Forecasting All-Cause Mortality: Leveraging Cause-of-Death Data through Neural Networks

School of Medical Sciences
UNSW Medicine and Health
University of New South Wales

Center for Big Data Research in Health
UNSW Medicine and Health
University of New South Wales

School of Risk and Actuarial Studies
UNSW Business School
University of New South Wales

Mamiya Adachi

Under supervision of:

Dr. Katja Hanewald

Dr. Andrés Villegas

August 5, 2022

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Health Data Science

Declaration

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signed: **Mamiya Adachi**
Date: 5th August 2022

Abstract

This research study aims to develop a new mortality forecasting model which leverages causes of death data through neural networks. We focused on developing a new convolutional neural network (CNN)-based model (referred to as the “new model” in this research study) since some past studies showed the CNN’s dominant forecast performance. We assessed the performance with the conventional Lee-Carter (LC) models and other CNN-based models with simpler structures. We used the publicly available US mortality data from 1959 to 2019 for model development and assessment. Our performance metrics for evaluating the forecasting models were based on the accuracy of the forecast, the duration for training and forecasting, and goodness of fit by forecast vs actual plots by age. We showed that the newly proposed CNN-based model is capable of outperforming the conventional LC models and CNN-based models with simpler structures. Further, we measured the time required for training both the LC and CNN-based models, and quantified the gain in accuracy at the cost of increased training time.

Acknowledgements

Among all the people who supported me during my research study, I would like to thank three people here since many have helped me during and outside the research time. I wish I could have expressed my gratitude to the others in person, but please allow me to make a simple greeting hereby.

I would like to thank my supervisors, Dr. Andrés Villegas and Dr. Katja Hanewald. It was my first time experiencing actuarial science-related research study, but I could survive joyfully with their incredible support. This experience is now important because I learned more options for my future path. I would like to also thank Dr. Bronwyn Haasdyk for supporting me as a course convener. Her feedback helped me throughout this research project.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation

According to the World Health Organization (2021), the percentage of the global population considered ageing (i.e. over 60 years old) is increasing. It is expected to reach 22% by 2050, with the ageing population trend previously observed in high-income countries now appearing significantly in low- and middle-income countries. Although enhanced longevity does signify more life opportunities, a large proportion of ageing populations will experience gradually decreasing physical and mental capacities (World Health Organization, 2021). This will potentially increase the cost of caring for this group, undermining the affordability of social security administration. This makes forecasting mortality rates and how they trend over time critical for social security planning and policymaking. Additionally, private-sector insurance companies need to forecast mortality to optimize their financial products for customer needs (e.g. price and corresponding reserves). Although previous studies (see Section 2) have introduced and assessed new forecasting models, these studies have not combined neural networks with cause-of-death information. Therefore, this study adopts that approach to develop a new mortality forecasting model which leverages causes of death data through neural networks.

1.2 Research aims and question

Our review of previous studies (details in Section 2) reveals that one significant gap in the literature is the limited use of cause-of-death as a feature variable during modelling. Another gap is the use of cause-of-death in neural-network-based models. Considering the feasibility of the research study within the limited timeframe, this research study aims to answer the question:

Can a convolutional-neural-network-based mortality forecasting model that includes cause-of-death as a feature variable outperform non-convolutional-neural-network-based models?

We respond to this question by developing several models based on models of the Lee-Carter (LC) type. We use US mortality data to train, validate, test and compare the models to assess their forecasting performance. We expect the convolutional neural network (CNN) model to perform without specifying the complex non-linear links between the feature variables. The autonomous detection of such links will improve the efficiency of the modelling process and minimize potential errors such as the unrecognition of the key causes of death. We have decided to focus on CNN models because of their demonstrated performance in previous studies (e.g. Perla et al., 2021). We use LC-type models as base models due to the short research study timeframe and the popularity and reliability of such models.

It is also worth mentioning that the US mortality data include cause-of-death data of sufficient quality, making it suitable for responding to the research question proposed.

1.3 Outline of the remaining sections

Section 2 reviews the results of previous research studies. Section 3 describes the methodology of our study. Section 4 describes the dataset used in our study. Section 5 describes the exploratory data analysis (EDA) of our dataset. Section 6 discusses the forecast performance of each model. Our new model partially outperforms the (conventional) non-CNN-based models and has the potential to yield even better forecast accuracy. Finally, Section 7 concludes the paper by further detailing the new model's performance and suggesting possible future research avenues.

2. Literature review

Various previous research studies have attempted to derive statistical models to explain and forecast how mortality rates change over time. These models have recently been integrated with different neural networks to improve forecasting performance. Assessment of selected previous research studies provides insight into what has been studied and what remains as gaps. Criteria for inclusion in this literature review are relevance to our research focus (see Section 1.1) and recentness.

Table 1 summarizes our review of these research studies, mapping four study features: the base model, the feature variable reduction method, the neural network and the feature variables.

	Lee and Carter, 1992	Wilmoth, 1995	Tabeau et al., 1999	Booth and Tickle, 2008	Cairns et al., 2009	Lyu et al., 2020	Perla et al., 2021	Richman and Wüthrich, 2021	C. Wang et al., 2021
Base model									
Lee-Carter (LC) model – original	✓			✓	✓	✓	✓	✓	✓
LC model – cohort extension (i.e. Renshaw and Haberman (RH))				✓	✓				✓
LC model – APC (age-period-cohort) extension (by Currie (2006))				✓	✓				
LC model – ACF (Augmented Common Factor) extension (i.e. Li-Lee model)						✓	✓	✓	
LC model – CAE (Common Age Effect) extension							✓	✓	
B-splines and P-splines (by Currie, Durban, and Eilers (2004))					✓				
Cairns-Blake-Dowd (CBD) model					✓			✓	✓
Heligman-Pollard model (including modifications)			✓	✓					
Gompertz model (including modifications)			✓						
Multiexponential models (*developed by various authors)				✓					
GLM				✓					
Neighboring prediction model									✓
Other		✓			✓				
Feature dimension reduction									
SVD (Singular Vector Decomposition)	✓			✓		✓	✓	✓	
PCA (Principal Component Analysis)				✓			✓	✓	
Embedding (a neural network layer)							✓		✓
Neural networks									
FCN (feed-forward fully connected neural network)							✓	✓	
RNN (Recurrent neural network)							✓		
LSTM (Long-short term memory)							✓		
CNN (Convolutional neural network)							✓		✓
Involved feature variable									
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓
Time (year)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gender		✓	✓			✓	✓	✓	✓

	Lee and Carter, 1992	Wilmoth, 1995	Tabeau et al., 1999	Booth and Tickle, 2008	Cairns et al., 2009	Lyu et al., 2020	Perla et al., 2021	Richman and Wüthrich, 2021	C. Wang et al., 2021
Cohort (birth year)			✓						
Area (country / region)			✓		✓	✓	✓	✓	✓
Cause-of-death		✓	✓	✓		✓			
Usage of sub-population (to forecast another sub-population)				✓		✓	✓	✓	✓

Table 1 **:** *The base models, feature variable reduction methods, neural networks and feature variable types used by each research study. Research studies are listed in chronological order. While the first six studies do not involve neural networks, the remaining three do. Components are marked as "assessed" ("✓") if assessed by that study or if the study refers to another study concerning the listed component and uses that as a basis for the method (modelling), discussion and conclusion.*

2.1 Base model

"Base models", or statistical models used in previous studies, can forecast mortality rates by themselves, but they can also be combined with neural networks, as shown in previous research (e.g. Perla et al., 2021).

Table 1 indicates that previous studies have often referred to the LC model and its extensions (i.e. LC-type models). Those models with neural networks also often use the LC-type models as a foundation. The frequent usage is due to the simplicity, interpretability and relatively promising performance of such models. This also means that non-LC-type models are less frequently used in research studies, producing a research gap. Additionally, it would be beneficial to conduct further studies about developing models with both the goodness of fit and forecasting accuracy, especially given several studies (e.g. Tabeau et al., 1999; Cairns et al., 2009) have concluded that no single model is clearly dominant over the others.

2.2 Feature dimension reduction (method)

The real-world datasets sometimes include various redundant features, which can increase computational cost, potentially slowing the process. To overcome such issues, feature matrices are often converted into simpler matrices that can still effectively represent the important characteristics of the original feature matrices (e.g. an age-time mortality matrix into the matrices of the smaller dimension and values using SVD).

Table 1 indicates that most of the previous studies, especially among those published post-2000, employ at least one feature variable reduction method regardless of neural network usage. This is reasonable because the number of original feature variables often exceeds the number of feature variables intended for consideration by the base models. Thus, there is no distinct gap in the recent research. Instead, using those listed feature reduction methods in any studies with modeling is essential.

2.3 Neural networks

Neural networks are a subset of machine learning. They are statistical models with a structure simulating how neural signals in the human brain are processed while learning (Bertolaccini et al., 2017). We can construct a neural network to capture the complex relationships between feature variables (e.g. age, gender, time (year), birth year and cause-of-death) in a manner similar to image recognition or text processing (Perla et al., 2021).

Although Table 1 shows that the usage of neural networks other than CNN represents a possible research gap, studies by Perla et al. (2021) and Wang et al. (2021) have already demonstrated the

superior capability of CNNs.

2.4 Feature variables involved in the study

None of the studies using neural-network-integrated models (i.e. Perla et al., 2021; Richman and Wüthrich, 2021; Wang et al., 2021) have employed cause-of-death data, making it a research gap. Additionally, there are some critical drawbacks to the usage of cause-of-death data in older studies (Tabeau et al., 1999; Booth and Tickle, 2008) due to the low quality and limited availability of cause-of-death data and the complex interdependence structure between causes of death (Lyu et al., 2020).

2.5 Contributions to the literature

As mentioned, our study aims to develop a mortality forecast model that can produce more accurate forecasts by combining a (convolutional) neural network with cause-of-death data. Using the causes of death as input for the neural network models addresses the extant research gap. We also assess the performance in terms of computation time for both the training and forecasting steps. Given our study's limited timeframe, we elaborate on potential avenues for future research (e.g. further investigation of optimal hyper-parameters, comparison of the performance of a greater variety of forecast models and performance assessment using data from different countries).

3. Methodology

Our data processing, modelling and analysis program has been developed using Python 3 on a Jupyter Notebook. The code is inspired by Brownlee (2021), Q. Wang et al. (2021), and Case Studies 6 and 9 from Schweizerische Aktuarvereinigung (n.d.). We store the Jupyter Notebook at https://github.com/MannyAdc/ForecastModel_LC_ML. The rest of this section describes exploratory data analysis, model development and comparison methodologies.

Furthermore, please note the versions of certain key packages used in this study:

- scikit-learn: 0.19.2
- statsmodels: 0.12.1
- TensorFlow: 2.5.0

3.1 Exploratory data analysis

We conduct the exploratory data analysis mainly graphically, and built on Villegas et al. (2021). The visualization includes age-standardized death rate (ASDR) over time by causes of death (CoD) and age-period-cohort (APC) mortality rate (age-specific) over time for all causes. We also conduct the primary data assessment and cleaning as detailed in Section 4 and the Appendix.

The study by Villegas et al. (2021) uses the same data as our study, and we have grouped causes of death (CoD) according to Table 3.2 of that research, yielding six broad groups of causes, namely, circulatory diseases, neoplasms, respiratory diseases, digestive system diseases, external causes, and other causes. We expect this grouping to enhance analysis interpretability and data processing efficiency.

3.2 Model development

We develop our study's models based on the four components mentioned in Section 2, yielding the main new model and three model variants for comparison summarized in Table 2. The variants differ according to the presence or absence of a CNN or cause-of-death data. Note that Variant 4 and 5 are

different only by hyper-parameters and how the training data have been applied (See Section 3.2.4 for details).

Model Component	New model	Variant 1	Variant 2	Variant 3	Variants 4 & 5
Base model	LC-type	LC-type	LC-type	LC-type	LC-type
Feature dimension reduction	Yes (embedding)	Yes (SVD)	Yes (SVD)	Yes (embedding)	None
CNN usage	Yes	No	No	Yes	Yes
CoD as a feature	Yes	No	Yes	No	No

***Table 2**: List of models developed for this study showing the model components included and excluded*

We use LC-type models as the foundation for our study. The mathematical definition of the most basic LC model is

$$\log(m_{x,t}) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}$$

Based on Lee and Carter (1992) and Richman and Wüthrich (2021), we define each term as follows:

- $m_{x,t}$ is the mortality rate at age x in year t.
- α_x is the average log mortality rate specific to age x.
- β_x is the rate of the log mortality change over time at age x. β_x is normalized to $\Sigma_x \beta_x = 1$.
- κ_t is the time index (year-to-year change) of the mortality in year t. κ_t is normalized to $\Sigma_t \kappa_t = 0$.
- $\epsilon_{x,t}$ is the error.

All the models except Variant 4 and 5 in Table 2 take the feature dimension reduction methods. We use those methods to fit the parameters and reduce the feature dimensionalities, as Lee and Carter (1992) and Richman and Wüthrich (2021) also mentioned.

3.2.1 Variant 1

Variant 1 is the simplest of the models developed. We develop and train the basic LC model separately for females and males using the following fitting and forecasting steps:

- Aggregate the cleaned data at the granularity levels of year and age.
- Split the aggregated data into training and test datasets based on year range (training: 1959–2007, test: 2008–2019).
- Fit the basic LC model using the training data.
  - We fit $\beta_x \kappa_t$ by decomposing (i.e. SVD) $\log(m_{x,t}) - \alpha_x$ under the assumption $\epsilon_{x,t} = 0$.
  - The decomposed matrices are noted as USV^T. Given the following characteristics, we define β_x as the first column of U, and κ_t as the first scalar of S multiplied by the first row of V^T.
    - The dimension of U is equivalent to the square of the length of age.

- ✧ The $\boldsymbol{S}$ is a diagonal matrix with dimensions given as the length of age (row) by the length of year (column).
- ✧ The dimension of $\boldsymbol{V^T}$ is equivalent to the square of the length of year.

- Obtain the $\boldsymbol{\kappa_t}$ forecast values and their 95% prediction intervals at each age for the test years using the **ARIMA(0, 1, 0)** model (equivalent to the random walk with drift (RWD)) proposed by Lee and Carter (1992).
- Obtain the $\mathbf{log}(\boldsymbol{m_{x,t}})$ forecast values of each age for the test years by calculating $\boldsymbol{\alpha_x + \beta_x \kappa_t}$ based on the values derived by following the above steps.
- Obtain the prediction intervals of the $\mathbf{log}(\boldsymbol{m_{x,t}})$ forecast values based on the previous steps.

3.2.2 Variant 2

Variant 2 includes CoD as a feature in addition to age and time (year). For this variant, we assume that each CoD is independent by fitting an independent LC model to each CoD. The modelling steps are as follows:

- Develop and train the basic LC model separately for each CoD and gender, following the same steps as Variant 1.
- Aggregate the $\mathbf{log}(\boldsymbol{m_{x,t}})$ of each CoD by summing those values for each CoD separately by age and gender.

3.2.3 Variant 3

Variant 3 is an LC-type model that uses a CNN. All our study's neural network models have been built using functions of the TensorFlow package (Abadi et al., 2016). The modelling procedure builds on the work of Perla et al. (2021) and Tam (2021) and is detailed as follows.

Data preparation

We aggregate the cleaned data at the granularity levels of year and age. Although CNNs are not originally intended for time series data (but instead for image processing), we expect a CNN to process time series data via its capacity to process multidimensional data with potentially complex correlations. Therefore, for training the models with CNN, we generate ten years of feature (input) data before the target (output) data (one year) by bootstrapping the data in the 11-year window from the dataset to allow our models to learn the time-series relationship of the data. It is worth mentioning that we have transformed the target data using the MinMaxScaler() function of the scikit-learn package (Pedregosa et al., 2011) in the range [0, 1]. We have decided to apply the transformation based on Perla et al. (2021) because MinMaxScaler() can scale the data but still output values outside the specified range (between 0 and 1) in cases where an outlier occurs. We have applied the scale separately for each gender to reflect the different scales of data for each gender.

Instead of addressing the data for each gender completely separately, we used both sets of data to train and validate only one model in order to grasping the potential correlations between the genders. We have arranged the gender data as a separate input, requiring the information to be in numerical form using the LabelEncoding() function of the scikit-learn package (Pedregosa et al., 2011).

We split the aggregated data into training, validation and test datasets based on the year range (training: 1959–1994; validation: 1995–2007; test: 2008–2019). The validation data is equivalent to 25% of all the data before the test years. We intend such data not only for validation but also to avoid using it for training to prevent data leakage. As mentioned by Natrajan et al. (2018), data leakage

between train and test datasets commonly occurs during the train/test split of time series data.

Model structure

As Figure 1 shows, Variant 3 features two separate branches for the different inputs (the log mortality rates and genders).

- Convolution branch (for the log mortality rate matrix)
  - The branch starts with the convolution layer after the input layer. The "(None, 10, 81)" shown at "rate_matrix" in Figure 1 represents the feature dimension the branch can take as an input. "None" indicates any dimension acceptable (i.e. the number of input matrices in our case), 10 indicates the number of years in an input matrix, and 81 indicates the number of ages from 20 to 100. The convolution layer creates the feature map from the input matrix using the specified dimension and number of filters. After running the model training multiple rounds, we decide the hyper-parameter values: the convolution filter number = 30, kernel size = 10, activation function = linear. Then, we normalize the feature data (log mortality rates) using the BatchNormalization() layer.
  - The next layer is the max-pooling layer, which further reduces the input matrix's feature dimension divided by the specified pool size. After running the model training for multiple rounds, we decide to use pool size = 10 (hyper-parameter).
  - The layer after max-pooling intentionally drops some of the nodes in the neural network for generalization at the rate = 10^{-2} (hyper-parameter).
  - To decide the hyper-parameters, we compare the forecast performance of each model (mean square error by years and forecast vs actual figures).
- Categorical variable branch (for gender variable)
  - This branch is for the numerically encoded gender data input from the "Data Preparation" process described. The critical layer in this branch is the embedding layer, which outputs another numerical vector representation of the original input. Embedding is a technique widely used to handle categorical variables. After running the model training for multiple rounds, our study uses 10 as the output dimension (hyper-parameter).

Next, the model has the section of the neural network layers that processes information from both branches. The first layer of the section concatenates the outputs from the two branches described, applying node-dropout rate of 10^{-2} (hyper-parameter). Comparing forecast performances with several activation functions leads to the choice of the activation function "linear".

Finally, the model is compiled after "forecast_rate (Reshape)" in Figure 1. For the optimizer setting, we use the ADAM (adaptive moment estimation) optimizer with the learning rate 10^{-3} (hyper-parameter at the default value), based on assessing the forecast vs actual figures from multiple rounds of model training. Alongside the learning rate setting, we run 300 epochs with 30 steps per epoch.

Generating forecast values

We obtain the $\mathbf{log}(\boldsymbol{m_{x,t}})$ forecast values of each age using the predict() function in a step-wise manner for the test years. More specifically,

- The first-year forecast is based on the actual data (initial input) in the window period (i.e. 11 years = 10 (input years) + 1 (output year)).
- The second-year forecast is based on part of the actual data (all except the first year of the

window period mentioned above) and the first-year forecast.

- All forecasts for later years are calculated recursively in the manner described but with a moving window period.

Please note the following:

- The raw output (forecast from the model) is back-transformed to the original scale from the [0, 1] scale at every forecast year.
- We do not obtain the prediction intervals for all the neural network models in our study due to time limitations. This represents a means of improving future studies.

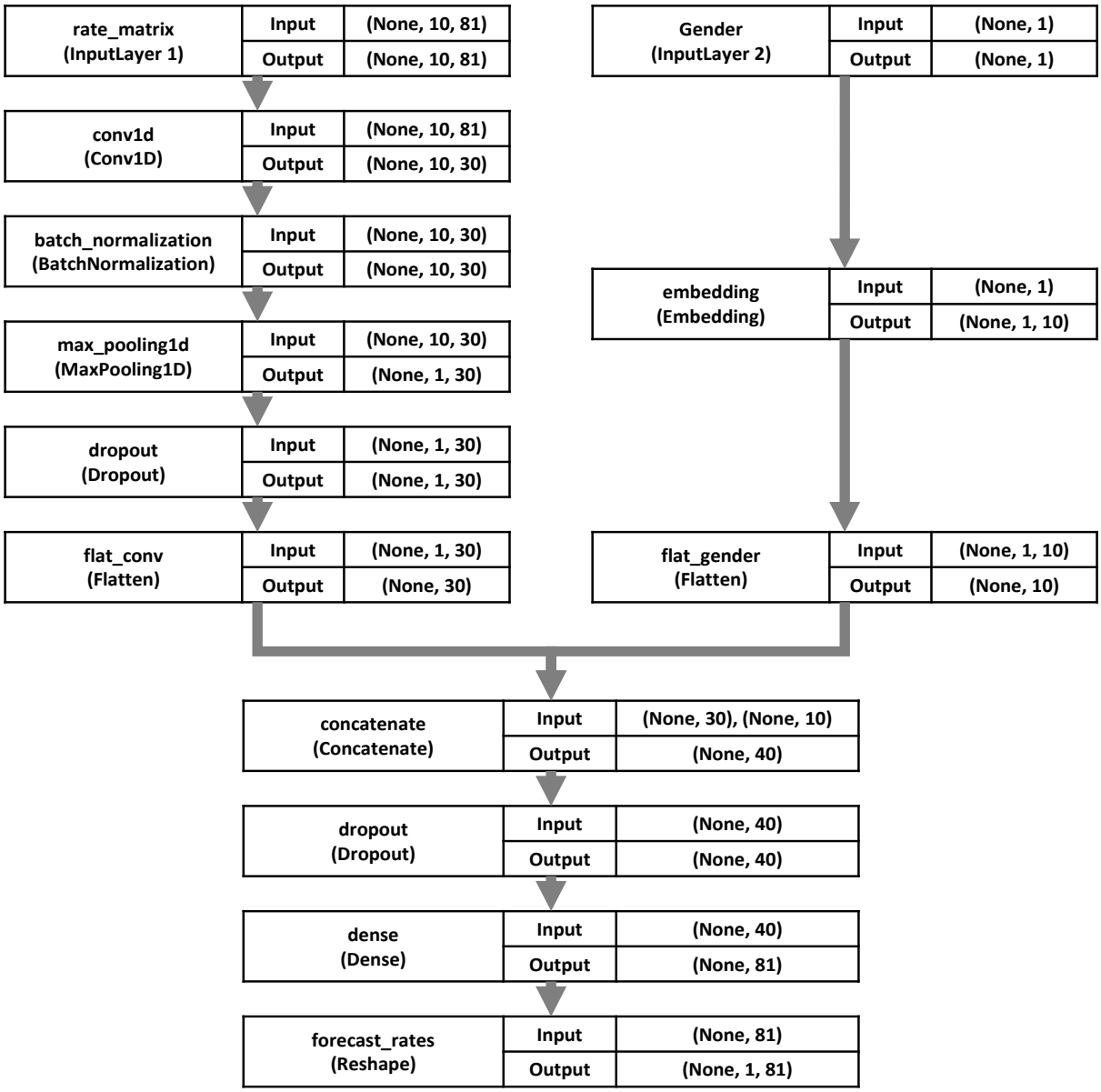


***Figure 1**: Graphic representation of the Variant 3 architecture*

Interpretation of the structure of neural network models and its relevance to the LC model

Perla et al. (2021) have described their conversion of the mathematical formula of a neuron in a neural network model at a specific age, time, region and gender into a formula resembling the basic LC model. Accordingly, the model includes time and region information in addition to age and time. We can apply this interpretation generally to any neural network model by adopting the mathematical formula of a neuron's operation $\boldsymbol{z_j(x) = \sigma(\omega_{j,0} + \langle W_j, x \rangle)}$, where $\boldsymbol{z_j(x)}$ is the output (i.e. the new feature to be passed to the ascending neurons), $\boldsymbol{x}$ is the input feature, $\boldsymbol{\sigma}$ is the activation function, $\boldsymbol{\omega_{j,0}}$ is the bias matrix, and $\boldsymbol{W_j}$ is the weight matrix.

Similar to Perla et al. (2021), our formula describes how a neuron calculates a specific age $\boldsymbol{x}$ and time $\boldsymbol{t}$ using CoD and gender information $\boldsymbol{i = (CoD,\ g)}$. Please note that we intentionally included CoD here to reflect our study aim of including CoD as a feature. $\boldsymbol{W_x}$ is the weight matrix such that $\boldsymbol{W_x = (W_x^{CoD}, W_x^g, W_x^f)}$, where $\boldsymbol{f}$ represents the mortality history, $\boldsymbol{z(CoD, g, U_t^{(i)})}$ is the concatenated information of CoD, gender and mortality rate matrix. Then, we write the mortality rate at the specific year, age, and $\boldsymbol{i}$ as $\boldsymbol{\hat{y}_{x,t}^{(CoD,g)}}$, yielding

$$\boldsymbol{\hat{y}_{x,t}^{(CoD,g)} = \sigma(\omega_{x,0} + \langle W_x, z(CoD, g, U_t^{(i)}) \rangle).}$$

We can rewrite that formula as,

$$\boldsymbol{\sigma^{-1}\left(\hat{y}_{x,t}^{(CoD,g)}\right) = \omega_{x,0} + \langle W_x^{CoD}, z_{CoD}(CoD)\rangle + \langle W_x^g, z_g(g)\rangle + \langle W_x^f, z_f\left(U_t^{(i)}\right)\rangle.}$$

$\langle \cdot,\cdot \rangle$ is a scalar product. We interpret the components $\boldsymbol{\omega_{x,0} + \langle W_x^{CoD}, z_{CoD}(CoD)\rangle + \langle W_x^g, z_g(g)\rangle}$ as $\boldsymbol{\alpha_x^{(i)}}$, and $\boldsymbol{\langle W_x^f, z_f\left(U_t^{(i)}\right)\rangle}$ as $\boldsymbol{\langle \beta_x^{(i)},\ \kappa_t^{(i)} \rangle}$ in the basic LC model. $\boldsymbol{\alpha_x^{(i)}}$ is primarily determined by $\boldsymbol{\omega_{x,0}}$ as an overall constant by age that is adjusted by the remaining components of $\boldsymbol{\alpha_x^{(i)}}$. The components for $\boldsymbol{\alpha_x^{(i)}}$ are all time-independent but age-dependent. The component for $\boldsymbol{\langle \beta_x^{(i)},\ \kappa_t^{(i)} \rangle}$ has one age-dependent part and one time-dependent part. Therefore, even if we use only the mortality history by age and time, we can write the formula as $\boldsymbol{\sigma^{-1}(\hat{y}_{x,t}) = \omega_{x,0} + \langle W_x^f, z_f(U_t)\rangle}$, which retains the formula of the basic LC model: $\boldsymbol{\alpha_x + \langle \beta_x,\ \kappa_t \rangle = \alpha_x + \beta_x \kappa_t}$.

3.2.4 Variants 4 and 5

During the Variant 3 development process, we also prepare three models that only have the CNN part of Variant 3; two Variant 4 (one for each gender) and one Variant 5 (trained using data for both genders). These three models are developed to observe the forecast capability and characteristics of the CNN itself and the forecast performance when the target (output) data is not scaled (i.e. MinMaxScaler() is absent).

Figure 2 shows the model architecture of Variant 4 for both genders. We have applied different hyper-parameter values for each model to separately optimize forecast accuracy. Specifically, the only different hyper-parameter from Variant 3 is the convolution filter number, which is 50 for females and 10 for males.

Variant 5 features the same architecture with the hyper-parameter values of the male version of Variant 4. The purpose of Variant 5 is to know whether it is possible to train one standard model for multiple population groups.

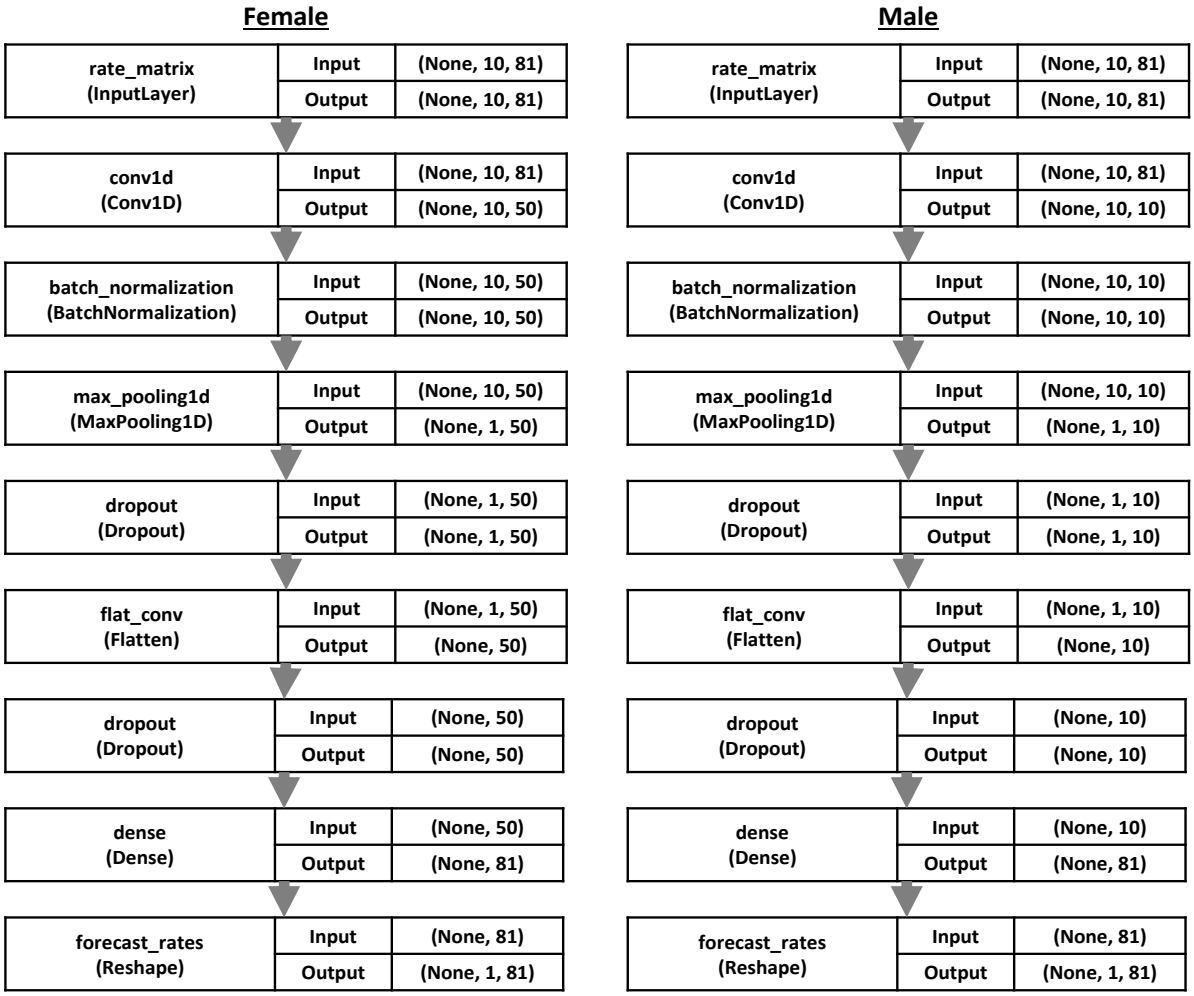


***Figure 2**: Graphic representation of the Variant 4 architecture*

3.2.5 New model

Our new model includes all of the components listed in Table 2. The principles of data preparation and model structure are identical to Variant 3, with the differences as follows and the model architecture shown in Figure 3.

Data preparation

As in the case of the gender-specific data, we arrange the CoD data as a separate input, represented by LabelEncoding() of the scikit-learn package (Pedregosa et al., 2011). To scale the target (output) data, we apply the scale separately for each gender and CoD to reflect the different scales of the data for CoD and the two genders.

The input branches

We add a separate input branch for the CoD data.

The hyper-parameters

We establish the following hyper-parameters based on observing multiple rounds of model training.

- The node dropout rate of the dropout layer in the convolution branch is set at 0.3.
- The pool size of the max-pooling layer is set at 5.
- The output dimensions of embedding layers for the gender and CoD branches are set at gender = 10 and CoD = 30. As recognized by Perla et al. (2021), the dimension values need not be consistent with the number of gender and CoD categories.
- The node dropout rate of the dropout layer after the concatenate layer is set at 0.3.
- The learning rate is set at 10^{-4} for the optimizer setting after "forecast_rate (Reshape)". We also set the epochs at 600 to balance the epoch with the learning rate adjustment.

The forecast value aggregation

After obtaining the forecast values for each CoD, we sum them to calculate the all-cause mortality rate for each gender.

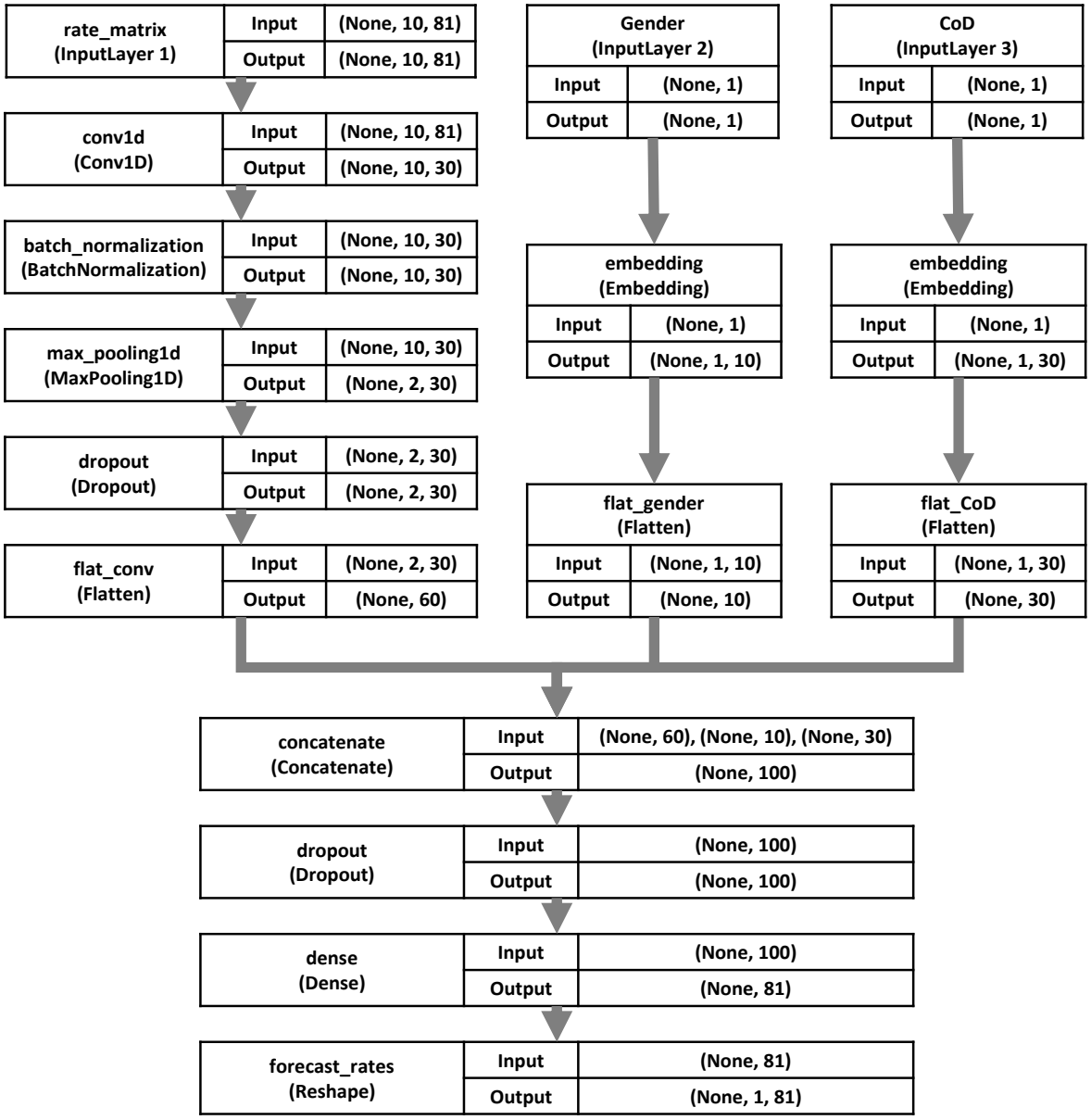


***Figure 3**: Graphic representation of the new model's architecture*

3.3 Model comparison

Following Q. Wang et al. (2021), we compare the models based on mean square error (MSE) loss and computing time. Due to time constraints, for the new model and Variants 3, 4 and 5, we have recorded only the results from the first run (i.e. no measure to compensate for the randomness of the initial parameters of the neural networks). One approach to overcoming such randomness involves following Perla et al. (2021) in taking an average value from the multiple model training runs. This represents an opportunity for future research.

3.4 Limitations

This research study's time constraints demanded limiting the number of hyper-parameter combinations and epochs to try while training the neural-network-based models. Similarly, as Section 3.3 mentioned, we record only the first-round training and testing results for the neural-network-based models.

4. Overview of the data

Three sources of data contribute to the baseline datasets for forecast modelling and analysis.

4.1 Death counts

4.1.1 From 1959 to 2016

We use the data used in Villegas et al. (2021), which represent the cleaned version of the publicly available data at Centers for Disease Control and Prevention (n.d.). The data in Villegas et al. (2021) comprises the death counts aggregated by year (1959–2016), age (0–100), gender (female/male), Human Mortality Database Cause of Death (HMD CoD) group (92 groups) and CoD category. Note that we do not include the US overseas territories (e.g. Puerto Rico and the US Virgin Islands) in our study's datasets. Equivalent but only partially cleaned data from the CDC is available from the National Bureau of Economic Research (n.d.). The causes of death in the CDC data are coded based on the International Classification of Diseases (ICD), which has been revised three times: ICD7 à ICD8 in 1968, ICD8 à ICD9 in 1979, and ICD9 à ICD10 in 1999.

As Section 3.1 mentioned, Villegas et al. (2021) summarize the CDC data based on Table 3.2 of their research study, which further classifies six broad CoD groups from the HMD CoD data: circulatory diseases, neoplasms, respiratory diseases, digestive system diseases, external causes and other causes. This system of categorization is helpful for its simplicity and interpretability. We slightly modify the data by omitting unnecessary rows and columns (see Figure 14 in the Appendix for details of the procedure). Consequently, 1,077,872 records exist, each of which contains the sum of death counts of the variable combinations listed in Table 3.

4.1.2 From 2017 to 2019

We also use publicly data available from the National Bureau of Economic Research (n.d.). The data comprises mortality microdata that tabulate individual death records by age, year, and CoD for the US population, based on the ICD10. As with the data for the period 1959–2016, the data for each year are cleaned and summarized according to the CoD groups described by Villegas et al. (2021). The general procedure is as follows (see Figure 14 in the Appendix for details of the procedure).

- Keep only the required columns (i.e. year, age, (main) cause of death and gender) from the original dataset.

- Given there are few records with ages over 100, transform these to 100. Next, aggregate the records by summing the columns retained. The number of records becomes less than 3% of the original.
- Map and replace each record's CoD with the category designated by HMD CoD and Villegas et al. (2021)). Then, aggregate the records by summing the remaining columns. The number of records becomes less than 0.5% of the original.
- Distribute the death counts of unknown age to known ages. The decrease in the number of records is minor, given the number of remaining records.

After cleaning the data, we have 43,408 records. Each record has the sum of death counts of a variable combination (not a single variable) listed in Table 3.

Variable	Description	Data type	Allowable entries
Year	The calendar year in which the death occurred	int64	1959–2019
Age	The age of the person when they died	float64	0–100
Cohort	The birth year cohort of the person in the data	float64	"Year" minus "Age"
Group	A number indicating one of the 92 cause-of-death categories used by the HMD	float64	1–92
ICD	A number indicating the ICD version	int64	7–10
Gender	The name of the person's gender	object	Female, Male
Deaths	The number of people who died	float64	Any number >= 0
Diseases	The name of diseases corresponding to "Group"	object	Strings such as "Tuberculosis"
Level 1	The higher (i.e. broader) CoD grouping categories detailed by Villegas et al. (2021)	object	Strings such as "External causes"
Level 2	The lower (i.e. more specific) CoD grouping categories detailed by Villegas et al. (2021). Note that our study does not use this category but left it in the dataset to show the data availability.	object	Strings such as "Other causes"

Table 3: *Glossary of post-cleaning death-count data*

4.2 Exposure-to-risk (i.e., population)

From 1959 to 2019, we use the US portion of the publicly available data from the HMD included in Shkolnikov et al. (2021). After cleaning the data by omitting the years before 1959, stacking gender columns into one column, transforming ages above 100 to 100 and aggregating the data according to the remaining variables (see Figure 15 in the Appendix for details of the procedure), 18,483 records exist, with each record containing the sum of exposures of a variable combination listed in Table 4.

Variable	Description	Data type	Allowable entries
Year	The calendar year that the population was counted	int64	1959–2019
Age	The age of the sub-population group	float64	0–100
Gender	The name of the person's gender	object	Female, Male
Exposures	The total headcount of the population in the US not including the US overseas territories (e.g. Puerto Rico and the US Virgin Islands)	float64	Any number >= 0

***Table 4**: Glossary of the post-cleaning US population data*

5. Exploratory data analsyis

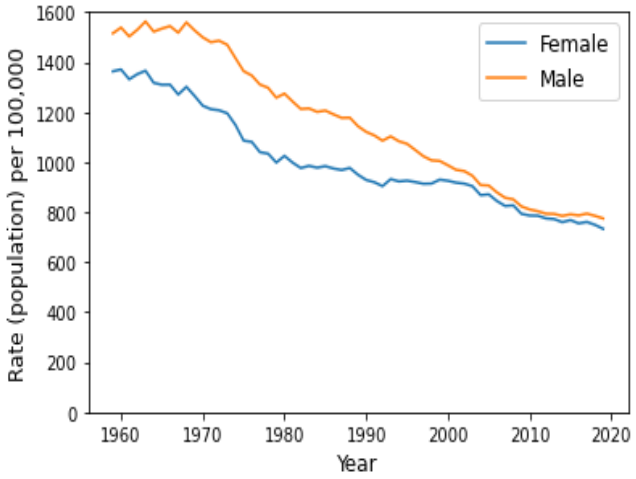


***Figure 4:** ASDR (headcount per 100,000) for all causes, 1959–2019*

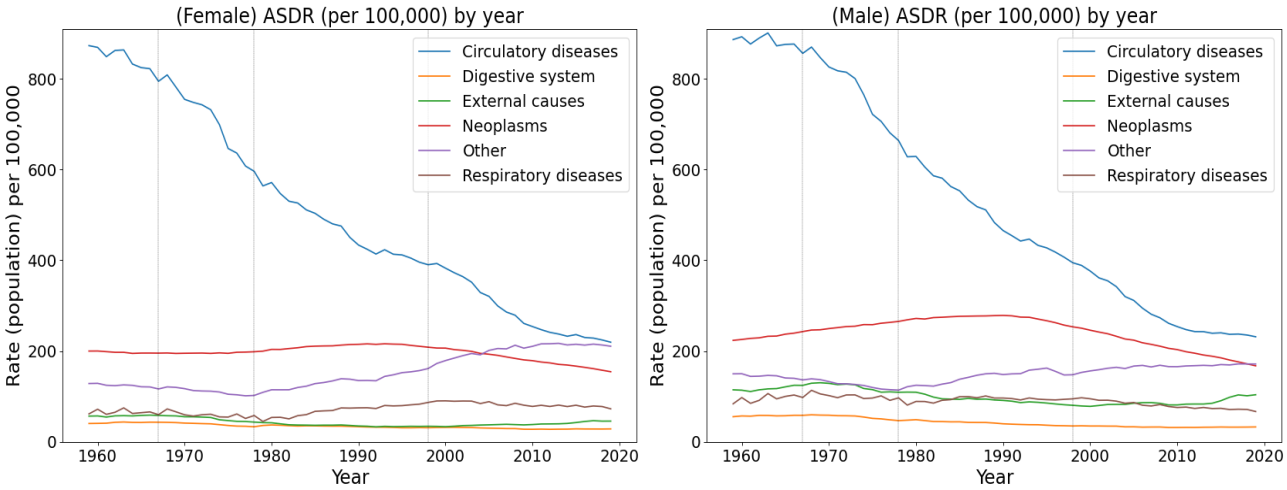


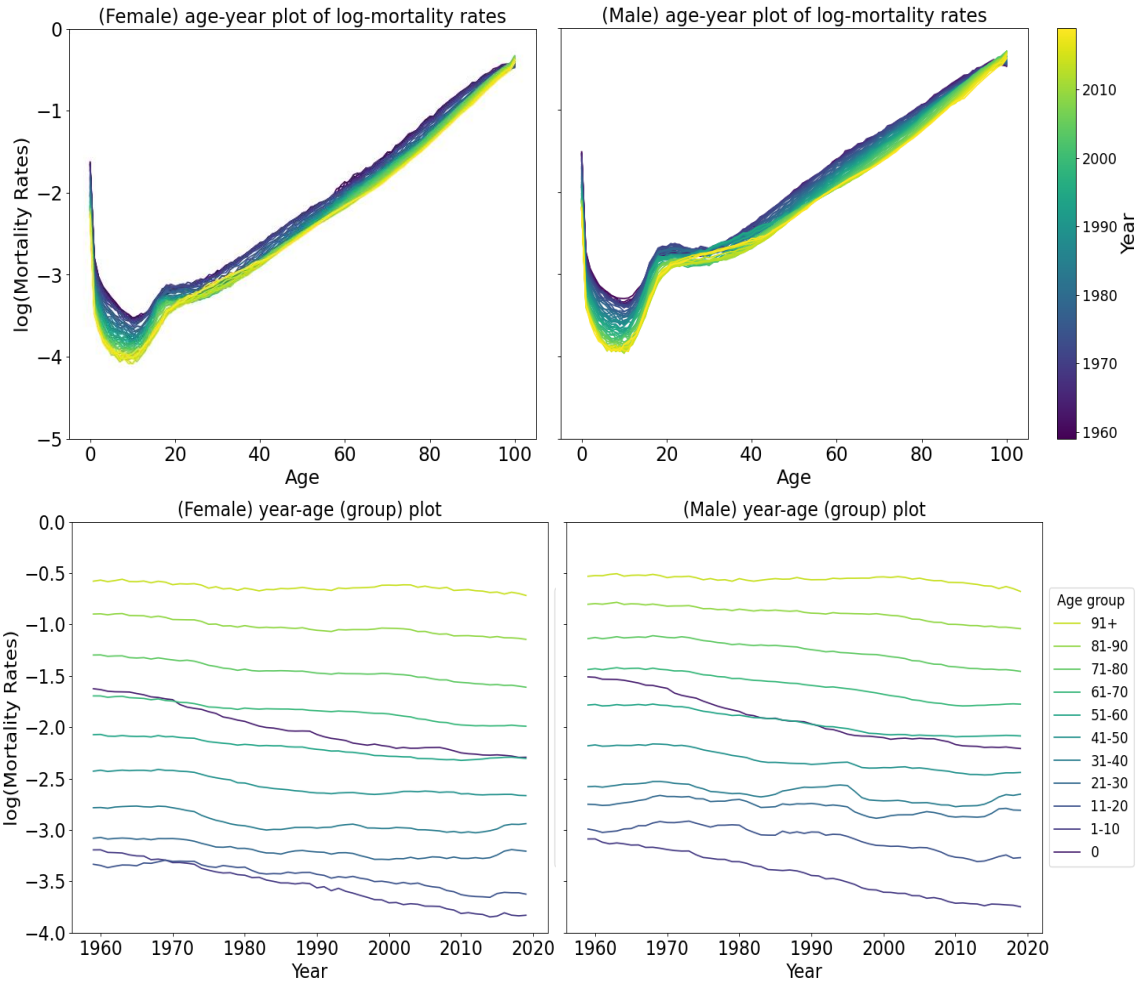
***Figure 5:** ASDR (headcount per 100,000) by year grouped by Level 1 causes of death, 1959–2019*

Figure 4 shows the age-standardized death rate (ASDR; headcount per 100,000) for all causes for females and males from 1959 to 2019. The ASDRs of both genders have decreased: (female) from 1,363 in 1959 to 733 in 2019 and (male) from 1,514 in 1959 to 774 in 2019 (unit: headcount per

100,000). The difference between genders has become smaller as the time has approached 2019.

Figure 5 shows that the leading causes of death have changed over time, with the changes different for females and males. Decreased circulatory diseases are the major contributor to decreased ASDRs for both genders: (female) from 876 in 1959 to 220 in 2019 and (male) from 888 in 1959 to 232 in 2019 (unit: headcount per 100,000). Decreased neoplasms are the second-biggest contributor, although the rate was increasing for both genders until the early 1990s. Meanwhile, the impact of "other causes" has increased for both genders, becoming the second-biggest cause of death in 2019, changing from 129 in 1959 to 211 in 2019 for females and from 150 in 1959 to 172 in 2019 for males (unit: headcount per 100,000). It would be worth probing the more specific causes captured by "other causes".

Mortality forecasting models must grasp such changes, given the nature of the data as a time series. It is also worth noting that, given the smooth lines of the plots in Figure 5-2, ICD changes have not drastically changed the ASDR breakdown by CoD.



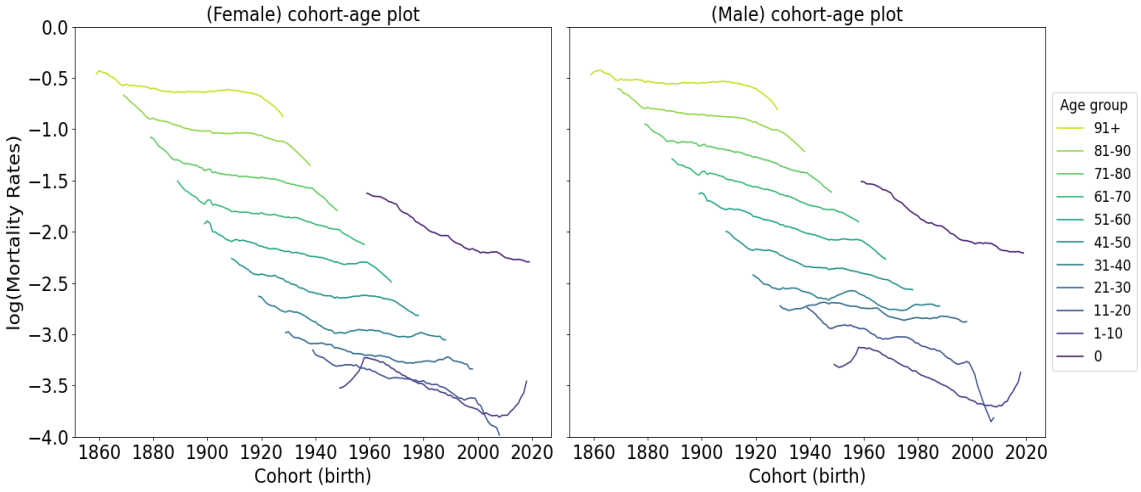


Figure 6: *APC plots of age-specific log-mortality rates for all causes, 1959-2019*

Figure 6 visualizes the log-mortality rates for three effects: age, period (time in years) and cohort (birth year). Age-year plots indicate that the general trends of the log-mortality rates are consistent for both genders for all periods; high rates for newborn infants, lowest rates around the age of 10, a rapid increase until around the age 20 and constant increases for the remaining ages. Another general trend is males having higher log-mortality rates at younger ages (10–60) than females. Generally, log-mortality rates have decreased in recent years. However, one notable phenomenon in the age-year plots is the higher log-mortality rates observed among those between the ages of around 20 and 40 in recent years (since 2015). This could influence the forecasted mortality rates.

Cohort-age plots display a general trend consistent with the age-year and year-age (group) plots, namely, the more recent the birth cohort, the lower the log-mortality rate. Meanwhile, the recent cohorts (since 2010) in the age group 1–10 show an acute increase for both genders. This could influence the forecast mortality rates.

Overall, given the younger age groups (0–20) show the distinct trends described, we focus on the data for individuals aged 20 and above for the rest of this study to avoid potential “noise” during model training.

6. Result and discussion

6.1 Overview

Table 5 summarizes the computing time and total MSE of the models. The smaller the MSE, the higher the forecast accuracy. From an accuracy standpoint (MSE from forecast vs actual of log mortality rates), the new model outperforms some variants, including Variant 2 (i.e. a conventional LC model). The new model does not outperform Variant 1 (i.e. the simplest LS model). However, it is worth mentioning that the new model still has the potential to yield better results because the training loss score in the right graph of Figure 13 has not yet plateaued; it would likely decrease after 600 epochs. The number of epochs represents how many times a model can be trained within one training. A model can be trained until the loss scores (training and validation) reach the plateau.

Variant 3 outperforms Variants 1 and 2 for females but not males. The left graph of Figure 13 shows that both the training and validation loss scores are stable at the final epoch (300) and are unlikely to decrease further. Therefore, more investigation with robust combinations of different hyper-parameters would be beneficial.

The new model and Variants 3 and 5 are the CNN-based models trained with the combined data for CoD and the two genders. Of these models, Variant 5 does not have the branch structures to separately consider categorical variables (i.e. gender and CoD) and considers the training data's target (output) without scaling. The MSEs in Table 5 show that those with the branch structures and the scaled target of the training data outperform Variant 5. Note that Variant 5 is less likely to perform well after the longer training time (i.e. larger epochs). The middle graph of Figure 13 shows the risk of overfitting with larger epochs due to validation loss beginning to increase slightly after epoch 20.

All of the neural-network-based models require more computing time for training. It is reasonable that structurally complex models require more computation time. Therefore, we should always be aware of the time cost in the event of a future study involving many training and assessment rounds (e.g. testing how various combinations of hyper-parameters affect model performance). The tendency for longer computation time for more complex models also applies to forecast time. Still, the scale is sufficiently small that the consequence is likely negligible when only generating forecast data.

Model	Computing time (in minute)		MSE total (in 10^{-3})	
	Training	Forecast	Female	Male
New model	215.00	85.83 x 10^{-3}	53.98	113.90
Variant 1	2.98 x 10^{-3}	9.73 x 10^{-3}	36.18	34.21
Variant 2	14.18 x 10^{-3}	35.83 x 10^{-3}	160.03	133.17
Variant 3	79.12	14.98 x 10^{-3}	33.78	167.38
Variant 4 (for female)	51.98	18.00 x 10^{-3}	67.06	Not applicable
Variant 4 (for male)	51.92	18.00 x 10^{-3}	Not applicable	147.97
Variant 5	52.83	35.55 x 10^{-3}	527.66	85.62

***Table 5**: Summary of computing time and total MSEs of models (the smaller the MSE, the higher the forecast accuracy)*

6.2 Forecast vs actual

Figure 7, Figure 8 and Figure 10 show the forecast vs actual plots for Variants 1, 2 and 4. The dotted lines from 2008 are the forecast lines. We train the three Variants separately for the data for each gender, and we used the data of different categories to train one model for the new model and Variants 3 and 5. The forecast lines generated by Variant 4 clearly differ between genders and generally resemble Variants 1 and 2 except in the case of certain older ages (i.e. 70 and 80). This means that both the conventional LC models and the CNN model can generate forecasts within the range of accuracy between Variants 1 and 2, with Variant 4's accuracy depending on the number of epochs. This represents an opportunity to enhance future studies.

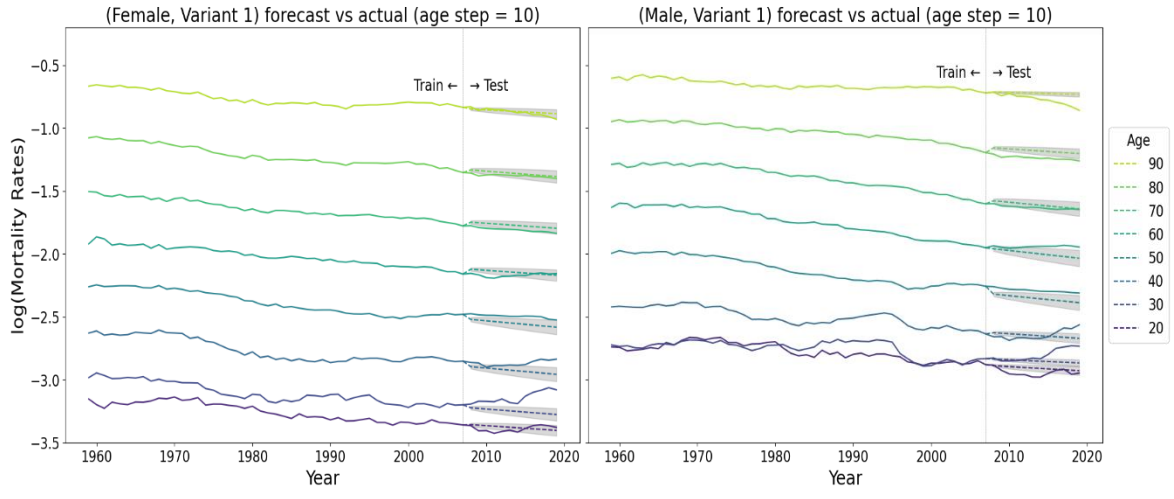


***Figure 7**: Forecast vs actual log mortality rate over time for Variant 1*

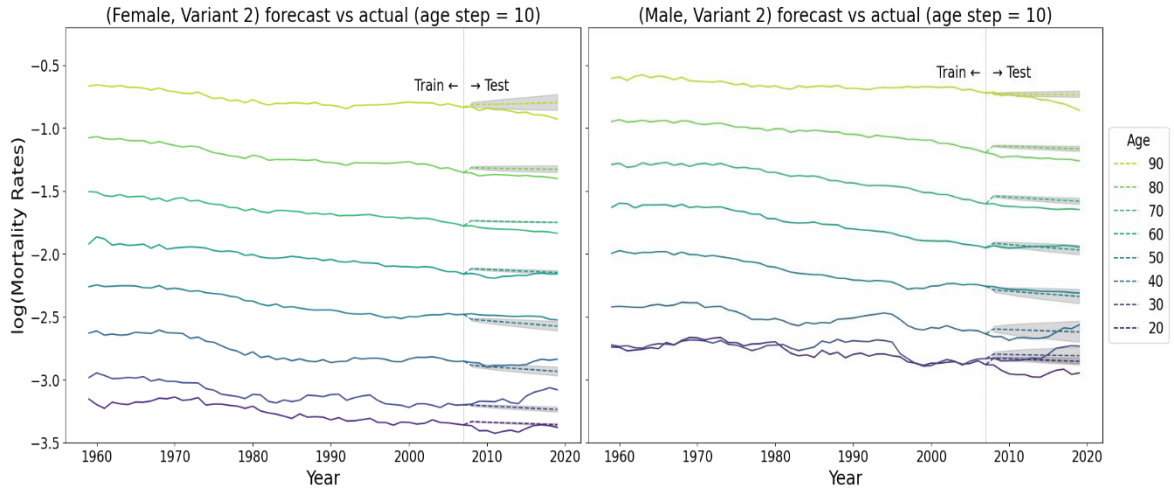


***Figure 8**: Forecast vs actual log mortality rate over time for Variant 2*

Figure 9, Figure 11 and Figure 12 show the forecast vs actual plots for Variants 3 and 5 and the new model. Variant 5 does not successfully differentiate between females and males, although some differences in the plots do appear. Observations during the development of Variant 3 reveal that using the branch structures for the categorical variables (i.e. gender and CoD) and MinMaxScaler() for the training data's target (output) contributes to the forecast differentiation.

Although the trend of the forecast values produced by the new model and Variants 3 and 5 are generally more optimistic than the actual values for most age groups, we do see some improvement in the forecast accuracy during the training and testing of different combinations of hyper-parameters, including the kernel size of the convolution layer and longer training times. This represents another opportunity to improve future studies.

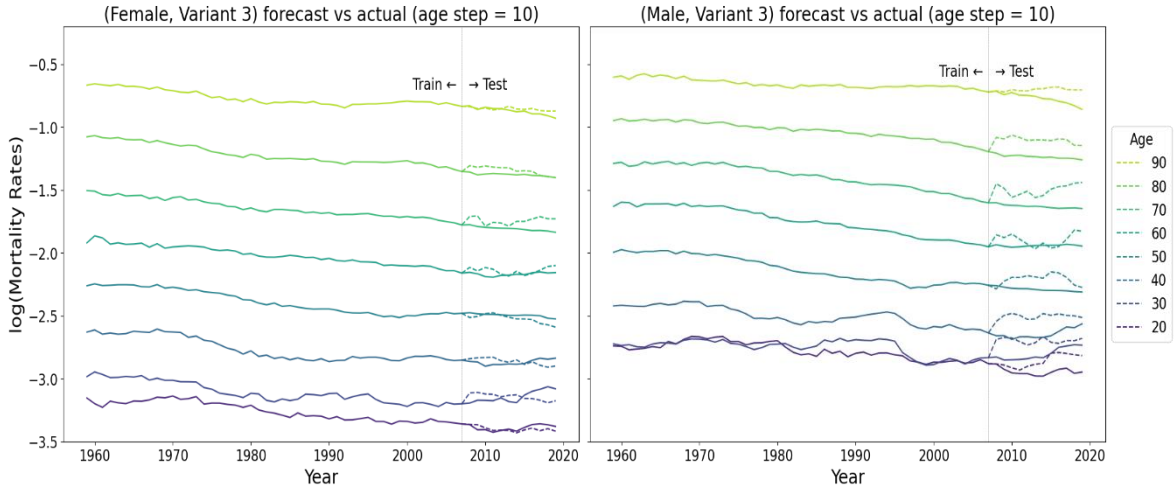


***Figure 9**: Forecast vs actual log mortality rates over time for Variant 3*

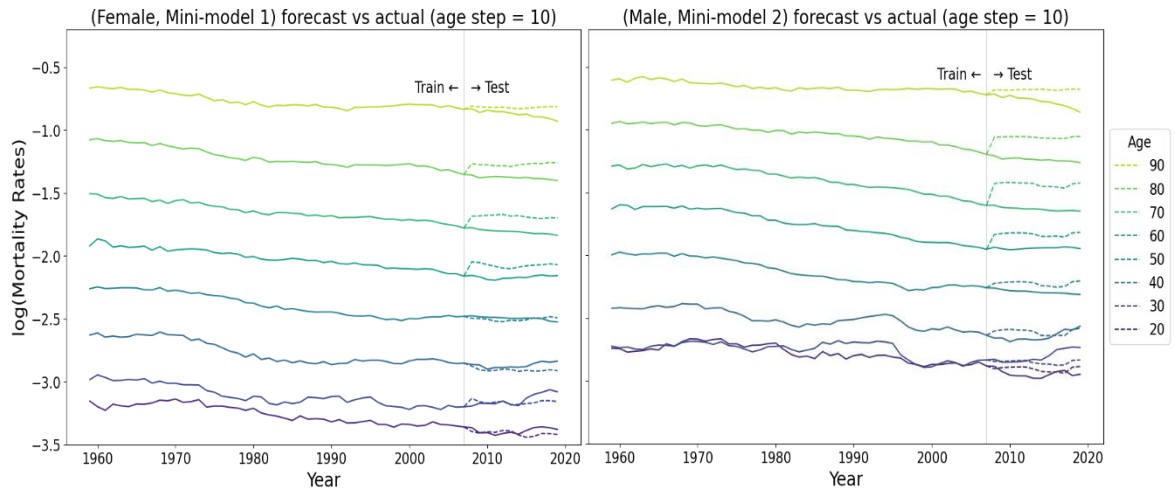


***Figure 10**: Forecast vs actual log mortality rates over time for Variant 4*

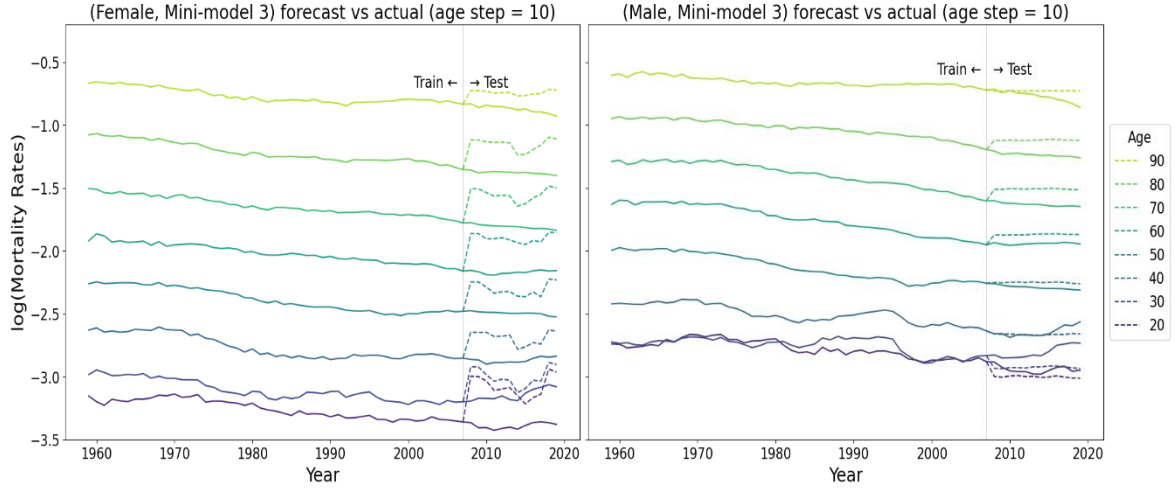


***Figure 11**: Forecast vs actual log mortality rates over time for Variant 5*

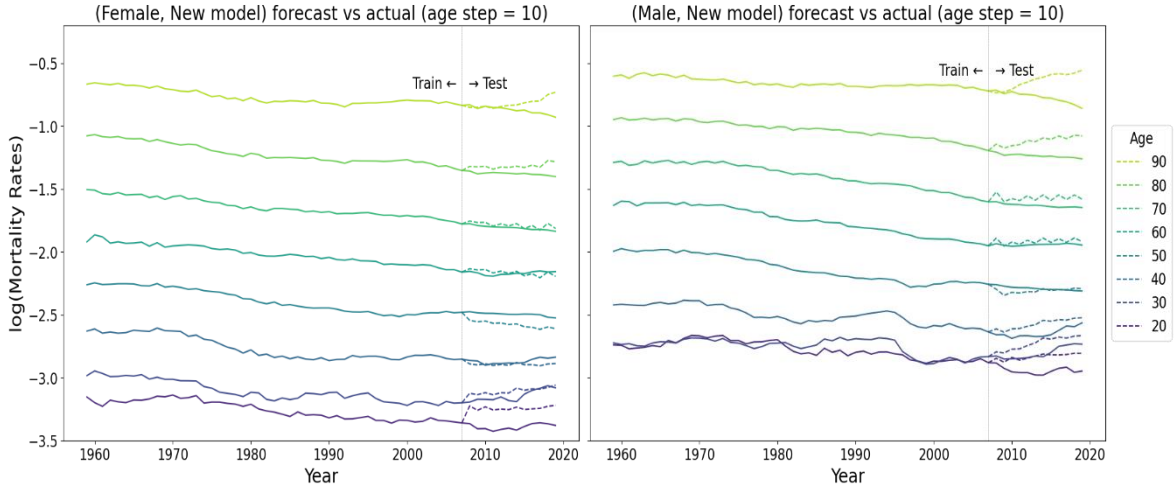


***Figure 12**: Forecast vs actual log mortality rates over time for the new model*

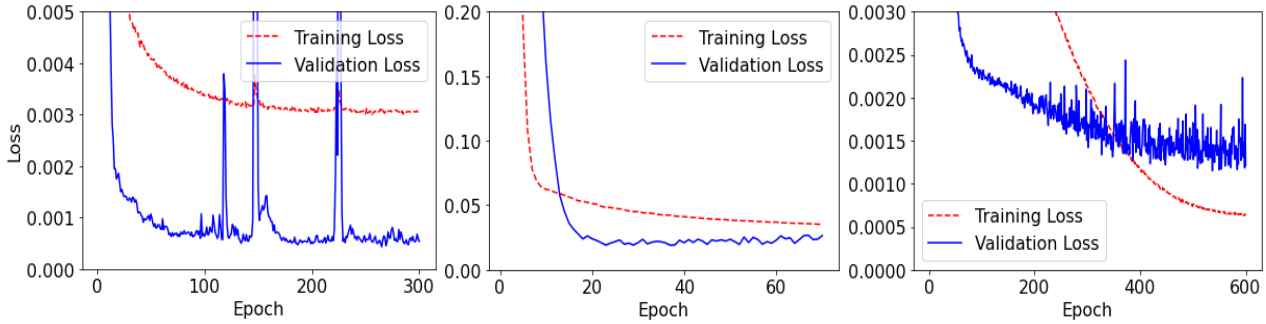


***Figure 13**: MSE-based loss score by epochs of training and validation.*
Left: Variant 3; middle: Variant 5; right: new model.

7. Conclusion

Overall, our research study assessed the performance of the CNN-based model with CoD usage as a feature variable by comparing it to the conventional LC and simpler CNN-based models. The assessment criteria were primarily the duration (training and forecast) and MSE (accuracy), and goodness of fit by forecast vs actual plots by age. We used the US mortality data between age 20 and over from 1959 to 2019 to see the model performance as a starting point for future research.

The new model outperforms some variants, including the conventional LC model using the data for each CoD and gender. Given that the new model has the potential to yield still better results (i.e. the loss value likely keeps decreasing) if we train with more epochs, we contend that our new neural-network-based model can more accurately forecast the all-cause mortality rate from cause-of-death-specific data. Our assessment also shows the time cost for model training. Due to the nature of the data (i.e. yearly mortality rate is typically updated only once a year or less), this time cost is negligible. However, we must address the time cost for all possible urgent scenarios (e.g. quick and dirty decision-making with a small RAM computer under the short deadline for a business case).

Future research should investigate robust combinations of different hyper-parameters, given we have seen forecast accuracy being sensitive to hyper-parameter changes (e.g. kernel size in the convolution layer). Using inception models (Szegedy et al., 2015), which combine CNNs with different kernel sizes, could ease hyper-parameter assessment while the models enhance forecast accuracy. Another way to enhance the robustness of the performance analysis would be to use data from multiple other countries for model training and testing. Although this step has been taken by previous studies, comparing the new model with models other than (convolutional-)neural-network-based models could also enhance robustness. Finally, investigating feature importance could be used to interpret the neural network model estimates.

8. References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. (2016, May 31). *TensorFlow: A system for large-scale machine learning*. arXiv.org. Retrieved July 1, 2022, from https://arxiv.org/abs/1605.08695

Bertolaccini, L., Solli, P., Pardolesi, A., & Pasini, A. (2017). An overview of the use of artificial neural networks in lung cancer research. *Journal of thoracic disease*, *9*(4), 924–931.

Brownlee, J. (2018, November 12). *How to Develop Convolutional Neural Network Models for Time Series Forecasting.* Machine Learning Mastery. https://machinelearningmastery.com/how-to-develop-convolutional-neural-network-models-for-time-series-forecasting/

Brunton, S. (2020, January 20). *Singular Value Decomposition (SVD): Overview* [Video]. YouTube. https://www.youtube.com/watch?v=gXbThCXjZFM

Brunton, S. (2020, January 20). *Singular Value Decomposition (SVD): Mathematical Overview* [Video]. YouTube. https://www.youtube.com/watch?v=nbBvuuNVfco

Brunton, S. (2020, Janurary 28). *Principal Component Analysis (PCA)* [Video]. YouTube. https://www.youtube.com/watch?v=fkf4IBRSeEc

Booth, H., & Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science, 1(2)*, 3–43.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., & Balevich, I., 2009. A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States. *North American Actuarial Journal*, 13(1), 1–35.

Center for Disease and Control Prevention. (n.d.). *Public-Use Data Files*. https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm#Mortality_Multiple

Koehrsen, W. (2018, October 2). *Neural Network Embeddings Explained*. Towards Data Science. https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419), 659–671.

Lyu, P., de Waegenaere, A., & Melenberg, B. (2020). A multi-population approach to forecasting all-cause mortality using cause-of-death mortality data. *North American Actuarial Journal*, 25(sup1), S421–S456.

National Bureau of Economic Research. (n.d.). *Mortality Data - Vital Statistics NCHS Multiple Cause of Death Data*. Retrieved April 1, 2022, from https://www.nber.org/research/data/mortality-data-vital-statistics-nchs-multiple-cause-death-data

Natrajan, S., Sathish, G., & Jeyakumar, B. (2018), Data wrangling and data leakage in machine learning for healthcare. *International Journal of Emerging Technologies and Innovative Research*, 5(8), 553–557.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Perla, F., Richman, R., Scognamiglio, S., & Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7), 572–598.

Richman, R. & Wüthrich, M. V. (2021). A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science,* 15(2*)*, 346–366.

Schweizerische Aktuarvereinigung. (n.d.). *Actuarial Data Science.* https://www.actuarialdatascience.org/ADS-Tutorials/

Shkolnikov, V., Barbieri, M., & Wilmoth J. (2021, March 17). *The United States of America, Exposure to risk (period 1x1)*. The Human Mortality Database. https://www.mortality.org/hmd/USA/STATS/Exposures_1x1.txt

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2015.7298594

Tabeau, E., Ekamper, P., Huisman, C., & Bosch, A. (1999). Improving overall mortality forecasts by analysing cause-of-death, period and cohort effects in trends. *European Journal of Population,* 15(2), 153–183.

Tam, A. (2021, November 20). *Using CNN for financial time series prediction*. Machine Learning Mastery. https://machinelearningmastery.com/using-cnn-for-financial-time-series-prediction/

Villegas, A.M., Bajekal, M., Haberman, S., & Zhou, L. (2021, September). Analysis of Historical U.S. Population Mortality Improvement Drivers 1959–2016. Society of Actuaries Research Institute. https://www.soa.org/globalassets/assets/files/resources/research-report/2021/2021-historical-us-population-mortality-improvement-drivers.pdf

Wang, C. W., Zhang, J., & Zhu, W. (2021). Neighbouring prediction for mortality. *ASTIN Bulletin*, 51(3), 689–718.

Wang, Q., Hanewald, K., & Wang, X. (2021). Multistate health transition modeling using neural networks. *Journal of Risk and Insurance,* 89(2), 475–504.

Wilmoth, J. R. (1995). Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies, 5(4)*, 293–319.

World Health Organization. (2021, October 4). *Ageing and health*. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

9. Appendix

9.1 Glossary

Here we describe and clarify certain information mentioned in the main text.

9.1.1 Feature variable reduction method

The following methods appear in Table 1.

- **SVD (Singular Vector Decomposition)**: This is a widely used data processing algorithm that decomposes the (feature) matrix into three matrices. Each of those three matrices has vectors or values that are hierarchically arranged by the importance of each vector or value. Such a hierarchy is determined by the ability to explain the correlations of values within the original matrix. The hierarchy allows us to omit vectors and values of less importance, reducing the original matrix to a simpler representation (Brunton, 2020).
- **PCA (Principal Component Analysis)**: This is the statistical interpretation of SVD, assuming

that the (feature) matrix has a statistical distribution. It can use SVD while processing the data. However, PCA (1) centres the values of each feature and (2) calculates the vectors (principal components) that work as the coordinate systems (i.e. axes) most capable of explaining the original matrix's covariance (Brunton, 2020).

- **Embedding (NN layer)**: This is a technique for converting discrete (categorical) variables into low-dimensional vectors of continuous values. An "embedding layer" does this as a hidden layer within a neural network (Koehrsen, 2018), which is especially useful when dealing with many categorical variables or variables with a large number of categories. In such cases, the popular one-hot encoding approach produces many variables, increasing the computational burden (Perla et al., 2021).

9.1.2 Neural networks

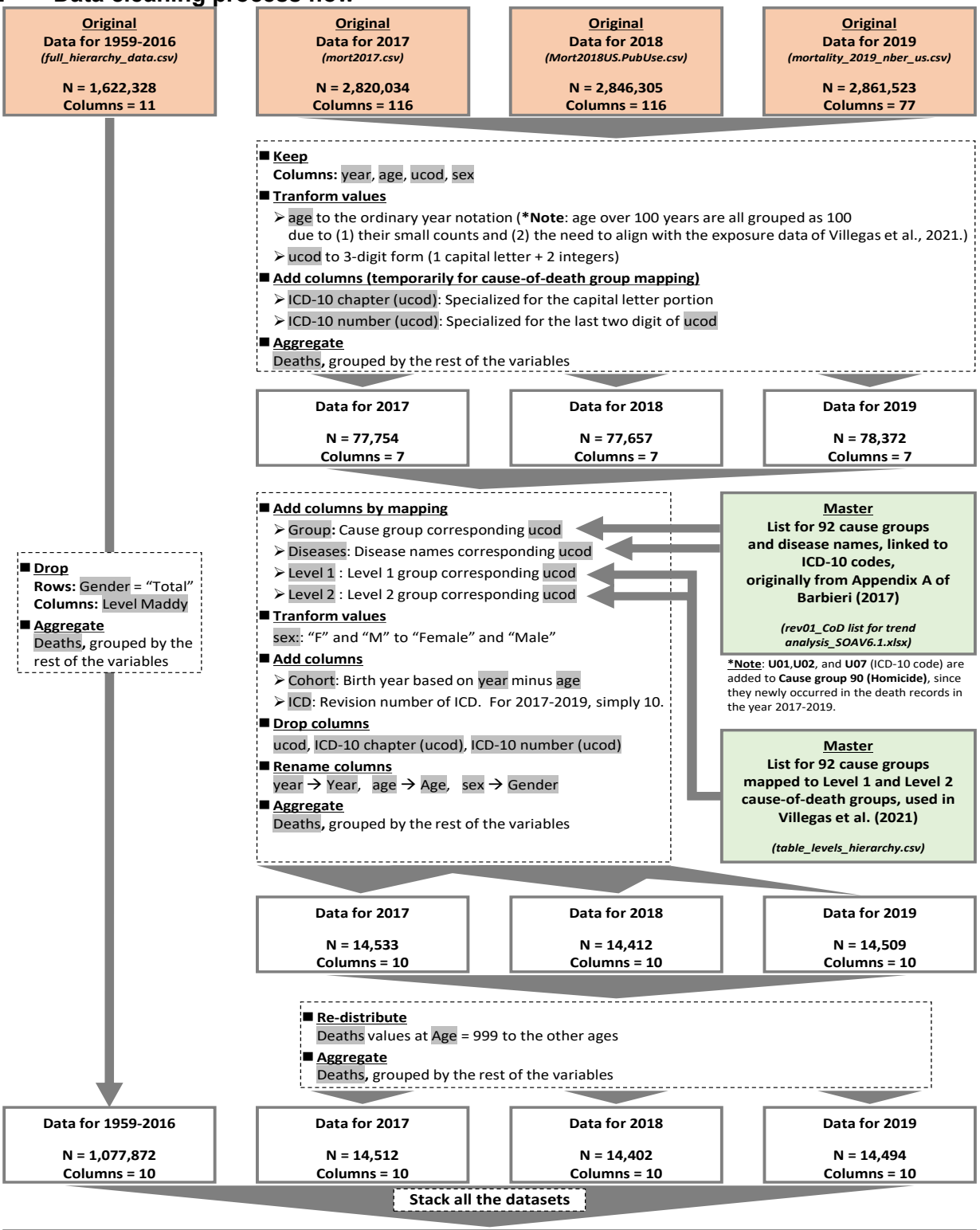
We only use CNNs in our project. However, it is necessary to describe the basics of neural networks to establish the foundational knowledge.

The basic structure of a neural network is layers of nodes. These are classified as an input layer, one or more hidden layers, and output layers. Each node connects to another node and has an associated weight, bias and threshold (of activation). We can construct a neural network to capture the complex relationships between feature variables (e.g. age, gender, time (year), birth year and cause of death) in a manner similar to image recognition or text processing (Perla et al., 2021).

Hence, the neural networks other than CNNs mentioned in Table 1 can be briefly described as follows:

- **FCN (feed-forward fully connected neural network)**: Each layer in a network is connected to every part of the previous layer. This is the most basic neural network form.
- **RNN (Recurrent neural network)**: This is designed to process sequential or time-series data, which depend on the information from previous parts of the sequence to compute the current output. It is achieved by including additional connections that cyclically (i.e. recurrently) connect the hidden layers.
- **LSTM (Long-short term memory)**: In addition to the cyclic component of RNNs, LSTM has a "memory" cell that stores and releases long-term information via a system of sub-networks called gates. LSTM performs well in sequential data processing applications, such as handwriting recognition and speech recognition.
- **CNN (Convolutional neural network)**: Similar to FCN, this neural network has a feed-forward structure. However, it can process multidimensional (i.e. grid-like or spatial topology, such as images) data. For a time series, a one-dimensional CNN is usually adequate. The other characteristics differentiating it from an FCN are as follows:
  - A CNN only connects neurons to a local region of the input array. This characteristic significantly reduces the number of parameters that must be learned, making it more computationally efficient and easier to train.
  - Weights (= a type of parameter) are shared across local regions by applying the same filter to the entire input array.
  - Such weight sharing enables the extraction of similar features in different regions of the input array.

9.2 Data cleaning process flow



Final

Data for 1959-2019 (Exported and saved as *"(Cleaned)_USA_Mortality_1959_2019.csv"*)

***N = 1,121,280**

Variable name	Year	Age	Cohort	Group	ICD	Gender	Deaths	Diseases	Level 1	Level 2
Non-null value count	1,121,280	1,121,280	1,121,280	1,121,280	1,121,280	1,121,280	1,121,280	1,121,280	1,121,280	1,121,280
Data type	int64	float64	float64	float64	int64	object	float64	object	object	object

Figure 14: *Flow chart representing the primary data assessment and cleaning process for the death records*

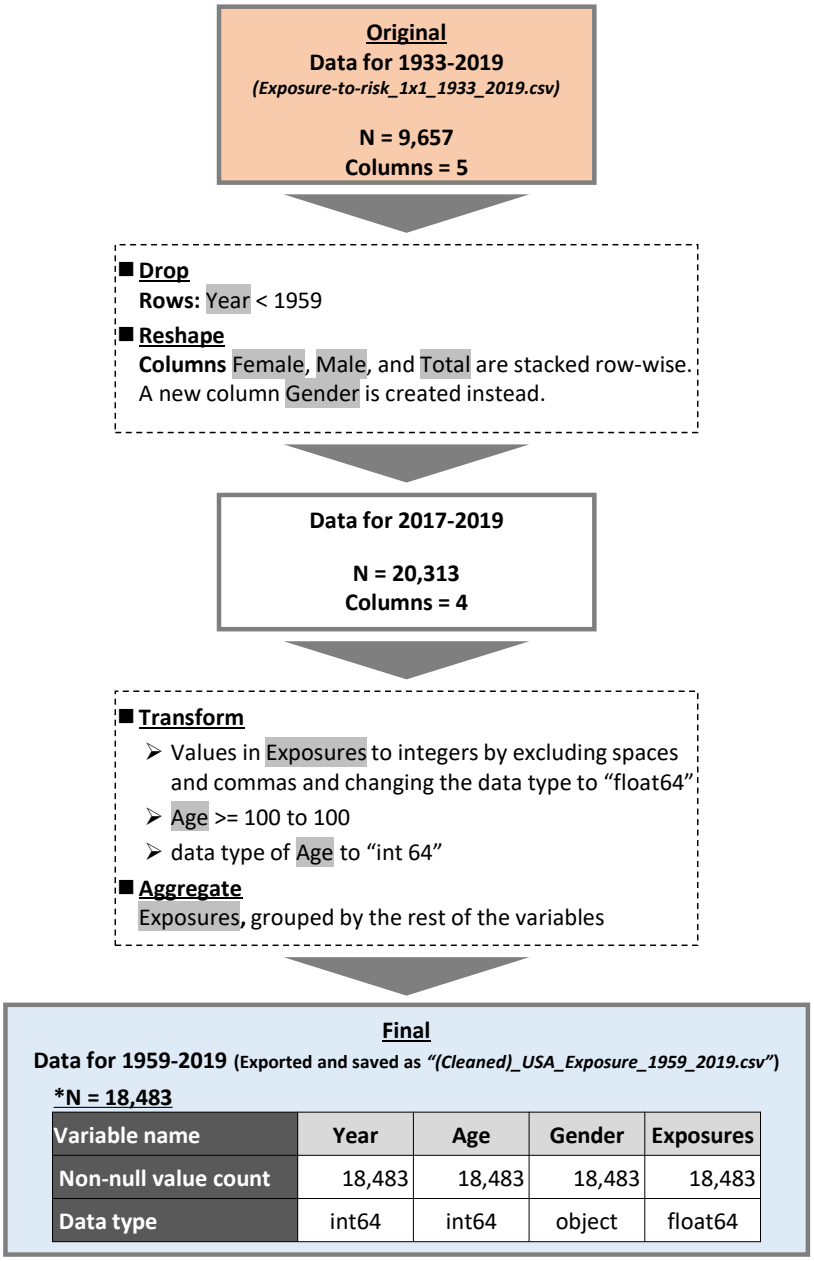


***Figure 15**: Flow chart representing the primary data assessment and cleaning process for the exposure data*

We have also assessed the consistency of our study data and the data from Villegas et al. (2021) using the 2016 US data. Specifically, we separately compared the death counts and the exposure counts. Based on the comparison result in Table 3, we decided to proceed with our data cleaning steps as the differences are negligible.

	Cleaned	Existing	Difference
Death count	2,749,864	2,744,248	5,616
Exposure count	322,851,497	323,369,352	▲517,855

***Table 6:** Comparison of the death and population counts from the newly cleaned dataset and the existing dataset. Both data represent the 2016 US datasets*

9.3 Supplementary plots for checking the Lee-Carter model

The plots in this section show the $\boldsymbol{\alpha_x}$, $\boldsymbol{\beta_x}$, and $\boldsymbol{\kappa_t}$ values of the conventional LC models (i.e. Variant 1 and 2). They allow us to check whether the fitting process of $\boldsymbol{\alpha_x}$, $\boldsymbol{\beta_x}$, and $\boldsymbol{\kappa_t}$ completed properly.

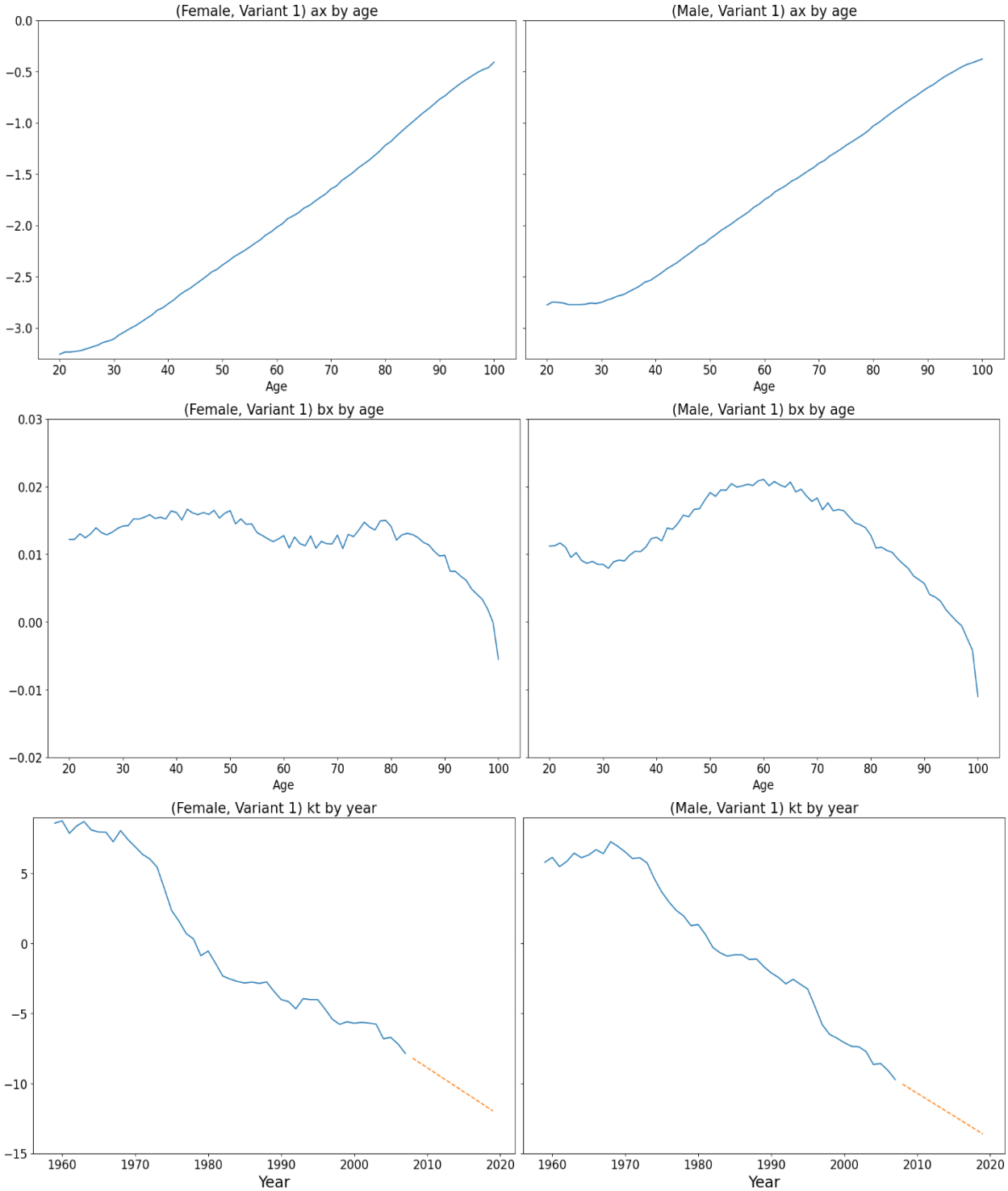


***Figure 16**: Plots of α_x , β_x , and κ_t of Variant 1*

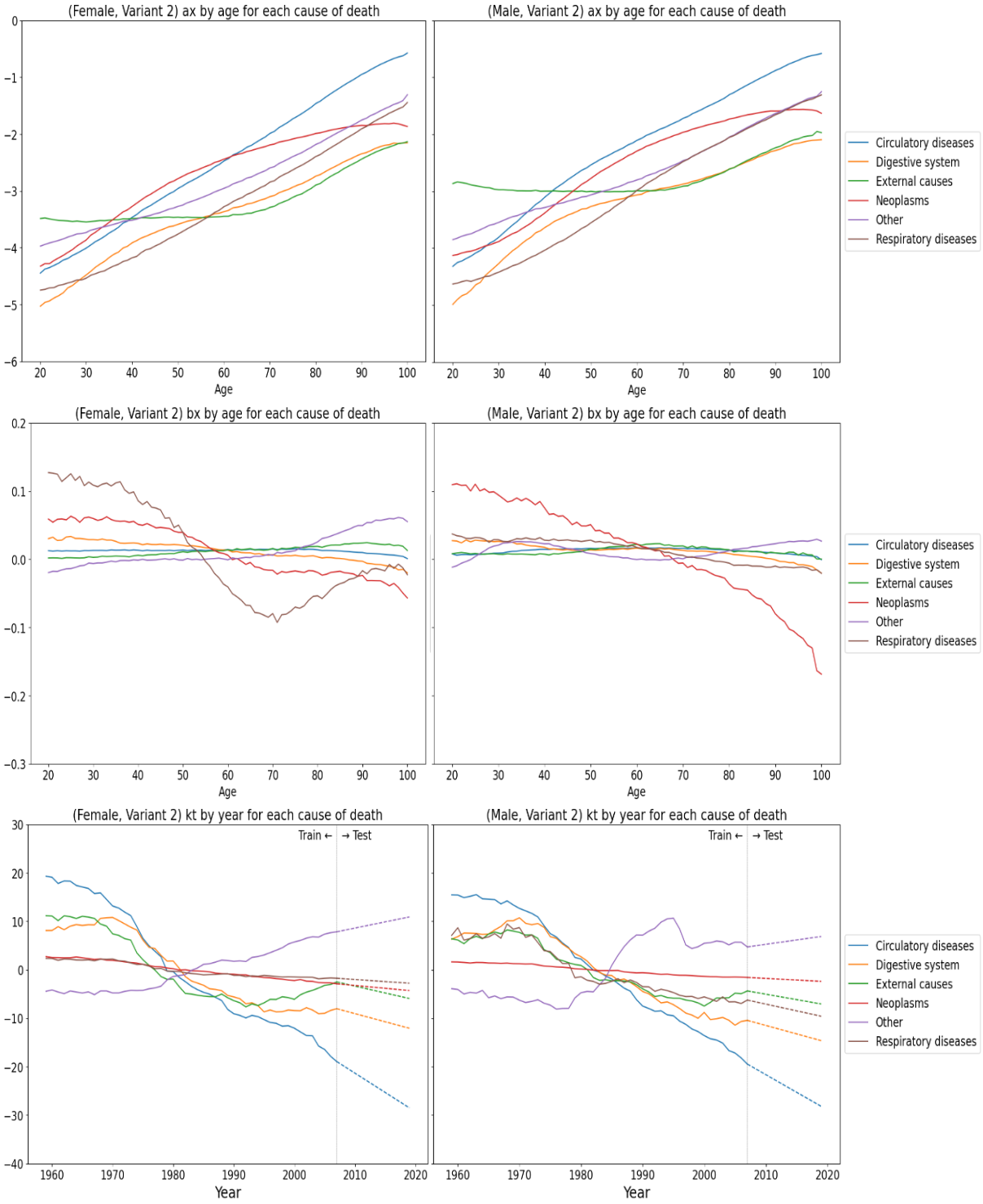


***Figure 17**: Plots of α_x , β_x , and κ_t of Variant 2*