International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Impact of Varying Datasets for Prediction of COVID- 19 Cases

Zakarya A Mohamed Zaki[1], Aisha Hassan Abdalla[2]

[1, 2]Department of ECE, Fac. of Eng., International Islamic Univ. Malaysia (IIUM), Jalan Gombak, 53100 Kuala Lumpur, Malaysia.
[1]Email: *zaltalib92[at]yahoo.com*
[2]Email: *aisha[at]iium. edu. my*

Abstract: ***COVID - 19 has been identified as a global pandemic, and many experiments are applying various numerical models to anticipate the virus's likely growth under development. It is responsible for the emergence of the highly contagious illness. It is impacting millions of people throughout the globe. It has created a change in the research community's orientations for identification, analysis, and control via the application of different statistical and predictive modelling methodologies. These numerical models are examples of decision - making techniques that depend significantly on data mining and machine learning to create predictions based on historical data. In order to make smart judgments and create strong strategies, policymakers and medical authorities need reliable forecasting techniques. These studies are carried out on a variety of small scale datasets including a few hundreds to thousands of records. This study uses different sets of datasets consisting of COVID - 19 instances recorded on a daily basis in Iraq, together with socio - demographic and health related attributes for the region. The primary goal of this research is to see what is the impact of varying datasets for daily forecasting of COVID - 19 instances using deep learning forecasting tools. The predictive modeling for daily COVID - 19 infection cases involved using neural network architectures like enhanced hybrid model built using a Convolutional Neural Network and a Long Short - Term Memory network (EH - CNN - LSTM. Prior to the modeling, appropriate procedures were used to prepare the data and identify any seasonality, residuals, and trends. The model is trained and tested on various splits of the dataset. It is discovered that the higher the amount of training data, the better the predicted performance. Mean Absolute Percentage Error (MAPE), Mean Squared Logarithmic Error (MSLE), and Root Mean Squared Logarithmic Error (RMSLE) are used to evaluate the predictive performance.***

Keywords: COVID - 19, Forecasting, Prediction, and Deep Learning

1. Introduction

The new coronavirus emerged in late 2019 is the consequence of exposure with the severe acute respiratory syndrome -

coronavirus - 2 (SARS - CoV - 2) (Samuel Lalmuanawma, 2020). It has expanded internationally from late 2019, resulting in a protracted epidemic. COVID - 19 dissemination is based on inter - individual physical proximity and breathing particle transfer. Corona virus are a broad virus group that have been linked to illnesses spanning from cold or flu to far more serious illnesses. There have been two more outbreaks caused by coronavirus and the most recent virus discovered in Wuhan, China, is known as SARS - COV - 2, and it causes COVID - 19 (Samuel Lalmuanawma, 2020).

The first case of an unidentified pneumonia was report in Wuhan, China on December 31, 2020. Since then, the frequency of corona virus kept increasing, along with the mortality rate. It took only thirty days to spread to the whole country (Samuel Lalmuanawma, 2020). Because COVID - 19 spreads between person to person, artificial intelligence assisted electronic gadgets (Bhaskar et al., 2020) can perform a critical role in stopping the virus's transmission. As the function of healthcare epidemiologists has grown, so has the prevalence of digital medical records. The growing accessibility of digital clinical information gives a significant potential in medicine including both research and pragmatic implementation to enhance healthcare.

Throughout the past years, machine learning has been used excessively in several problems including ecommerce (Rath, 2022), sports (Richter et al., 2021), and healthcare (Qayyum et al., 2020). Time series forecasting is also one of the main areas for machine learning algorithms (Ahmed et al., 2010), because efficient forecasting may lead to better trading returns and enhance utilization of healthcare infrastructure. Many researchers now a days are focusing on hybrid approaches such as CNN - LSTM as superior alternative for time series forecasting.

Most technologically advanced deep learning modelling techniques are based on ANNs, particularly CNNs, while they can incorporate probabilistic algorithms or latent constructs organized tier in deep generative designs like the endpoints in deep learning and deep Boltzmann automated systems (Salakhutdinov & Larochelle, 2010). Deep learning techniques can be used to solve unsupervised training problems (Károly et al., 2018). This is a significant advantage since unidentified input is much frequent than classified data. Deep belief networks (Hinton, 2009) are indeed an illustration of a deep architecture that may be learned unsupervised. If breadth of a deep neural network containing ReLU (Agarap, 2018) stimulation is higher than the incoming size, the system may estimate any Lebesgue integrable value (Burkill, 2004); if the dimension is less than or equal to the incoming dimension, the structure is not a universal probabilistic model.

Deep learning is where the probabilistic understanding comes from. It includes the optimization ideas of learning and evaluation, which are linked to matching and generalization, accordingly. The probability approach views the activating nonlinearity as a cdf. Dropout was introduced as a regularizer in neural networks because of the deterministic understanding. The deterministic approach was developed by scholars such as Hopfield, Widrow, and Narendra and promoted in questionnaires such as Bishop's.

2. Method

This section comprises of implementation, testing, and validation steps of the applied deep learning technique in this research and has a deep dive into all the phases of the workflow as well. The dataset used was filtered through pandas python library. All the rows were filtered out except the rows for Iraq. The features of the dataset were dropped using the same pandas library except for dates and new cases, as these were needed for building, testing and evaluating models.

The data was then converted to matrices using and NumPy arrays and split up between training and test sets. The features were then scaled using MinMaxScaler from scikit learn. Different models under evaluation in this work were trained using various classes and interfaces from scikit learn library. And then evaluated with various performance metrics provided by the same scikit learn library.

The dataset used is from WHO, COVID - 19 data, and 2022. This dataset includes daily entries from 24 - 02 - 2020 to 01 - 08 - 2022. The dataset includes description in the form of various columns about the country, region, and continent. Moreover, it has data for country wise new cases, deaths, cumulative cases, and cumulative deaths. It contains records for 216 countries and 890 days.

The dataset was read using pandas python library through its pandas. read_csv (Pandas, 2022) function and readily converted to time series. As only date, new cases of country Iraq were required to proceed with the research workflow, all the irrelevant data was dropped using the same pandas library. Below Figure 1 has the code snippet for reading and filtering dataset.

```
df = pd.read_csv('Iraq daily dataset.csv', usecols = ['date','new_cases', 'location'], header=0, infer_datetime_format=True,
parse_dates=['date'], index_col=['date'])
df = df.loc[df['location'] == 'Iraq']
```

Figure 1: Code for Reading and Filtering Data

Every machine and deep learning algorithm require data to be processed in some way to have better training and performance in terms of validation and prediction. Most of the machine and deep learning scientists consider this step to be even more crucial than building the model itself. Thus, it has been considered the most important step in the direction of model development.

The first step was converting the dataset into matrices with the help NumPy by using its numpy. array (Numpy, 2022). Figure 2 is how it is done by defining a function that returns two arrays.

```
def convert2matrix(data_arr, look_back):
    X, Y =[], []
    for i in range(len(data_arr)-look_back):
        d=i+look_back
        X.append(data_arr[i:d,0])
        Y.append(data_arr[d,0])
    return np.array(X), np.array(Y)
```

Figure 2: Converting data to matrices

Dataset was then split 80/20 with 80% in the training set and 20% in the test set for one set of models and 70/30 for the second one and 60/40 for the third one and 50/50 for the fourth set of models. Next, the train and test features were scaled using MinMaxScaler (Scikit) from scikit learn library as shown in Figure 3.

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler(feature_range=(0, 1))
trainX = scaler.fit_transform(trainX)
testX = scaler.transform(testX)
```

Figure 3: Feature Scaling

The following tools were used to carry out the research work swiftly with Python as the programming language.

a) Lenovo Ideapad, Windows 10 Pro, Processor Intel (R) Core (TM) i5 - 6200U, CPU[at]2.30GHz 2.40 GHz, RAM 20GB
b) Anaconda for python distribution.
c) Jupyter Notebooks hosted locally on server.
d) Pandas for data reading and processing.
e) NumPy for computations and matrices operations.
f) Scikit learn for feature scaling.
g) TensorFlow and Keras for deep learning.
h) Matplotlib and Seaborn for visualization.

3. Results

Eighty percent of the data is used to train the model, while the remaining twenty percent is used to evaluate the model's performance then the 70% of the data is used for training and 30% for testing after that sixty percent for training and fourty percent for testing and last experiment used 50% o the data for training and fifty percent for testing the performance of the model. When modeling using the train - test split, each of the four splits are trained and evaluated for prediction.

3.1 EH - CNN - LSTM

The training and testing loss of hybrid model's training is depicted in Figure 4 whereas the observed and predicted

values are plotted in Figure 5. Table 1 provides the evaluation results for training and testing data for three evaluation metrics.

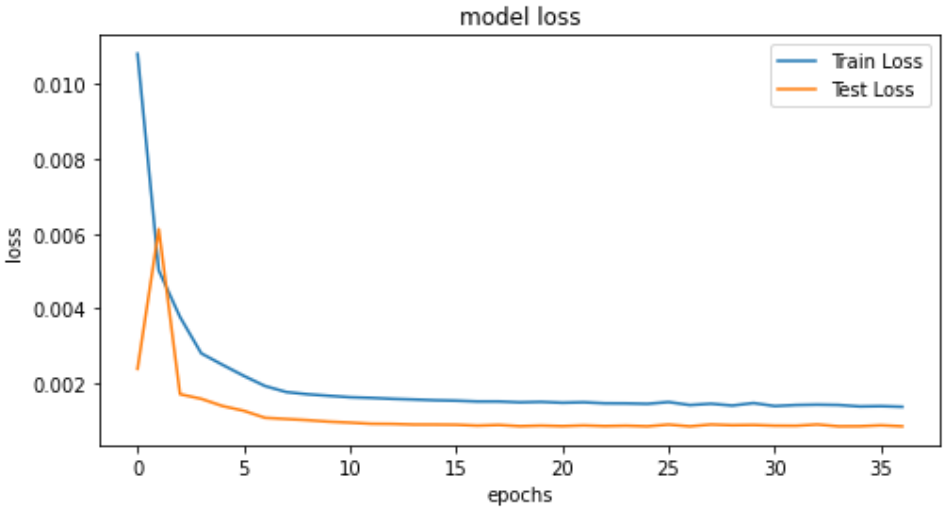

Figure 3: Training and testing loss of the hybrid model

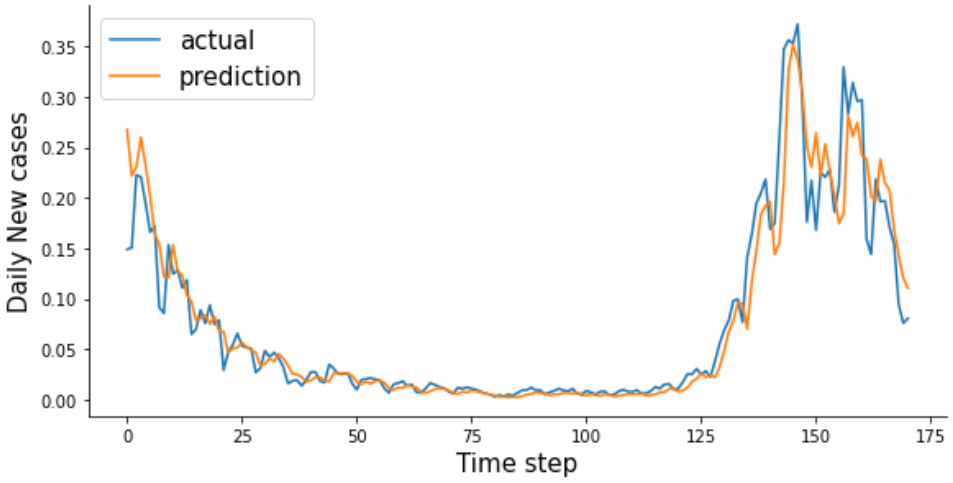

Figure 5: Plot of observed and predicted cases

Table 1: Performance of three evaluation metrics for training and testing sets

Evaluation metric	Training set	Testing set
Mean Absolute Percentage Error	70.06	5.28
Mean Squared Logarithmic Error	0.00	0.00
Root Mean Squared Logarithmic Error	0.03	0.02

3.2 Comparison of all data Splits

Table 2 summarizes the results of the all Scenarios and compares all of the potential models. The Scenarioal findings show that the hybrid model CNN - LSTM performs better when the training set is more than the testing set in predicting new instances of COVID - 19. Figures 6 - 9 show a similar pattern of predicted and observed instances for one month's forecast for all the datasets splits.

Table 2: Comparison of prediction model for all splits of datasets

Splits	Mean Absolute Percentage Error	Mean Squared Logarithmic Error	Root Mean Squared Logarithmic Error
50 - 50	9.39	0.00	0.04
60 - 40	5.25	0.00	0.03
70 - 30	5.67	0.00	0.03
80 - 20	5.28	0.00	0.02

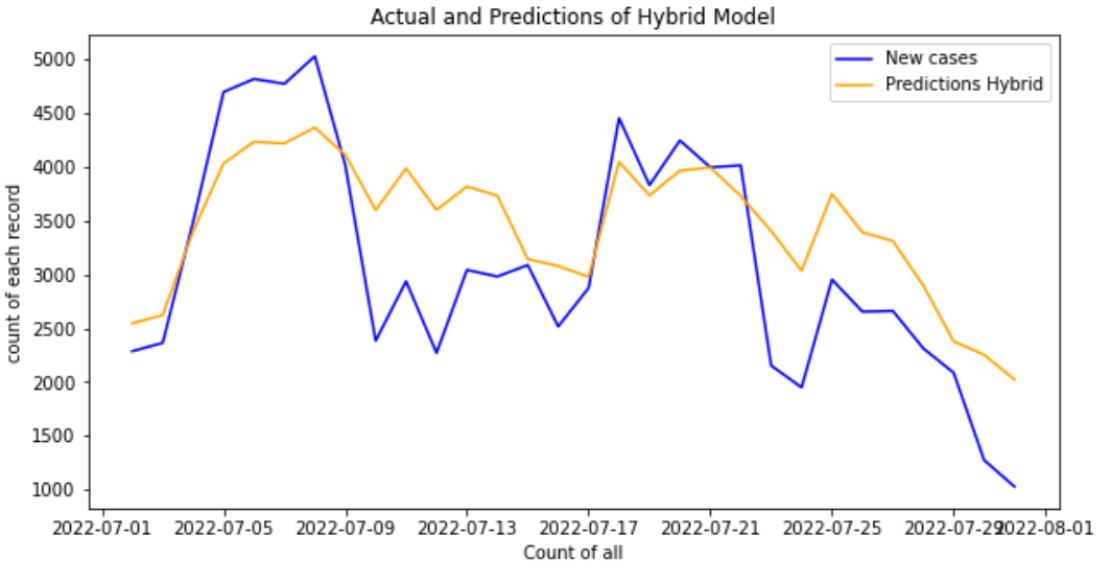

Figure 6: Plot of predictions and observations for first 50 - 50 split

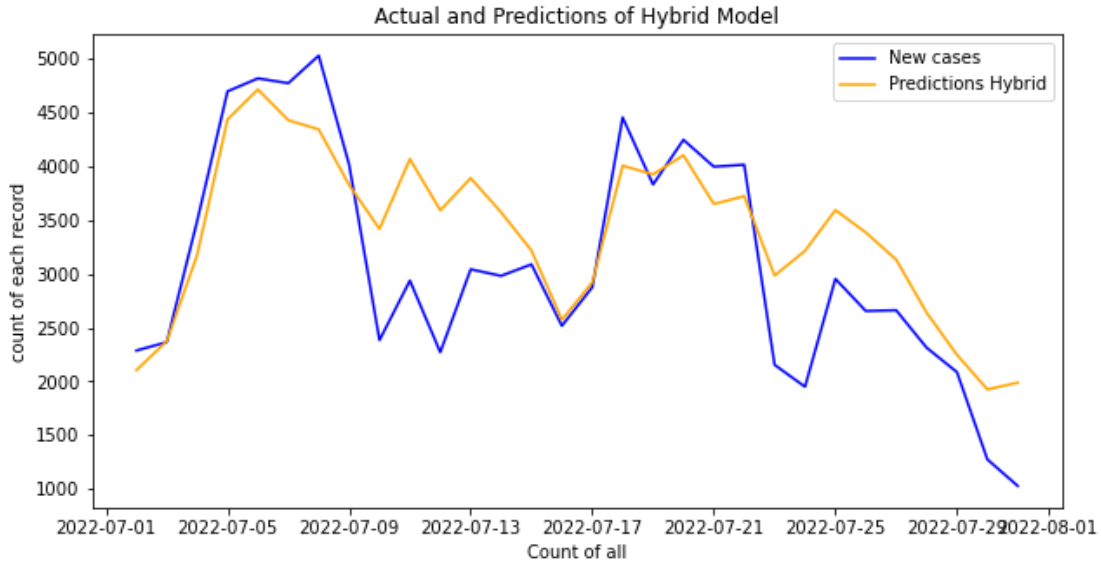

Figure 7: Plot of predictions and observations for first 60 - 40 split

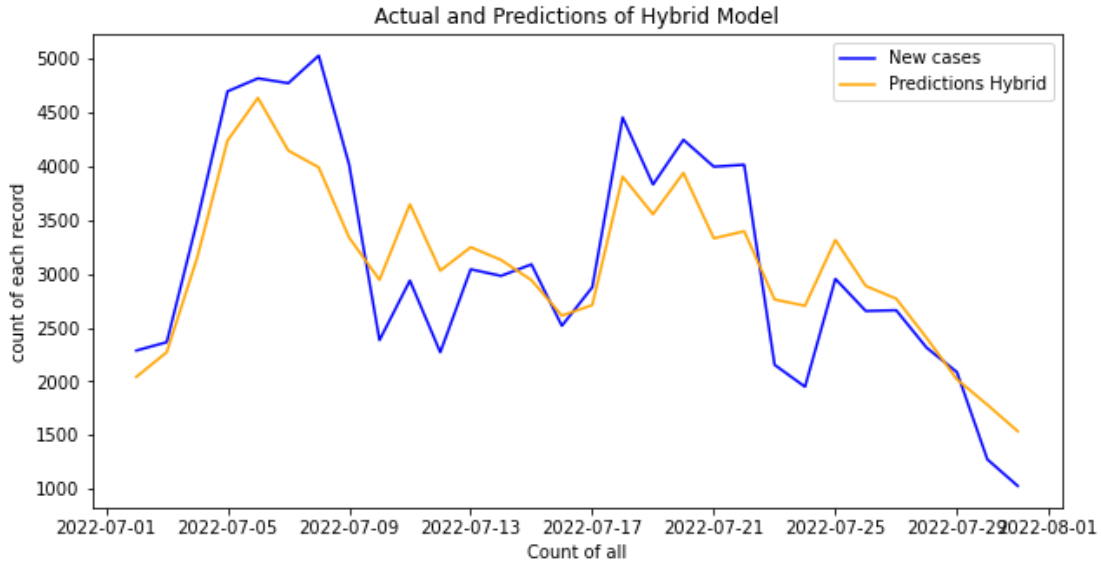

Figure 8: Plot of predictions and observations for first 70 - 30 split

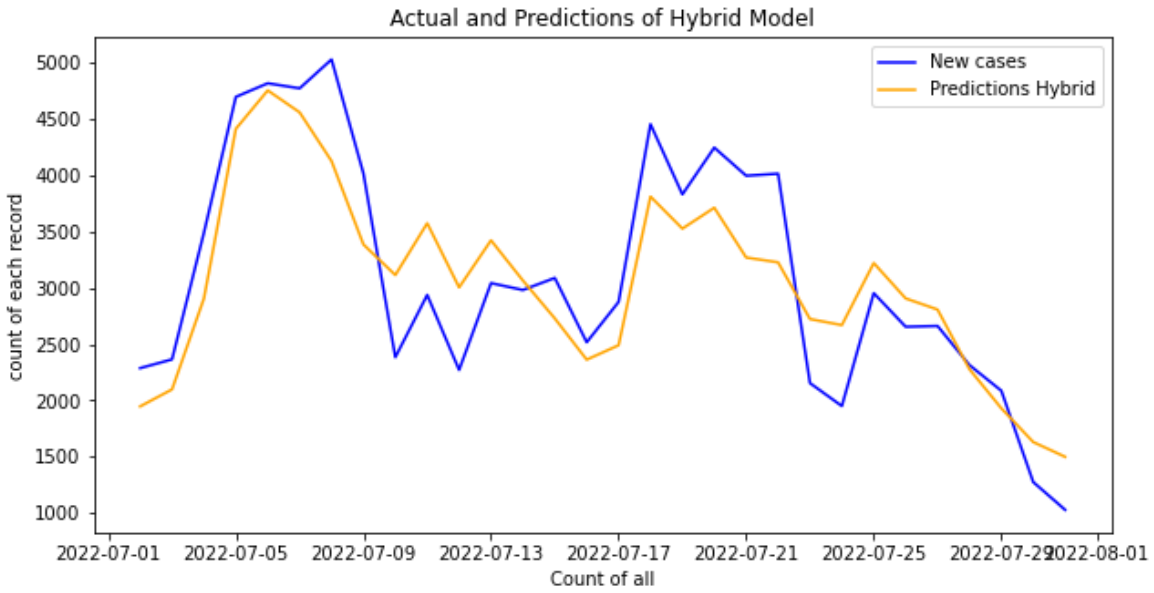

Figure 9: Plot of predictions and observations for first 80 - 20 split

4. Discussion

In order to get the Scenario results for the four different datasets splits, four different train - test splits of data were created for each scenario. The results using these four data splits are reported in the form of four scenarios, and they are helpful for evaluating the utility of a larger dataset size for improved prediction performance. The performance of the prediction algorithm changed as a result of a change in the train - test split, and the EH - CNN - LSTM method produced the best performance overall. EH - CNN - LSTM, which was trained on 80% of the data and evaluated on 20% of the data, achieved a MAPE of 5.28, MSLE of 0.00, and RMSLE of 0.02.

5. Conclusion

Disturbance to international commerce and economies all over the globe have resulted from the COVID - 19 pandemic. Governments and healthcare institutions are responding by taking aggressive actions to prevent the spread of disease. Each of these interventions may have a different effect on the spread and prevalence of COVID - 19 and the number of deaths caused by this virus, so it is essential to evaluate their relative effectiveness. In addition, the ability to foresee the influx of new cases enables policymakers and health experts to implement preventative measures in a timely manner. Prediction algorithms that can estimate the daily occurrence of instances can be developed with the use of machine learning, therefore facilitating the achievement of this goal.

The analysis of the dataset and the research requirements indicate that a time series forecasting algorithm can be trained to predict new cases of COVID - 19. A systematic methodology consisting of six stages is followed for research development. The data preparation is carried out after exploratory data analysis to transform it into a suitable for predictive modeling. To train a prediction algorithm, four datasets splits are trained and evaluated on the prepared dataset. The predictive performance of the model is especially effective when 80% of the data is used for model training and 20% is used for evaluation. The best performing model provided an MAPE of 5.28, MSLE of 0.00 and RMSLE of 0.02. The prediction of the model on new cases is performed and provided in the form of plot indicating trend of new cases and close prediction of the proposed model.

References

[1] Samuel Lalmuanawma, J. H., Lalrinfela Chhakchhuak. (2020). Applications of machine learning and artificial intelligence for COVID - 19 (SARS - CoV - 2) pandemic: A review. *Science Direct*.

[2] Bhaskar, S., Bradley, S., Sakhamuri, S., Moguilner, S., Chattu, V. K., Pandya, S., Schroeder, S., Ray, D., & Banach, M. (2020). Designing futuristic telemedicine using artificial intelligence and robotics in the COVID - 19 era. *Frontiers in public health*, 708.

[3] Rath, M. (2022). Machine learning and its use in e - commerce and e - business. In *Research Anthology on Machine Learning Techniques, Methods, and Applications* (pp.1193 - 1209). IGI Global.

[4] Richter, C., O'Reilly, M., & Delahunt, E. (2021). Machine learning in sports science: challenges and opportunities. *Sports Biomechanics*, 1 - 7.

[5] Qayyum, A., Qadir, J., Bilal, M., & Al - Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, *14*, 156 - 180.

[6] Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El - Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric reviews*, *29* (5 - 6), 594 - 621.

[7] Salakhutdinov, R., & Larochelle, H. (2010). Efficient learning of deep Boltzmann machines. Proceedings of the thirteenth international conference on artificial intelligence and statistics,

[8] Károly, A. I., Fullér, R., & Galambos, P. (2018). Unsupervised clustering for deep learning: A tutorial survey. *Acta Polytechnica Hungarica*, *15* (8), 29 - 53.

[9] Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, *4* (5), 5947.

[10] Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv: 1803.08375*.
[11] Burkill, J. C. (2004). *The lebesgue integral*. Cambridge University Press.