Machine Learning Approaches for Dengue Fever Prediction: Harnessing Evolutionary Algorithms

MachineLearner:
Justin Chiu, Joaquin Naverette, Audrey Chu, Alisha Mohsin
Center for Data Science
New York University
{jhc9986, jn2353, ac7816, am10823}@nyu.edu

Abstract

Dengue Fever, a mosquito-borne virus with transmission dynamics intricately linked to climate variables, is becoming more significant and pertinent amidst changing climatic conditions due to global warming. This study analyzes epidemiological and meteorological data in San Juan, Puerto Rico, and Iquitos, Peru, obtained from a DengAI competition, alongside curated data from Kaohsiung, Taiwan, to predict Dengue Fever incidence using multiple machine learning modes, with a specific focus on Neural Networks. Our results show that the NEAT (NeuroEvolution of Augmenting Topology) algorithm finds Neural Network architectures that best predict dengue fever cases. However, none of our models exhibit excellent generalizability, as their performances on the curated Taiwan data could be better. Future endeavors aim to further investigate the potential of the NEAT algorithm and enhance predictive accuracy across different regions worldwide.

1 Background

In typical cases, Dengue Fever causes a fever, rash, and joint pain. However, in severe cases, the virus can cause severe bleeding, low blood pressure, and death. Dengue Fever has an estimated mortality rate between 2 and 5% [8]. Since it is carried by mosquitoes, the transmission dynamics of the virus are intricately linked to climate variables [3], which are in turn influenced by climate change. The disease is most prevalent in tropical and subtropical regions like Brazil and Argentina, while it is rare in non-tropical regions like the continental United States. However, climate change is altering the location of the tropical belt. According to the Taiwan CDC, compared to 50 years ago, “The worldwide incidence of Dengue has risen 30-fold” [8]. Soon, many non-tropical areas may experience a similar climate, and this could drastically increase Dengue Fever cases in unexpected geographical locations, creating severe health implications for ill-prepared countries [4]. Moreover, understanding these relationships can aid in predicting and mitigating the impact of Dengue epidemics.

In the past, addressing Dengue Fever primarily involved reactive measures, such as vector control through the use of insecticides, environmental management to reduce mosquito breeding sites, and public health campaigns to raise awareness about preventing mosquito bites. However, these approaches often focused on controlling outbreaks after they occurred rather than proactively addressing the underlying factors driving the spread of the disease. Now, with a better understanding of the link between Dengue Fever transmission and climate variables, including temperature, precipitation, and humidity, there's a shift towards a more proactive and holistic approach [3]. This includes climate-informed surveillance and prediction, integrated vector management, cross-sectoral collaboration, community empowerment and education, and ongoing research and innovation [5]. By integrating these strategies, we aim to not only respond to outbreaks but also prevent them and build resilience in communities vulnerable to Dengue Fever.

2 Preliminaries

2.1 Data

The data utilized in this report comes from two sources. The first source is a DengAI competition [3] hosted on DrivenData, titled "DengAI: Predicting Disease Spread." This competition focuses on forecasting local epidemics of Dengue Fever in San Juan, Puerto Rico, and Iquitos, Peru, leveraging weekly meteorological data. Participants were given a training and testing set consisting of twenty predictor variables encompassing weather conditions, alongside the target variable: the weekly count of Dengue Fever cases for a given city. The Dengue surveillance data is provided by the U.S. Center for Diseases Control (CDC), the Department of Defense's Naval Medical Research Unit 6, and the Armed Forces Health Surveillance Center. The meteorological and climate data is provided by the National Oceanic and Atmospheric Administration (NOAA) [3]. The training data is given from 1990 to 2010, about 1400 rows, while the testing data includes 2008 to 2013. However, as the competition is ongoing, the number of cases for the testing data is not released. As a consequence, we had to split the training data they gave us into a testing and training set. This caused us to worry that the DengAI data alone would not be enough to be able to test our models and ensure that they are not overfitting. Therefore, we decided to create our own dataset for another city: Kaohsiung, Taiwan. We use the curated Taiwan data as a validation set to test the Out-of-Distribution (OoD) error of our models.

The weekly number of Dengue Fever cases in Kaohsiung was obtained using OpenDengue, a database of Dengue case counts for every Dengue-affected country globally since 1990. OpenDengue uses "a range of publicly available sources including ministry of health websites, peer-reviewed publications and other disease databases" [1]. The data for Taiwan specifically comes from the Taiwan Centers for Disease Control (CDC) under the Ministry of Health and Welfare. In tandem with this epidemiological data, meteorological information was acquired through the Visual Crossing Weather API [12]. For historical meteorological datasets, the platform relies upon the Integrated Surface Database (ISD) sourced from the National Oceanic and Atmospheric Administration (NOAA) [7]. The meteorological parameters utilized in our study were measured at the weather station located within Kaohsiung International Airport. Our final dataset contains data from 1998 through July of 2023.

2.2 Data Cleaning

We began cleaning our training data by examining the number of missing values in each column. Since the training dataset was not missing much data, we decided to drop any rows that did contain null values. After this, we looked at the correlations between variables, the correlation matrix is provided in Appendix A. We found a high correlation (> 0.7) between many variables. In light of this observation and anticipation of the computational demands of our evolutionary algorithms, we opted to execute Principal Component Analysis (PCA) to mitigate redundant information.

Our initial PCA analysis, conducted without removing any variables, identified that the highest eigenvalue was associated with the city designation, specifically San Juan or Iquitos. Since we wanted our models to be generalizable to any city, we decided to remove this column from our data and conducted a second PCA. The results of the second PCA are available in Appendix B. Following this refinement, we kept seven of the original twenty meteorological variables for model testing and training, resulting in an explained variance of 99.63%.

When obtaining the data for Kaohsiung, we matched only the seven columns retained post-PCA analysis. While the Dengue case count for Kaohsiung was given in a weekly format, the meteorological data was only available at a daily frequency. To synchronize the temporal granularity, we aggregated the weather data by computing the mean value for each column across a week. Subsequently, we retained only those weeks for which corresponding Dengue case information was available. This resulted in a dataset for Kaohsiung that mirrored our original training dataset exactly.

3 Models

3.1 Baseline Models

During our preliminary modeling phase for the Dengue Fever prediction, we examined the predictive capabilities of various baseline algorithms, including linear models, ensemble models such as XGBoost and Random Forest, and a primary Neural Network architecture. The Neural Network comprised three densely connected layers with ReLU activation, with 64 and 32 units in the first and second hidden layers, respectively. For our ensemble methods, we implemented GridSearchCV to find the optimal hyperparameters. Following this, XGBoost was configured with a learning rate of 0.1, a maximum depth of 3, and 50 estimators, while the Random Forest model had a maximum depth of 20, 'sqrt' for maximum features, a minimum of 4 samples per leaf, a minimum of 10 samples for split, and 200 estimators. We evaluated the models' performances using Mean Absolute Error (MAE), which we considered the most suitable metric for evaluation due to its holistic understanding of model performance and its resilience to the presence of outliers, which are common in the dataset. This approach allowed for a thorough comparison of the models' effectiveness in predicting Dengue Fever incidence, while also serving as comparison metrics for the novel NEAT approach implemented to forecast Dengue cases.

3.2 NEAT

In addition to the conventional Fully Connected Neural Network Architecture, we wanted to systematically explore the potential of alternative Neural Network structures. To accomplish this, we adopted the NEAT algorithm. NEAT (NeuroEvolution of AugmentingTopology) is an evolutionary algorithm proposed by Stanley and Miikkulainen [10] that incorporates the challenge of searching hyperparameters into its mechanism. This allows for the exploration of different structures while ensuring that the evolved networks are optimized for both structure and function. Furthermore, the algorithm has the capability of introducing non-traditional architectures such as Skip and Sparse connections. An overview of these two alternative structural choices is provided in Appendix A. In this paper, we leveraged NEAT in two distinct models:

Model A: Most of the mutation settings are enabled during the NEAT algorithm's execution, representing a more traditional NEAT usage approach.

Model B: Only the topology structure is allowed to be mutated throughout the evolutionary algorithm.

In Model B, for every generation, each member's topology information is passed into a PyTorch Neural Network Class that can dynamically create a Fully Connected Neural Network based on the provided information and train the weights through back-propagation. The motivation behind creating an additional design through Model B is to investigate the importance of alternative Neural Network architectures on the performance of this dataset.

In the following sections, we will explain the general settings of each phase within the algorithm. The detailed numeric settings are listed in Appendix D.

3.2.1 Population Initialization

Both models will begin with a population of simple Neural Networks, where the input layer is directly connected to the output layer. During each generation, each member of the population will carry a different genome object, representing a different network archetype.

3.2.2 Neural Mutation

The NEAT Python library requires the specification of the following factors' mutation rates:

1. Activation Function
2. Bias
3. Connection Add/Removal Rates
4. Node Add/Removal Rates
5. Response
6. Weight

In Model A, all factors except Factors 1 and 5 can be mutated. Factor 1 is disabled to ensure consistent comparison with Model B, which incorporates the ReLU activation function in its PyTorch Neural Network. Factor 5 is disabled as it is an external factor in the node function and does not affect the mutation process.

Conversely, only Factor 4 can be mutated Model B since it directly influences the network's architecture. The bias and response remain constant, while the weight is trained through backpropagation.

While the mutation rate and power are increased to ensure that the algorithm explores as many structure archetypes as possible, the remaining settings are unchanged from the example settings provided by NEAT's Python documentation [2].

This adjustment ensures that both models effectively explore the range of possible architectures while maintaining consistency and adhering to established mutation guidelines.

3.2.3 Elitism

Elitism is implemented to maintain continuous improvement within the current population by allowing only the top 20% of the population to "survive." This ensures that the fitness of the current generation surpasses that of the previous one. Furthermore, the genome structure of only the two best-performing individuals remains unmodified to preserve their advantageous traits, while the structures of the remaining individuals change.

3.2.4 Selection

To ensure valid comparisons between Models A and B and our baseline models, the fitness criterion is MAE. The evolutionary process is terminated when either 1) a member surpasses the Error threshold, indicating optimal performance, or 2) the generation count reaches its predetermined limit. This termination criterion ensures that the evaluation process is comprehensive and allows for effective comparison between different model iterations.

4 Results

4.1 Baseline Model Performance

All of our baseline models' performances, including the basic Neural Network, were first assessed on data from a test set derived through a train-test split on the DengAI data. We specifically examined the performance of our basic Neural Network compared to linear regression and ensemble baseline models. These comparisons were crucial in understanding the effectiveness of Neural Network-based approaches in predicting Dengue Fever incidence. Surprisingly, as can be seen in Figure 1, we found that the Neural Network exhibited the highest error on the testing set compared to the other models, despite neural architectures often being favored for capturing intricate relationships.

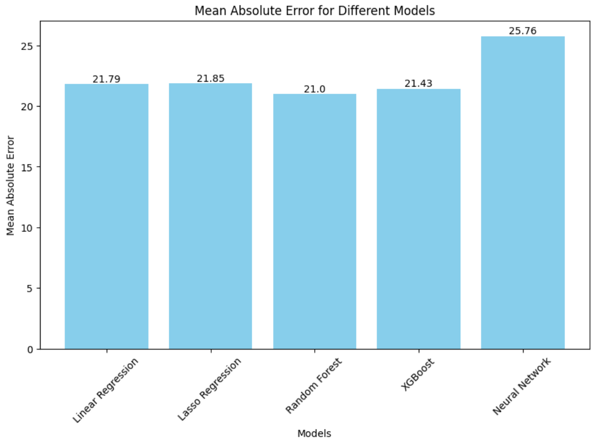

Figure 1: Baseline model Test Set Performance

4.2 NEAT Performance

Table 1 presents our NEAT-based models' performances on the same training and testing data derived from the DengAI competition that we used to evaluate the baseline models:

	Model A	**Model B**
Training Set MAE	18.09	17.78
Testing Set MAE	19.05	23.73
Generation Count	10	1
Time	5 minutes	30 seconds

Table 1: Results for NEAT-based Models A and B.

Training and Testing Set MAE: This column indicates the MAE for both the training and testing phases of the top-performing model.
Generation Count: This reflects the number of cycles the evolutionary algorithm required to yield the optimal solution.
Time: This denotes the total real-time duration needed for the algorithm to identify the winning model.

By setting the fitness threshold at 18.1, both Model A and B were able to find a "winning member" in a reasonable amount of time. We observed that Model B yielded a lower Training Set MAE for Model A, but Model B also leads to a higher Test Set MAE than MAE. This could indicate that Neural Network structures with Skip and Sparse connections have better generalizability over traditional Fully Connected Neural Network architecture.

We encountered a few drawbacks to the usage of NEAT. Setting the fitness threshold to 18.1 allowed us to obtain a winning structure within a reasonable time frame. However, as the best-recorded fitness value decreases, it becomes exponentially harder for the population to undergo a beneficial mutation that further reduces the MAE. When we attempted to set the MAE threshold below 17, both Model A and Model B took an extraordinary amount of time to compute. In Appendix C, Figure 6 shows the relationship between fitness and generation. While the algorithm might eventually discover a genome architecture meeting this lower threshold, time constraints and the absence of parallel computing resources prevented us from running the models for extended periods to see if this is true.

4.3 Out-of-Distribution Performance: Baseline and NEAT Models

To further evaluate our models' generalizability across diverse geographic regions and determine if the basic Neural Network's performance remained consistent, we conducted validation testing on our Taiwan data. This analysis offered valuable insights into the robustness and transferability of the models. There is a notable difference in the performance of models on the testing set with that on the Taiwan validation set. As illustrated in Figure 2, the Neural Network model exhibited an OoD MAE of 93.05 for the Taiwan data, the smallest error out of all of our baseline models – directly opposing our results from the test set. Meanwhile, the next best baseline model, Random Forest, had a 108.88 OoD MAE.

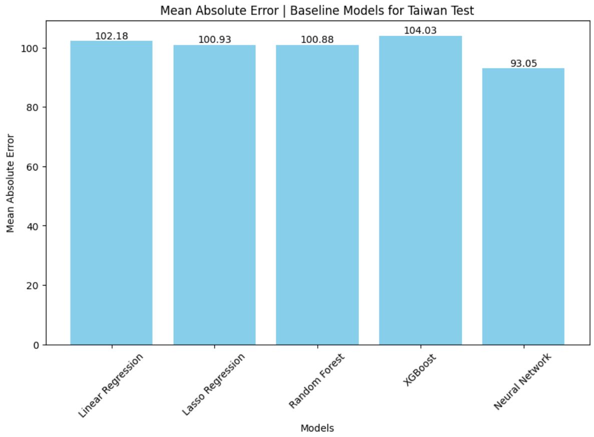

Figure 2: Baseline Out-of-Distribution Performance

This unexpected outcome suggests several possibilities: it could imply that Dengue transmission dynamics behave differently in different regions, highlighting a need for region-specific modeling, rather than a one-size-fits-all approach. It may also hint at potential complexity in the relationships between Dengue incidence and meteorological variables, which may vary across different geographical locations. However, there is also a possibility of discrepancies or errors in the Taiwan dataset, warranting further investigation into data accuracy and reliability. Nevertheless, neural networks appear to be a promising model for predicting Dengue cases. In the next section, we will delve into our findings from NEAT and discuss potential architectures.

Similar to the baseline models, the best models produced by NEAT show limited generalizability. Model A had an OoD MAE of 87.49, while Model B had an OoD MAE of 89.48. Despite MAE being far higher than we would like, both NEAT models exhibit superior generalizability compared to all of our baseline models. This highlights the potential of the Neural Network architectures suggested by the NEAT algorithm, indicating that it warrants further exploration. It's also important to note that Model A also outperformed Model B on this OoD test as it did on the testing set, suggesting that a traditional, Fully Connected Neural Network model may be preferable for Dengue Fever incidence prediction in future studies.

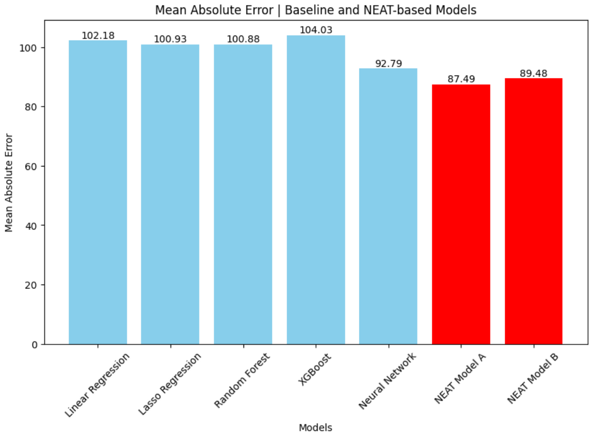

Figure 3: Baseline and NEAT OoD Performances

The generally suboptimal performance across all of our models can likely be attributed to our exclusive focus on environmental factors, overlooking human activities. For example, due to its geographic location, Taiwan receives a significant number of international tourists from other Dengue-endemic regions in Southeast Asia. This influx can introduce various strains of the Dengue virus into the country, complicating efforts to control the disease and leading to unpredictable spikes in cases. Additionally, the accuracy of our Taiwan data, sourced from a non-governmental organization, may also be a contributing factor to the poor model performance compared to the data obtained from the OpenAI Competition.

5 Conclusion

Our study has explored various machine learning models to estimate Dengue Fever cases and has attempted to generalize these findings using external data from Taiwan. We have limited the number of predictive features to climate-related variables, aiming to shed light on the impact of climate change on dengue fever cases. While we have achieved moderate success with NEAT in our initial dataset, relying solely on climate-related variables may not suffice for generalizing this model beyond the Latin American region. Furthermore, our study faces several limitations. Due to challenges in data collection, we were unable to utilize all variables strongly linked to climate, which could potentially enhance model performance. Additionally, we have yet to fully explore the capabilities of the NEAT algorithm due to constraints in resources and time.

6 Future Work

In future research, we aim to expand our investigation into the capabilities of the NEAT algorithm. Specifically, we plan to leverage parallel computation to reduce the time required for the algorithm to identify solutions with an MAE below 17. Additionally, we intend to explore the specific impact of skip and sparse connections by developing a model similar to Model B but with these connections explicitly integrated into the network architecture. This approach will allow us to understand how these connections alone can influence the performance and efficiency of neural networks. Furthermore, we plan to explore the possibility of including additional climate-change-related features to better characterize the severity of Dengue Fever, aiming to improve the generalizability of our model.

References

1. Clarke, J., Lim, A., Gupte, P., Pigott, D. M., van Panhuis, W. G., & Brady, O. J. (2023). OpenDengue: Data from the OpenDengue database (Version 1.2). Figshare. https://doi.org/10.6084/m9.figshare.24259573.

2. CodeReclaimers, LLC. (2019). NEAT-Python documentation (Version 0.92). Retrieved from https://neat-python.readthedocs.io/.

3. DrivenData. (2024). DengAI: Predicting Disease Spread Competition. Retrieved from https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/.

4. Hales, S., de Wet, N., Maindonald, J., & Woodward, A. (2002). Potential effect of population and climate changes on global distribution of dengue fever: An empirical model. Lancet, 360(9336), 830–834. https://doi.org/10.1016/S0140-6736(02)09964-6.

5. Jain, S., & Sharma, S. K. (2017). Challenges & options in dengue prevention & control: A perspective from the 2015 outbreak. The Indian Journal of Medical Research, 145(6), 718–721. https://doi.org/10.4103/ijmr.IJMR_1325_16.

6. Majeed, M. A., Shafri, H. Z. M., Zulkafli, Z., & Wayayok, A. (2023). A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention. International Journal of Environmental Research and Public Health, 20(5), 4130. https://doi.org/10.3390/ijerph20054130.

7. NOAA National Centers for Environmental Information. (2001). Global Surface Hourly Kaohsiung. NOAA National Centers for Environmental Information.

8. Taiwan Centers for Disease Control. (2024). *Dengue Fever*. Retrieved from https://www.cdc.gov.tw/En/Category/ListContent/bg0g_VU_Ysrgkes_KRUDgQ?uaid=9_Oq7OYHa-l8BO5iUwyVvQ.

9. Salim, N. A. M., Wah, Y. B., Reeves, C., et al. (2021). Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. Scientific Reports, 11, 939. https://doi.org/10.1038/s41598-020-79193-2.

10. Stanley, K. O., & Miikkulainen, R. (2002). Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation*, 10(2), 99-127. Available at https://nn.cs.utexas.edu/downloads/papers/stanley.ec02.pdf.

11. Tian, N., Zheng, J.-X., Li, L.-H., Xue, J.-B., Xia, S., Lv, S., & Zhou, X.-N. (2024). Precision Prediction for Dengue Fever in Singapore: A Machine Learning Approach Incorporating Meteorological Data. Tropical Medicine and Infectious Disease, 9(4), 72. `https://doi.org/10.3390/tropicalmed9040072`.

12. Visual Crossing Corporation. (2024). Visual Crossing Weather (1998-2023). [data service]. Available at `https://www.visualcrossing.com/`.

A Appendix

Appendix A

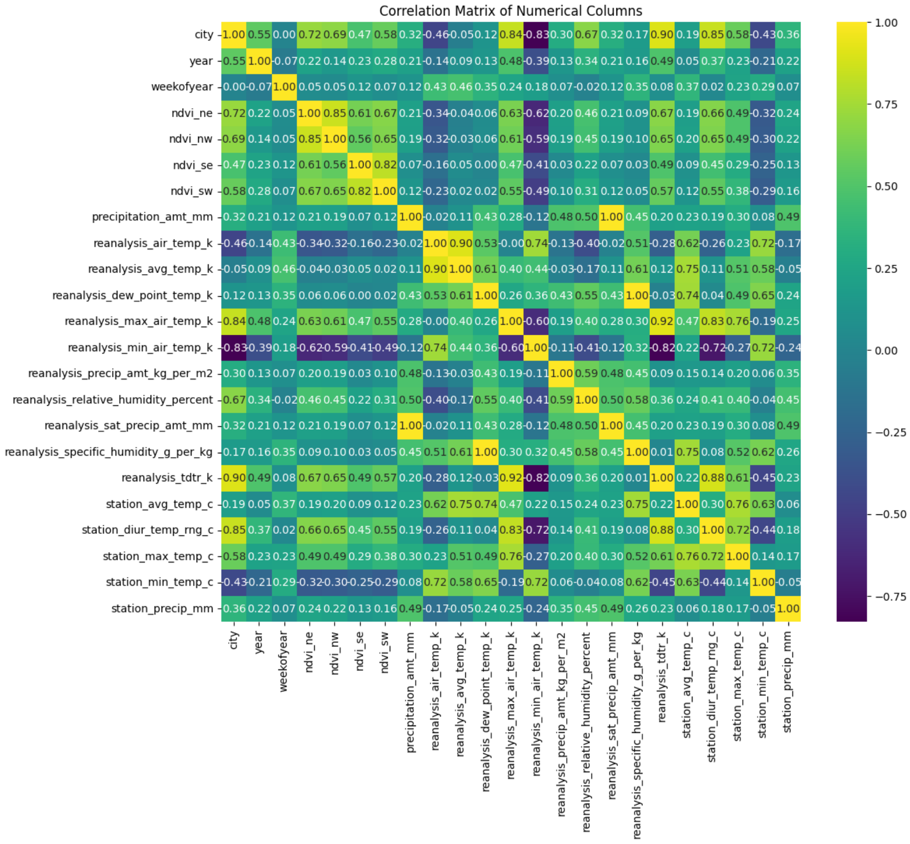

Figure 4: Correlation Matrix of Raw Data

Appendix B

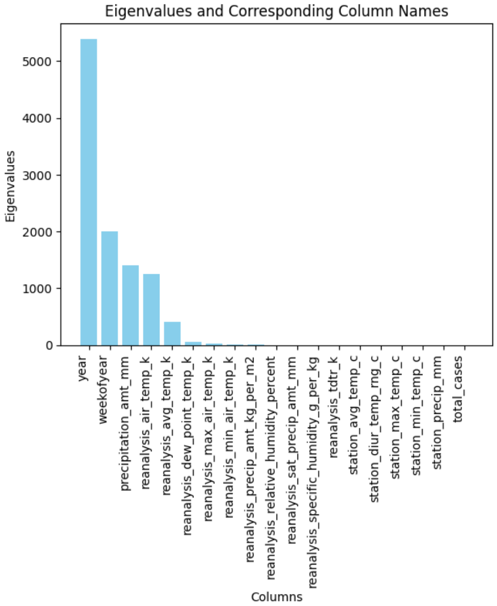

Figure 5: PCA Results

Appendix C

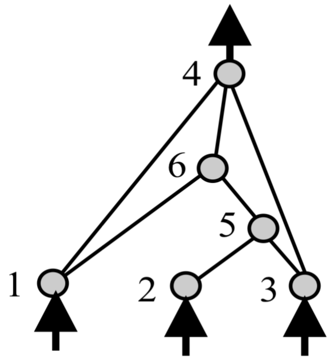

Figure 6: An alternate Neural Network Architecture

Figure 6 illustrates the use of skip and sparse connections within a neural network. Nodes 1, 2, and 3 serve as input nodes, while nodes 5 and 6 are located in the hidden layer, and node 4 is the output node. A skip connection allows a node to bypass adjacent layers and directly connect to a more distant layer; in this example, node 3 bypasses nodes 5 and 6 to connect directly to node 4. Sparse connections are characterized by the absence of connections where they would exist in a traditional Fully Connected Neural Network. For instance, node 2 does not connect to node 5, illustrating this sparse connectivity.

Appendix D

	Model A	**Model B**
fitness_threshold	-18.1	-18.1
pop_size	7500	250
activation_default	relu	relu
activation_mutate_rate	0.0	0.0
aggregation_default	sum	sum
aggregation_mutate_rate	0.0	0.0
bias_mutate_power	0.5	0.7
bias_mutate_rate	0.1	0.0
bias_replace_rate	0.0	0.0
conn_add_prob	0.5	0.0
conn_delete_prob	0.5	0.0
enabled_default	True	True
enabled_mutate_rate	0.01	0.01
node_add_prob	0.2	0.9
node_delete_prob	0.2	0.6
response_mutate_power	0.0	0.0
response_mutate_rate	0.0	0.0
response_replace_rate	0.0	0.0
weight_mutate_power	0.5	0.1
weight_mutate_rate	0.8	0.0
weight_replace_rate	0.1	0.0
elitism	2	2
survival_threshold	0.2	0.2

Table 2: The settings used to set up for model A and model B. This table does not present all the available settings, but only the settings that are of interest to our study here.

Appendix E

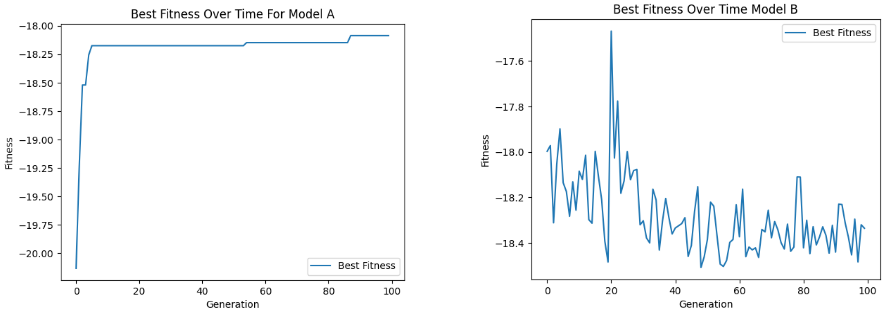

Figure 7: A Fitness vs Generation Graph

The graphs above demonstrate the best MAE fitness progression throughout 100 generations, with the MAE threshold set as -17. Neither Model A nor Model B were able to reach the threshold but with two different behaviors: We see Model A's performance monotonically decrease at an decreasing rate as the MAE gets smaller, while Model B's performance is more irregular but becomes progressively worse (Higher MAE) than previous generations.