Cancer Growth Forecasting

Introduction

Cancer proliferation is an inherently complex process, influenced by a myriad of factors such as the tumor microenvironment and the host immune system. Despite significant strides in cancer research, accurately modeling the temporal evolution of tumors remains a challenge. In this project, I sought to gain a deeper understanding of cancer growth dynamics that could enable decision-making for personalized cancer therapies. This study determines predictions using data from murine models of a metastatic variant of human triple-negative breast carcinoma. To examine tumor volume measurements over time, I applied three predictive models: Transformer, Artificial Neural Network, and Linear, to forecast the volume measurements for the next time step.

Methodology

I conducted a machine learning analysis using real tumor growth data sourced from murine experiments. This data involved a metastatic variant of human triple-negative breast carcinoma, as detailed in Vaghi (2020). The objective was to perform time series forecasting on tumor volume, for which three distinct models: a Transformer, an artificial neural network (ANN), and a linear model were implemented.

The purpose of these three models was to cover the gamut of bias variance tradeoff. In other words, I applied models ranging from the most general to the most specific. The Transformer, making the fewest assumptions about the data, was implemented to address potential nonlinear cancer growth trends. Conversely, I applied the linear model to characterize more straightforward, potentially linear aspects of the data.

The Transformer model, conceptualized in the seminal work "Attention is All You Need" by Vaswani et al., utilizes a neural network architecture that is particularly adept at handling complex non-linear sequential data. Its core feature, the attention mechanism, assesses the importance of each element of the input data in generating the output. Our architecture includes a causal Decoder, designed on the premise that previous data points influence subsequent ones. The model, trained to predict the nth data point given the previous n-1 points, is evaluated using mean absolute error (MAE). MAE is particularly useful in our study for managing the high variability at later time points, thus mitigating the impact of outliers on the model.

While the Transformer makes less assumptions than the ANN model, the ANN model however is still able to learn non-linear functions. The biggest gain from using the ANN model compared to the Transformer is that ANN is still able to learn non-linear functions while having increased data efficiency (i.e. the model can be fitted well onto a small data set). While the Transformer took into account all n-1 data points, the ANN model only took into account the very previous datapoint, further improving upon data efficiency. I expected the ANN model to perform best, as I predicted it would balance the bias variance tradeoff the best. In other words, I predicted the Transformer would overfit to the data, while the linear model would not be able to capture any non-linearities present.

Lastly, I implemented a linear model, conceptually simpler than its counterparts, that predicts outputs as a linear function between the last value and time sample. This model makes very strong assumptions about the underlying relationships between the data points over time. This makes it the most data efficient, but also makes it unsuitable for many datasets. For example, if the cancer growth was exponential and not linear, then the linear model would be incorrect.

Results and Discussion

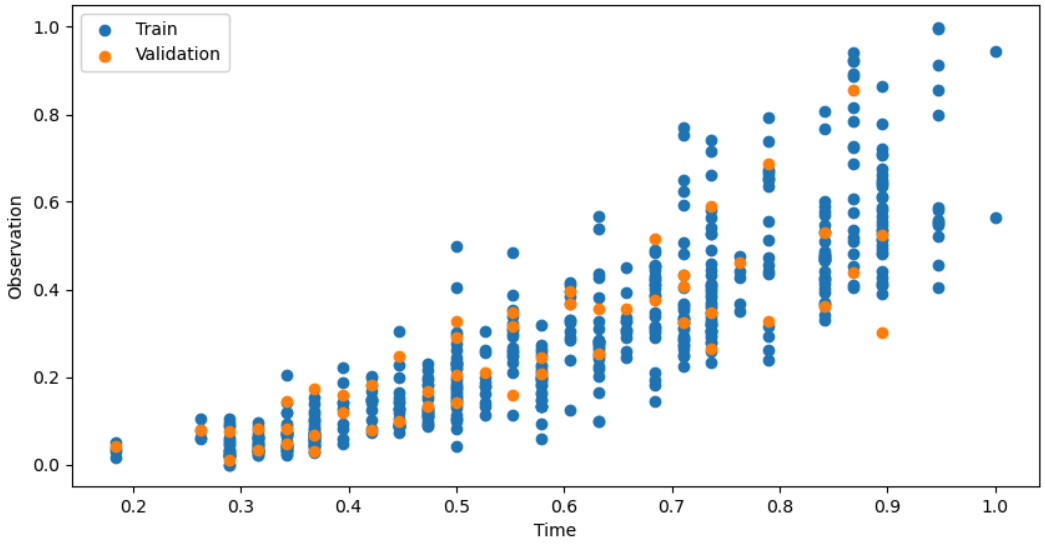

Figure 1 Normalized input data of training and validation Dataset

The data included 66 mice models where tumor volume (mm^3) with 5 time points range that corresponded to the day of tumor measurement post tumor implantation. Of these, 60 mice were used to create the training dataset, and the remaining 6 constituted the validation dataset. I normalized the cancer growth trajectories linearly for both datasets by dividing each data point by the respective maximum value (Figure 1). According to the MAE results on the validation dataset, ANN model outperforms both the Transformer and Linear (Table 1). This matches literature in that it has been shown by Zeng 2020 that Transformers are not reliable in time series forecasting in comparison to other models. Moreover, ANN outperforming meets our hypothesized expectation in that it would outperform both the general Transformer model and the specific linear model. The MAE of 83.49 for the ANN model indicates the average magnitude of error in predicting the subsequent time point based on previous data. To contextualize the MAE results, Figure 2 compares the predicted values from the ANN model against the actual values. This demonstrates the model's accuracy in closely mirroring the observed data and that a MAE of the ANN model is low. These results confirm our initial hypothesis that the ANN model would exhibit superior forecasting accuracy over the Transformer and linear models.

Table 1 MAE values of three difference machine learning model forecasting tumor growth over time

	Mean Absolute Error of Validation Data
Transformer	88.84
Artificial Neural Network	**83.49**
Linear	85.10

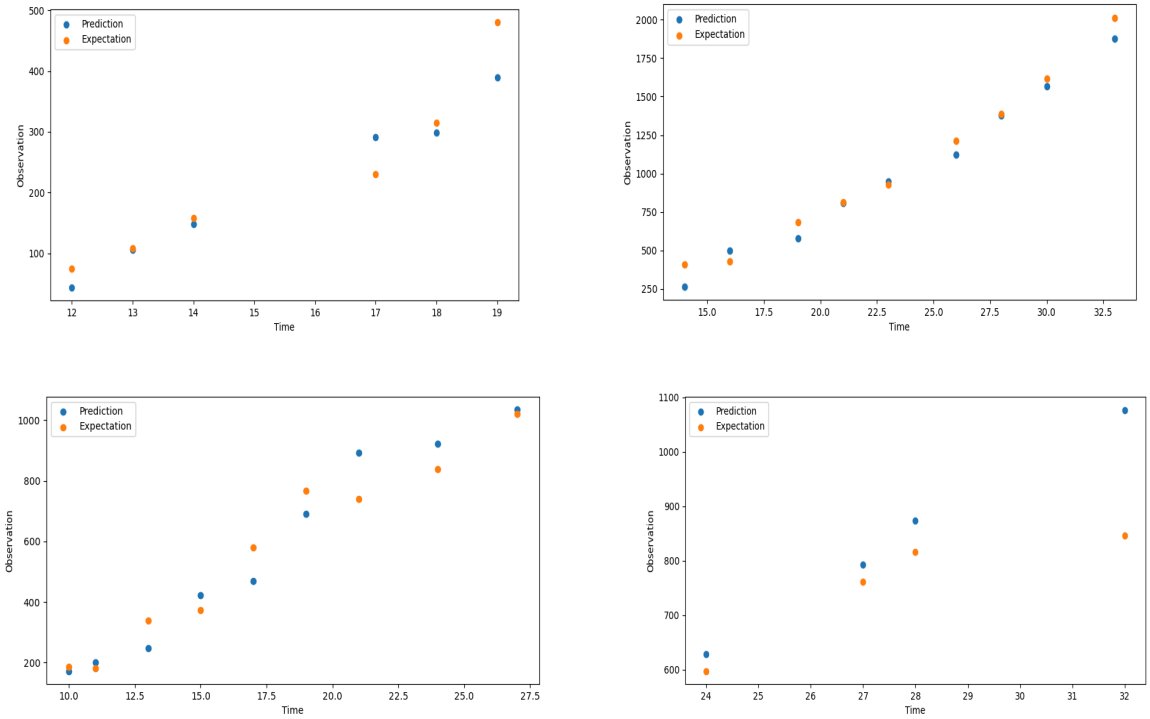

Figure 2 Prediction vs Expectation of volume of cancer tumor (observation) by time (days after implantation) for the ANN model

Citations:

1. Vaghi C, Rodallec A, Fanciullino R, Ciccolini J, Mochel JP, et al. (2020) Population modeling of tumor growth curves and the reduced Gompertz model improve prediction of the age of experimental tumors, PLoS Comput Biol, 16, p. e1007178.
2. Zeng, Ailing, et al. "Are transformers effective for time series forecasting?." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. No. 9. 2023.