Comparative Analysis and Prediction of Malignant Tumor Mortality in China Based on LSTM Models

Fan Yang
School of Computer Science and Engineering, Central South University, Nan Chang, China

Keywords: Malignant Tumor, LSTM Model, Mortality Rate.

Abstract: The prediction and analysis of the mortality rate of malignant tumors is a hot topic of great importance and urgency in the world. However, the dynamic mortality prediction model commonly used in the medical field still has some limitations in the nonlinear structure of mortality research. Therefore, based on the survey data of "Mortality rate of major diseases in selected urban/rural areas of China" from 2008 to 2021, this research will utilize the Long and short-term memory networks (LSTM) model to validate the model and predict the mortality rate in 2022. This research will also quantitatively analyze and explore the reasons for the mortality rate differences arising from regional differences and gender differences. The prediction results show that the mortality rate in urban areas in 2022 may increase compared with the value in 2021, while the mortality rate in rural areas will decrease, and the difference in the trend is only reflected in the regional differences. Furthermore, the trends in mortality rates over the years show a general decline in urban areas and an increase in rural areas, with the urban mortality rate being lower than the rural rate after 2020, and the male mortality rate being much higher than the female mortality rate.

1 INTRODUCTION

Malignant tumors, one of the most dangerous diseases nowadays, always occupy the first place on the list of causes of death in many countries (Sung et al. 2021). Rapid economic development, increasing environmental pollution, and changes in people's lifestyles have increased the incidence of malignant tumors. China National Cancer Center released a new report on malignant tumor statistics, which reveals that the number of new cases of malignant tumors in China reached 4,064,000 in 2016, and the total number of deaths reached 2,414,000 (Rongshou et al. 2022). The development of society and the economy are now seriously hampered by malignant tumors, which also pose a serious threat to people's lives and health. The importance and urgency of malignant tumor-related research have become self-evident. The mortality rate is an important index in malignant tumor-related research, and the prediction and analysis of malignant tumor mortality rate is also a hot topic. Commonly used mortality prediction models are static and dynamic mortality prediction models. Currently, the dynamic mortality prediction model represented by the Lee-Carter model is commonly used in the medical field. However, the dynamic mortality prediction model can only portray the linear relationship between the influencing factors and the mortality rate, and there are some limitations in the nonlinear structure of the mortality rate research (Gang et al. 2022). When it comes to machine learning algorithms, the nonlinearity of the kernel function and activation function can realize the nonlinear mapping of the data, which enables machine learning algorithms to have excellent nonlinear learning ability. Moreover, for the time series data, the Long and short-term memory networks (LSTM) model always performs excellently. In this research, the prediction of malignant tumor mortality and the analysis of regional and gender differences will help understand and predict the development trend of malignant tumors in China, explore the sore points of malignant tumor deaths in China, and provide targeted preventive suggestions to the public from a multi-level perspective.

Based on the data from the "Mortality rate of major diseases in selected urban/rural areas of China" from 2008 to 2021, this research will validate the model and predict the mortality rate of malignant tumors in selected urban/rural areas of China in 2022 with the LSTM model. The paper will also analyze the

Yang, F.
Comparative Analysis and Prediction of Malignant Tumor Mortality in China Based on LSTM Models.
DOI: 10.5220/0012816200003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 178-183
ISBN: 978-989-758-705-4

differences in mortality rate by region and gender and the reasons for these differences. This study might be helpful for the Chinese public and medical institutions to understand and treat malignant tumors. This research will first use the data of the previous 12 years to predict the mortality rate in 2020 and 2021, thus verifying the LSTM model's applicability. Then, the LSTM model will be used to predict the total mortality rate in urban and rural areas and the mortality rate of each gender in 2022. Furthermore, the differences in mortality rate due to regional and gender differences will be quantitatively analyzed and the reasons for these differences will be investigated.

2 METHOD

2.1 Data Source

The data for this research come from a survey "Mortality rate of major diseases in selected urban/rural areas of China", published in the China Statistical Yearbook by the National Bureau of Statistics of the People's Republic of China (NBS). The survey utilized a stratified probability sampling method proportional to population size. Several large cities and small and medium-sized cities were selected as urban respondents, and several municipalities and county-level cities were selected as rural respondents. The survey conducted gender-specific mortality rate and death cause statistics on the top 10 major diseases among Chinese residents in recent years according to the International Classification and Statistics Standard-10 (ICD-10) for Diseases. Based on the theme of this study and the distribution of causes of death, this study excluded diseases that were not continuously listed as the top 10 major diseases in the survey, as well as diseases with small mortality values ranging from 0.01% to 1%. This study will use the mortality rate value in selected urban/rural areas from 2008 to 2021 as research data. Since malignant tumors account for the highest proportion of causes of death among the ten major diseases throughout the year, and the mortality rate is also relatively high, it is of great research significance and representativeness.

The data of this study include six indicators, including the total mortality rate of malignant tumors, male mortality rate, and female mortality rate in urban and rural areas, etc.

2.2 Model

Long and short-term memory networks are a specific kind of recurrent neural network (RNN). RNN is the most natural architecture for sequential data due to the chaining feature. However, since the same function will be combined with itself many times, RNN will face the problem of gradient vanishing or gradient explosion and long-term dependency during training. The long-term dependence issue brought on by RNN will be partially resolved by Hochreiter's unique unit design of the LSTM model, which also effectively realizes the modeling of time-series data (Hochreiter and Schmidhuber 1997). RNN models all have some repeating neural network modules, the traditional RNN only simply repeats a single neural network layer, while the LSTM repeats module comprises four interacting layers, including three sigmoid layers and one tanh layer, whose interactions tend to be very specific.

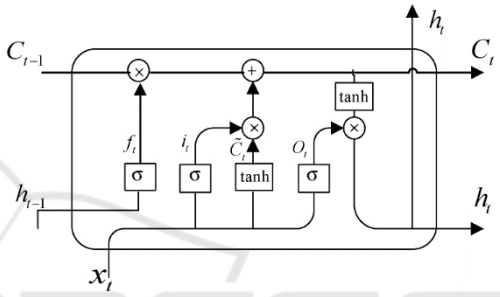

Figure 1: Cell of LSTM model.

As shown in Fig. 1, each LSTM cell contains three gate structures, the forgetting gate(f_t), the input gate(i_t), and the output gate(o_t), as well as a cell state to maintain and update the state (Han et al 2021). Where the forget gate is used to choose which information from the cell state to discard, the structure reads the previous output h_{t-1} and the current input x_t, then conducts a Sigmoid nonlinear mapping and outputs a vector f_t, which can be described as:

$$f_t = \left(W_{fh} h_{t-1} + W_{fx} x_t + b_f\right). \quad (1)$$

where W denotes the weights and b denotes the bias.

What information will remain in the cell state is determined by the input gate. It is divided into two sections. The sigmoid layer will first determine which values to update; the tanh layer will then generate a new vector of candidate values C_t and add it to the cell state. The relationship is stated as follows:

$$i_t = \left(W_{ih} h_{t-1} + W_{ix} x_t + b_i\right). \quad (2)$$

$$\tilde{C}_t = \tanh\left(W_{ch} h_{t-1} + W_{cx} x_t + b_c\right). \quad (3)$$

Next, the cell state will be updated from C_{t-1} to C_t as an input to the next cell, and the relationship can be expressed as:

$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t. \quad (4)$$

where ∘ denotes the Hadamard product.

The output gate is in charge of calculating the output value depending on the cell state. To begin, use a sigmoid layer to determine which part of the cell state will be output. Simultaneously, the cell state is processed by the tanh layer and multiplied by the output of the sigmoid gate, generating the value h_t, which is important for the computation of the following cell. This relationship can be stated as follows:

$$o_t = \sigma\left(W_{oh}h_{t-1} + W_{ox}x_t + b_o\right). \quad (5)$$

$$h_t = o_t \circ \tanh\left(C_t\right). \quad (6)$$

3 RESULT

3.1 Verification

To validate the LSTM model's applicability for forecasting malignant tumor mortality, this research will use the malignant tumor mortality data of urban areas from 2008 to 2019 to predict the corresponding mortality rates in 2020 and 2021. The predictions will also be compared with the true values to evaluate and validate the model's applicability based on Mean Squared Error (MSE). The models for applicability validation will include five mainstream regression prediction models: random forest, linear regression, decision tree, support vector machines (SVM), and LSTM. The prediction results are shown in Table 1:

Table 1: Mse of the Models.

Model	Random Forest	Linear Regression	Decision Tree	SVM	LSTM
MSE	7.012	9.179	1.542	8.321	0.813

The MSE obtained by the LSTM model is 0.813, whose magnitude is much smaller than the range of prediction values from 150 to 180, indicating good accuracy. Meanwhile, the MSE value of the LSTM model is much smaller than the corresponding values of the other four mainstream regression prediction models, which also indicates that the LSTM model has a more excellent performance than the other models in this research topic. Thus, it can be verified that the LSTM model is appropriate for the following prediction of malignant tumor mortality.

3.2 Prediction

3.2.1 Prediction of Urban and Rural Total Mortality Rates in 2022

In this research, the mortality rate of malignant tumors in urban and rural areas in 2022 will be predicted using the LSTM model. The predicted value of the total malignant tumor mortality rate in the urban region is 163.076, which is an increase of 2.76% compared to 2021. The predicted value of the total mortality rate for rural regions in 2022 is 154.456, which is a decrease of 7.54% compared to 2021. At the same time, urban regions' forecasts are 5.5% higher than rural regions' forecasts. The trend of the total mortality rate of malignant tumors in urban and rural regions from 2008 to 2022 is shown in Fig. 1, where the horizontal coordinate indicates the year and the vertical coordinate is the mortality rate (1/100,000):

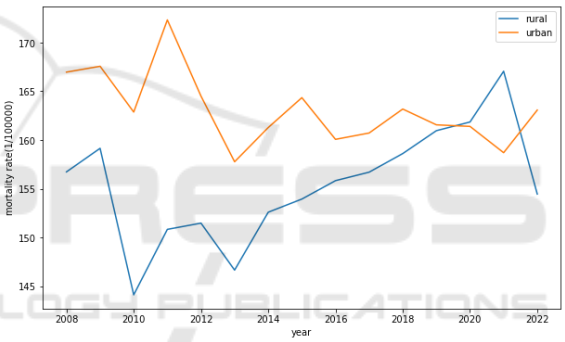

Figure 2: Overall urban and rural mortality rates (Picture credit: Original).

Urban areas show a general downward trend in the total mortality rate, with a predicted decrease of 2.33% in 2022 compared to 2008. Rural areas show a more mixed picture, with a 1.45% decrease in 2022 compared to 2008. Comparing the mortality rate of urban areas with the rural ones, the mortality rate of urban areas was higher than the rural areas until 2020, after which it reversed. Furthermore, between 2021 and 2022, urban and rural regions are expected to follow opposing trends, with urban areas rising and rural areas falling.

3.2.2 Predictions of Male/Female Mortality Rates in Urban and Rural Areas in 2022

Urban Area: The predicted male mortality rate for the urban area is 204.753, an increase of 2.33% from 2021. While the female mortality rate for the urban

area in 2022 is predicted to be 119.984, an increase of 2.76% from 2021. At the same time, the predicted value of the male mortality rate is 70.65% higher than the predicted value of the female mortality rate. The trend of male and female mortality rates in the urban area from 2008 to 2022 is shown in Figure 2, where the horizontal coordinate indicates the year and the vertical coordinate is the mortality rate (1/100,000):

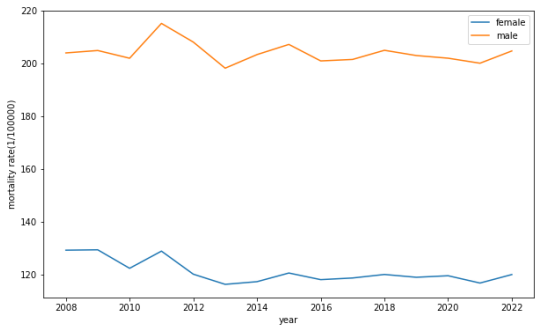

Figure 3: Male/female mortality rates in urban areas (Picture credit: Original).

Trends in male mortality rates are generally consistent with those of female mortality rates, both of which are on a downward trend. In the case of men, the predicted value of the mortality rate in 2022 is 0.37% lower than in 2008. The predicted value of the female mortality rate in 2022 is 7.15% lower than that in 2008. Male mortality is much greater than female mortality, and both are predicted to follow a similar upward trend between 2021 and 2022.

Rural Area: The predicted male mortality rate in rural areas is 197.797, which is 7.19% lower than the value in 2021. The predicted female mortality rate in rural areas in 2022 is 111.543, which is 6.35% lower than that in 2021. At the same time, the predicted value of the male mortality rate is 77.33% higher than the predicted value of the female mortality rate. The trend of male and female mortality rates in rural areas from 2008 to 2022 is shown in Figure 3, where the horizontal coordinate indicates the year and the vertical coordinate is the mortality rate (1/100,000).

Trends in male mortality rates are generally consistent with those of female mortality rates, both of which are on an upward trend. In the case of males, the predicted value of the mortality rate in 2022 is 3.32% lower than in 2008. The female mortality rate is predicted to increase by 4.19% in 2022 compared to 2008. The male mortality rate is much higher than the female mortality rate, and both are predicted to follow a similar downward trend between 2021 and 2022.

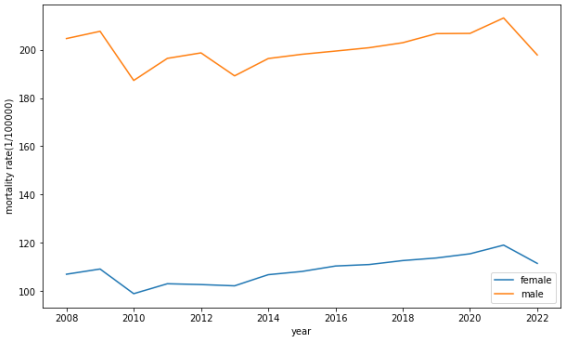

Figure 4: Male/female mortality rates in rural areas (Picture credit: Original).

4 DISCUSSION

The predicted trends in mortality rates for males and females in the different regions are consistent with the predicted trends of the total mortality rates in the corresponding regions, and the differences in the trends are only related to the differences in regional distribution, and not so much to the differences in gender. In general, the mortality rate in urban areas is decreasing, probably due to the development of medicine and increased health awareness. In rural areas, the mortality rate is on the rise, probably because the rural population aging problem is very serious (Rongshou et al 2018). Until 2020, the total mortality rate in urban regions was greater than in rural regions, but this trend reversed after 2020. This might be associated with the Human Development Index (HDI), a comprehensive assessment of regional human development in terms of health, education, and income, with free human development as the core concept (UNDP 1990). Research has shown that the HDI index is closely related to malignant tumors, and Khazaei discovered a positive correlation between the incidence of prostate cancer and the HDI index, and a negative correlation between the standardized death rate and the HDI index (Khzazei et al 2016). The HDI index in urban areas is higher than the value in rural areas, and rural areas lack medical services and cancer preventive understanding. All these factors can lead to the reversal of the mortality rate in rural areas and urban areas. The government can publicize cancer prevention in rural areas and promote healthcare reform to alleviate the imbalance of healthcare resources.

As for gender differences, the mortality rate of men is much higher than that of women, which may be due to the following reasons: (1) males and women have distinct physiological structures, and women

have greater estrogen levels than males, which may be a preventive factor for the development of some malignancies (Xiyi 2017); (2) the awareness of tumor prevention of women is stronger than that of men (Ni et al2019); (3) the level of exposure to risk factors of men is higher than that of women. For example, tobacco consumption and alcoholism are higher in men than in women. Chronic disorders such as hypertension and diabetes mellitus are also more prevalent in males than in women (Ye et al 2016 & Zengwu et l 2018). Moreover, there are differences in the cancer spectrum of different genders, for example, breast cancer and thyroid cancer are highly prevalent in women but have a better prognosis, whereas digestive tract cancers are highly prevalent in men but have a poorer prognosis, which may lead to a higher burden of malignant tumors in men than in women (Wei et al 2021). Therefore, men should raise their awareness of cancer protection and reduce their exposure to harmful substances such as tobacco and alcohol, which can effectively prevent cancer.

This research still has some limitations. First, the LSTM model requires a large amount of training data for training and prediction. Still, this research only studied the mortality-related data of malignant tumors during the 14-year period from 2008 to 2021, which is a small amount of data, and it may lead to inaccurate predictions. Second, the focus of the analysis in this research was mainly on regional and gender differences, and many other influencing factors were not taken into account, such as age, cancer spectrum distribution, and other factors. In the future, the model and analysis can be further refined in the above aspects and more detailed data are also indispensable to facilitate in-depth research on this topic.

5 CONCLUSION

The predictions in this research indicate that the total mortality rate and the male and female mortality rates in urban areas are likely to increase in 2022, whereas the corresponding mortality rates in rural areas are likely to decrease, and the differences in the trends are only related to regional differences. Second, the research analyzed the differences in the mortality rate by region and gender. For regional differences, the mortality rate in urban is declining, while the mortality rate in rural regions is growing, and after 2020, the value of the mortality rate in urban began to be lower than the corresponding mortality rate in rural areas. Possible explanations for the difference include a higher HDI index in urban regions, a deeper understanding of tumor prevention in urban areas than in rural ones, and a more serious population aging problem in rural areas. The difference by gender is reflected in the fact that the mortality rate for men is much higher than that for women. The reasons for this difference may include the different physiological structures of men and women, the higher awareness of tumor prevention in women than in males, the higher level of exposure to risk factors in men than in women, and the difference in the cancer spectrum. The analysis of the malignant tumor mortality rate, regional differences, and gender differences in this study can help to understand and predict the development trend of malignant tumors in China, uncover the pain points of malignant tumor deaths, and provide targeted preventive advice to the public from a multi-level perspective. However, this study also suffers from the limitations of too little data and not enough comprehensive investigation of factors. In the future, this research should make use of more sufficient data and improved models to include more factors into the investigation and conduct a multi-factor comprehensive analysis.

REFERENCES

H. Sung, J. Ferlay, R. L. Siegel, et al., “Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: a cancer journal for clinicians, vol. 71, pp. 209-249, 2021.

Z. Rongshou, Z. Siwei, Z. Hongmei, et al., “Cancer incidence and mortality in China, 2016,” Journal of the National Cancer Center, vol. 2, pp. 1-9, 2022.

Y. Gang, Y. Yanping, S. Chao, “Prediction of Mortality of Elderly Population with an Improved AE-LSTM Model,” Mathematical Theory and Applications, vol. 42, pp. 100, 2022.

S. Hochreiter, J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, pp. 1735-1780, 1997.

G. Han, S. Renju, M. Li, L. Wenyu, “A heart failure mortality prediction model based on AB-CNN-BILSTM,” Computer Applications and Software, vol. 38, pp. 37-42, 2021.

Z. Rongshou, G. Xiuying, L. Xueting, et al., “Analysis on the trend of cancer incidence and age change in cancer registry areas of China, 2000 to 2014,” Chinese journal of preventive medicine, vol. 52, pp. 593-600, 2018.

UNDP, “Human development report 1990,” New York and Oxford: Oxford University Press, pp. 10, 1990.

S. Khazaei, S. Rezaeian, E. Ayubi, et al., “Global prostate cancer incidence and mortality rates according to the human development index,” Asian Pacific Journal of Cancer Prevention, vol. 17, pp. 3791-3794, 2016.

J. Xiyi, H. Yunqing, Y. Ding, L. Qilong, C. Kun, J. Mingjuan, “Disparities of sex on cancer incidence and

mortality in Jiashan county, Zhejiang province, 1990-2014," Chinese journal of Epidemiology, vol. 38, pp. 772-778, 2017.

W.Ni, H. KaiYong, Y. Li, "Analysis on health literacy and its influencing factors about cancer prevention and control among urban residents in Guangxi," Chinese Journal of Disease Control & Prevention, vol.23, pp. 711-716, 2019.

R. Ye, Y. Qinghua, X. Jiying, et al., "Epidemiology of diabetes in adults aged 35 and older from Shanghai, China." Biomedical and Environmental Sciences, vol. 29, pp. 408-416, 2016.

W. Zengwu, C. Zuo, Z. Linfeng, et al., "Status of hypertension in China: results from the China hypertension survey, 2012–2015," Circulation, vol. 137, pp. 2344-2356, 2018.

C. Wei, C. Hongda, Y. Yiwen, L. Ni, C. Wanqing, "Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020," Chinese medical journal, vol. 134, pp. 783-791, 2021.