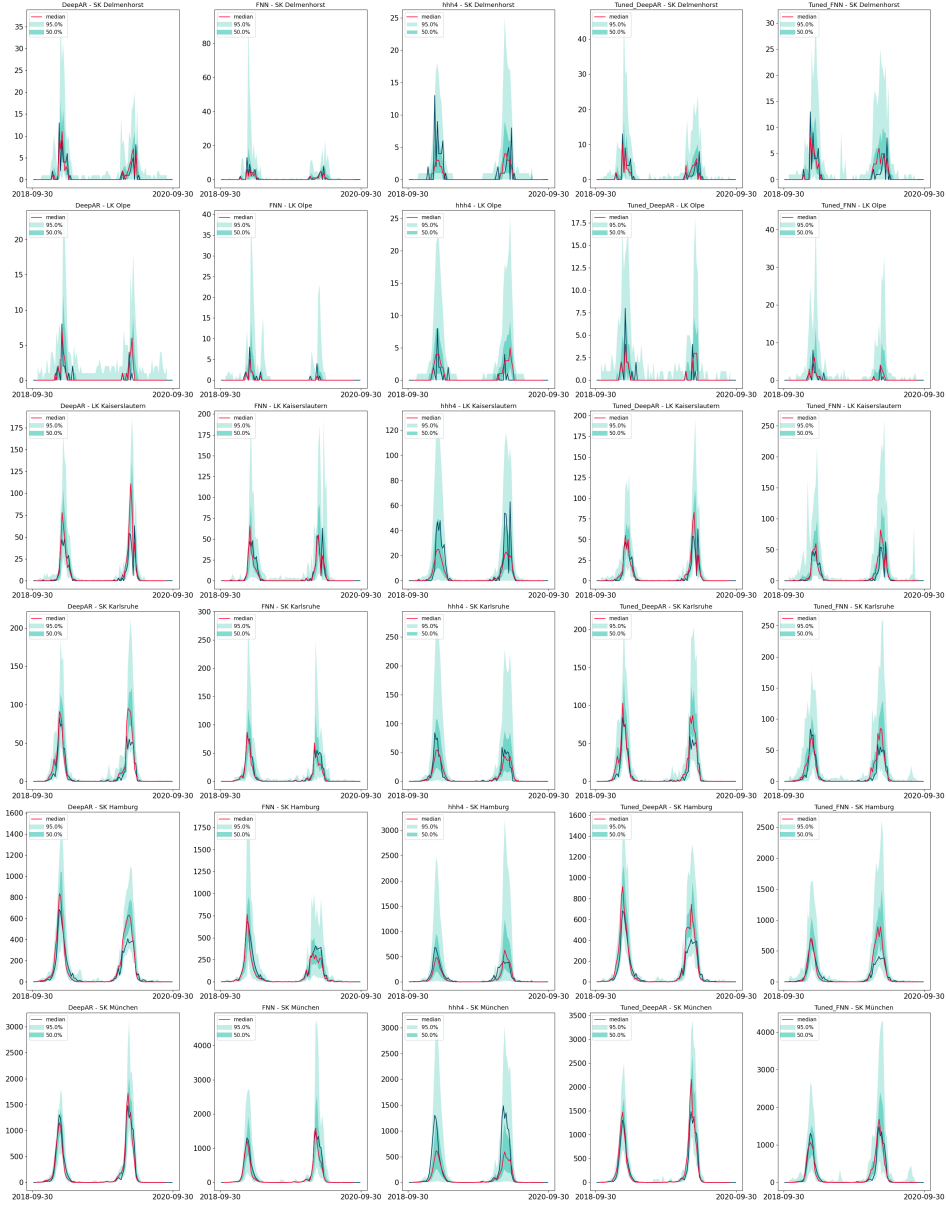


Figure 2: One-week ahead predictions of the default DeepAR, default FNN, hhh4, tuned DeepAR and tuned FNN model for selected regions on the test set. The red line represents the median of the predictions, while the dark and light green surfaces correspond to predictions that lie within the 50th and 95th prediction interval respectively. The blue lines equal the true underlying observations.

two DeepAR models produces the sharpest forecast. Furthermore, the default DeepAR model produces comparably larger PIs for periods outside peak regions, e.g. the predictions of *LK Olpe* are almost always larger than zero for the 95% PI. Furthermore, the hhh4 model best avoids this mistake, as its PIs are often zero for periods in between peaks. However, a flaw of the hhh4 model is that its median predictions are often too low, e.g. the median forecast for *LK Kaiserslautern* or *SK München*. Moreover, for *SK München* the 50% PI is also too low. In contrast, the other four models all produce more suited median forecasts for *SK München*. Although the baseline model does not seem to perform as poorly for *SK Hamburg*, it still produces the largest PIs out of the five models for this region. Overall, this suggest that the hhh4 model is potentially unfavorable for regions with relatively high case numbers. This is also supported by the fact that the lower bound of the baseline model's 95% PI is regularly close to values of zero during peak periods, hence its 95% PI ends up wider than those of the competing models. Aside from that, the 2020 peak of the time series of *SK Karlsruhe* and *SK Hamburg* seems difficult to forecast for some models, as they produce median forecasts and 50% PIs that are larger than the true underlying observations. Solely the default FNN and the baseline model avoid this mistake.

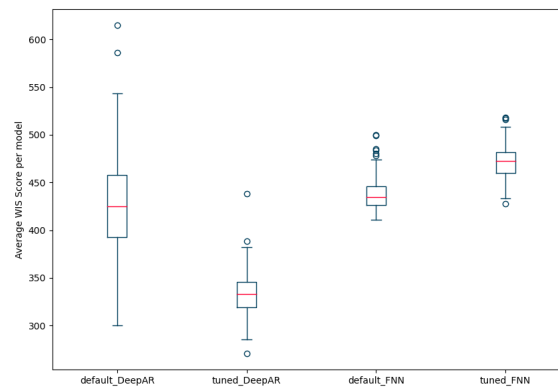


Figure 3: Illustration of the average one- to four-week ahead WIS Scores of 100 model runs on the test set of the default DeepAR, tuned DeepAR, default FNN and tuned FNN.

Figure 3 illustrates the WIS for 100 model runs of the tuned and default models. It indicates that the tuned DeepAR model is the best model out of the four models, as it has the lowest median, maximal and minimal WIS value out of the four models. In contrast, the tuned FNN model is the worst model for the same reasons. Altogether, the main results can be summarized as follows:

DeepAR outperforms the other models. The DeepAR model benefits from the tuning process and outperforms the other four models it competed with.

HP-tuning mainly improves longer forecasting horizons. Although the performance of the tuned DeepAR model improves for every week ahead forecast, the improvement

is more significant for longer forecasting horizons. Similarly, the tuned FNN's performance increases for longer forecasting horizons, however its short-term forecasts are even worse than those of the default FNN.

The hhh4 baseline is stable, but unprecise. During peak times, the predictions of the hhh4 model are not sharp, however it is not fluctuating as much compared to other models during low incidence times.

Lastly, I shortly want to highlight a few points, that further research studies could take into consideration:

Utilizing time series of similar diseases. The global modelling approach of the DeepAR model might also benefit from similar epidemiological time series, e.g. influenza time series from other countries or time series of other diseases through multi-pathogen forecasting. In fact, numerous studies suggest that the meningococcal disease has an impact on influenza infections [Je03].

DeepGLO. Although the DeepAR model is a global model, it does not incorporate information from other time series during inference. In contrast, Sen et al. [SYD19] introduced the DeepGLO model, which was designed to use all available time series during training and inference. Theoretically, this model could translate well to epidemiological time series forecasts.

Acknowledgements

I would like to express my deepest gratitude to Daniel Wolfram and Johannes Bracher for their exceptional supervision and guidance throughout the course of this thesis. Subsequently, I want to greatly appreciate Prof. Rainer Gemulla and Simon Forbat for their valuable discussions and insightful comments during the creation of this paper.

Bibliography

- [Ac12] Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.-H.; Liu, B.: Twitter Improves Seasonal Influenza Prediction. In: Healthinf. Pp. 61–70, 2012.
- [Br21] Bracher, J.; Ray, E. L.; Gneiting, T.; Reich, N. G.: Evaluating epidemic forecasts in an interval format. PLOS Computational Biology 17 (2), ed. by Pitzer, V. E., e1008618, 2021, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1008618.
- [Ca06] Casagrandi, R.; Bolzoni, L.; Levin, S. A.; Andreasen, V.: The SIRC model and influenza A. Mathematical Biosciences 200 (2), pp. 152–169, 2006, ISSN: 0025-5564, DOI: 10.1016/j.mbs.2005.12.029.
- [Cr22] Cramer, E. Y.; Ray, E. L.; Lopez, V. K.; Bracher, J.; Brennen, A.; Castro Rivadeneira, A. J.; Gerding, A.; Gneiting, T.; House, K. H.; Huang, Y.: Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. Proceedings of the National Academy of Sciences 119 (15), ISBN: 0027-8424 Publisher: National Acad Sciences, e2113561119, 2022.

- [En23] Engebretsen, S.; Palomares, A. D.-L.; Rø, G.; Kristoffersen, A. B.; Lindstrøm, J. C.; Engø-Monsen, K.; Kaminen, M.; Chan, L. Y. H.; Dale, Ø.; Midtbø, J. E.; Stenerud, K. L.; Ruscio, F. D.; White, R.; Frigessi, A.; Blasio, B. F. d.: A real-time regional model for COVID-19: Probabilistic situational awareness and forecasting. *PLOS Computational Biology* 19 (1), Publisher: Public Library of Science, e1010860, 2023, ISSN: 1553-7358, DOI: 10.1371/journal.pcbi.1010860.
- [Fe05] Ferguson, N. M.; Cummings, D. A.; Cauchemez, S.; Fraser, C.; Riley, S.; Meeyai, A.; Iamsirithaworn, S.; Burke, D. S.: Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437 (7056), pp. 209–214, 2005, ISSN: 0028-0836, 1476-4687, DOI: 10.1038/nature04017.
- [GBR07] Gneiting, T.; Balabdaoui, F.; Raftery, A. E.: Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69 (2), pp. 243–268, 2007, ISSN: 1369-7412, 1467-9868, DOI: 10.1111/j.1467-9868.2007.00587.x.
- [Gn23] Gneiting, T.; Wolfram, D.; Resin, J.; Kraus, K.; Bracher, J.; Dimitriadis, T.; Hagemeyer, V.; Jordan, A. I.; Lerch, S.; Phipps, K.; Schienle, M.: Model Diagnostics and Forecast Evaluation for Quantiles. *Annual Review of Statistics and Its Application* 10 (1), pp. 597–621, 2023, ISSN: 2326-8298, 2326-831X, DOI: 10.1146/annurev-statistics-032921-020240.
- [HBB21] Hewamalage, H.; Bergmeir, C.; Bandara, K.: Global Models for Time Series Forecasting: A Simulation Study, 2021, arXiv: 2012.12485[cs, stat].
- [HHH05] Held, L.; Höhle, M.; Hofmann, M.: A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling* 5 (3), pp. 187–199, 2005, ISSN: 1471-082X, 1477-0342, DOI: 10.1191/1471082X05st0980a.
- [Je03] Jensen, E. S.; Lundbye-Christensen, S.; Samuelsson, S.; Sørensen, H. T.; Carl Schønheyder, H.: A 20-year ecological study of the temporal association between influenza and meningococcal disease. *European Journal of Epidemiology* 19 (2), pp. 181–187, 2003, ISSN: 0393-2990, DOI: 10.1023/B:EJEP.0000017659.80903.5f.
- [Ju22] Jung, S.; Moon, J.; Park, S.; Hwang, E.: Self-Attention-Based Deep Learning Network for Regional Influenza Forecasting. *IEEE Journal of Biomedical and Health Informatics* 26 (2), Conference Name: IEEE Journal of Biomedical and Health Informatics, pp. 922–933, 2022, ISSN: 2168-2208, DOI: 10.1109/JBHI.2021.3093897.
- [Lo05] Longini, I. M.; Nizam, A.; Xu, S.; Ungchusak, K.; Hanshaoworakul, W.; Cummings, D. A. T.; Halloran, M. E.: Containing Pandemic Influenza at the Source. *Science* 309 (5737), pp. 1083–1087, 2005, ISSN: 0036-8075, 1095-9203, DOI: 10.1126/science.1115717.
- [PH11] Paul, M.; Held, L.: Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in medicine* 30 (10), ISBN: 0277-6715 Publisher: Wiley Online Library, pp. 1118–1136, 2011.
- [Ra20] Ray, E. L.; Wattanachit, N.; Niemi, J.; Kanji, A. H.; House, K.; Cramer, E. Y.; Bracher, J.; Zheng, A.; Yamana, T. K.; Xiong, X.; Woody, S.; Wang, Y.; Wang, L.; Walraven, R. L.; Tomar, V.; Sherratt, K.; Sheldon, D.; Reiner, R. C.; Prakash, B. A.; Osthus, D.; Li, M. L.; Lee, E. C.; Koyluoglu, U.; Keskinocak, P.; Gu, Y.; Gu, Q.; George, G. E.; España, G.; Corsetti, S.; Chhatwal, J.; Cavany, S.; Biegel, H.; Ben-Nun, M.; Walker, J.; Slayton, R.; Lopez, V.; Biggerstaff, M.; Johansson, M. A.; Reich, N. G.: Ensemble Forecasts of Coronavirus Disease 2019 (COVID-19) in the U.S. preprint, *Epidemiology*, 2020, DOI: 10.1101/2020.08.19.20177493.

- [Sa20] Salinas, D.; Flunkert, V.; Gasthaus, J.; Januschowski, T.: DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* 36 (3), pp. 1181–1191, 2020, issn: 01692070, doi: 10.1016/j.ijforecast.2019.07.001.
- [SLG19] Soliman, M.; Lyubchich, V.; Gel, Y. R.: Complementing the power of deep learning with statistical model fusion: Probabilistic forecasting of influenza in Dallas County, Texas, USA. *Epidemics* 28, p. 100345, 2019, issn: 17554365, doi: 10.1016/j.epidem.2019.05.004.
- [SYD19] Sen, R.; Yu, H.-F.; Dhillon, I.: Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting. Publisher: arXiv Version Number: 2, 2019, doi: 10.48550/ARXIV.1905.03806.
- [Ta19] Tapak, L.; Hamidi, O.; Fathian, M.; Karami, M.: Comparative evaluation of time series models for predicting influenza outbreaks: application of influenza-like illness data from sentinel sites of healthcare centers in Iran. *BMC Research Notes* 12 (1), p. 353, 2019, issn: 1756-0500, doi: 10.1186/s13104-019-4393-y.
- [WHO2323] World Health Organization, 2023, url: [https://www.who.int/newsroom/fact-sheets/detail/influenza-\(seasonal\)](https://www.who.int/newsroom/fact-sheets/detail/influenza-(seasonal)), visited on: 05/30/2023.
- [Wi96] Winkler, R. L.; Muñoz, J.; Cervera, J. L.; Bernardo, J. M.; Blattenberger, G.; Kadane, J. B.; Lindley, D. V.; Murphy, A. H.; Oliver, R. M.; Ríos-Insua, D.: Scoring rules and the evaluation of probabilities. *Test* 5 (1), pp. 1–60, 1996, issn: 1133-0686, 1863-8260, doi: 10.1007/BF02562681.
- [Wu18] Wu, Y.; Yang, Y.; Nishiura, H.; Saitoh, M.: Deep Learning for Epidemiological Predictions. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '18: The 41st International ACM SIGIR conference on research and development in Information Retrieval. ACM, Ann Arbor MI USA*, pp. 1085–1088, 2018, isbn: 978-1-4503-5657-2, doi: 10.1145/3209978.3210077.
- [Wu20] Wu, N.; Green, B.; Ben, X.; O'Banion, S.: Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case, 2020, doi: 10.48550/arXiv.2001.08317, arXiv: 2001.08317[cs, stat].
- [Ya20] Yang, C.-T.; Chen, Y.-A.; Chan, Y.-W.; Lee, C.-L.; Tsan, Y.-T.; Chan, W.-C.; Liu, P.-Y.: Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources. *The Journal of Supercomputing* 76 (12), pp. 9303–9329, 2020, issn: 1573-0484, doi: 10.1007/s11227-020-03182-5.