Ministry of Higher Education and
Scientific Research
University of Diyala
Department of Computer Science

Development of Prediction Model for Population Concentration using LSTM and GRU Algorithms

A thesis

Submitted to the Department of Computer Science \ College of the Science \ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Naseer Salah Abbas

Supervised By

Prof. ***Naji M. Sahib***

2021 A.D. 1442 A.H.

بسم الله الرحمن الرحيم

والذي هُوَ يُطْعِمُنِي وَيَسْقِينِ ﴿٧٩﴾

وَإِذَا مَرِضْتُ فَهُوَ يَشْفِينِ ﴿٨٠﴾

صدق الله العظيم

سورة الشعراء الاية (٧٩_ ٨٠)

Acknowledgment

First of all, praise is to GOD, the lord of the whole creation, on all the blessing was the help in achieving this research to its end.

I wish to express my thanks to my college (college of science), my supervisor, prof. Naji Motar Sahib for supervising this research and for the generosity, patience and continuous guidance throughout the work. It has been my good fortune to have advice and guidance from him. My thanks to the academic and administrative staff at the Department of the computer sciences.

Naseer Salah Abbas

Dedication

I would like to dedicate this work to:

To my candle that light my life My Mother.

To My father may God have mercy on him.

For his unlimited love, his support,

And Encouragement for me,

About everything in my life.

He is my first and last success.

To My Brothers and Sister.

To All My Friends.

I produce this work with all my love….

Naseer Salah Abbas

Abstract

The health crisis that attributed to the quick spread of the COVID-19 has impacted the globe negatively in terms of economy, education and transport and led to the global lockdown. The risk of the COVID-19 infection has been increased due to a lack of a successful cure for the disease. Thus, social distancing is considered the most appropriate precaution measure to control the viral spread throughout the world.

In this study, a model was proposed for predicting the people's movement to find out the proportion of social distancing in the short term (one day) using deep learning algorithms to take the necessary measures and precautions to control the COVID-19 infection. The proposed model consists of four phases: data collection, data pre-processing phase, prediction and evaluation stage and comparison phase. The dataset is obtained from 428 mobility reports, collected based on data from users selected for their Google Account location history for a country such as Iraq from 2020-02-15 to 2021-04-17 (428 days) stored in a comma-separated values file (CSV). Then, deep learning algorithms Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and hybrid model (GRU & LSTM) are applied to pre-processed data to predict the people's movement. They are compared using statistical measures: Mean absolute error (MAE) and root mean square error (RMSE) for performance measurement of these machine learning algorithms, where the lower the error rate, the better and more accurate the prediction performance of the model.

The results of the GRU are the sum of MAE = 0.4284 and sum of RMSE = 0.6479 for predict people's movement with training time equal to 35.951 sec, while the results of the hybrid model are the sum of MAE = 0.4360 and sum of RMSE = 0.6558 for prediction and the training time equal to 71.190 sec and the results of the LSTM are the sum of MAE = 0.4418 and sum of RMSE = 0.6618 for prediction and the training time equal to 123.799 sec.

These statistical measurement values indicate proposed model GRU outperformed all other models, it showed a solid performance to predict person path and movement in coronavirus pandemic and took little time to train compared to other algorithms, while the hybrid algorithm showed good performance and a short time in training compared with the LSTM model.

Social distancing and mixing between people play a vital role in the rapid spread of the COVID-19 pandemic. Therefore, our forecasted results of future trends of people's movement in society are very helpful for the country to control the pandemic and for purposes of social distancing guidance.

List of Contents

List of Abbreviations

Abbreviation	*Description*
Adam	*Adaptive Moment Optimizer*
AI	*Artificial Intelligence*
ANN	*Artificial Neural Network*
BPTT	*Backpropagation Through Time*
COVID	*Coronavirus*
CSV	*Comma Seprated Values*
CT	*Computed Tomography*
DL	*Deep Learning*
GRU	*Gated Recurrent Unit*
IoT	*Internet of Things*
LSTM	*long Short-Term Memory*
MAE	*Mean Absolute Error*
ML	*Machine Learning*
MSE	*Mean Square Error*
RMSE	*Root Mean Square Error*
RNN	*Recurring Neural Network*

List of Figures

List of Tables

List of Algorithms

Chapter One

Introduction

Chapter one

Introduction

1.1 Overview

Coronavirus (COVID-19) is a member of the Coronaviridae family of viruses. In every country around the world, a virus with no cure is wreaking havoc on people's lives, as well as financial and economic losses. The epidemic was declared a public health emergency and pandemic on January 30, 2020, by the World Health Organization (WHO). This virus causes exhaustion, dry cough, fatigue and respiratory problems among other things. Maintaining social distance is the safest way to avoid and slow down transmission. As a result, an automated detection system must be implemented to prevent the virus from spreading among people. The role of computer technologies in discovering Corona disease is large and effective. In the battle against the COVID-19 crisis, artificial intelligence is a key tool. Machine Learning (ML) and Deep Learning (DL) are two subdomains of artificial intelligence (AI). It has numerous uses in the areas of Computer Vision it aids in the diagnosis and prediction of that virus. Deep learning and machine learning techniques can be used to create alerts to maintain social distance, diagnose and treat COVID-19, track COVID-19 events, create dashboards, forecast and for another potential mechanism of control [1].

Coronavirus infection is the world's most global epidemic, since there is no effective vaccine or cure for this virus, non-pharmaceutical treatments such as contact tracing, hospitalization or self-isolation, quarantine, group lockout and social distancing are the only ways to reduce the spread of infection [2].

Machine Learning (ML) can be used to manage vast amounts intelligently and data forecast disease spread. To monitor the virus, forecast the epidemic's progress and devise strategies and policies to control its spread, cloud computing and machine learning can be used efficiently. Researchers will use Machine Learning (ML) and Artificial Intelligence (AI) to predict when and where the virus will spread and then alert those provinces to make the necessary preparations [3].

This thesis presents a model to predict the movement of individuals in society by using Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) algorithms, In addition to creating a hybrid algorithm from the two previous algorithms.

1.2 Related works

It is known that the Coronavirus appeared at the end of 2019 and continues until this thesis was written and there are many studies and research that appeared during this period dealing with the use of (ML) in contributing to finding solutions in the discovery and method of dealing with this virus or pandemic. As for finding the relationship between the people's movement and time, we did not find any direction for this type of research.

Many types of research that have used machine learning and artificial intelligence to contribute to finding solutions to this pandemic or virus as reviewed:

- **Wang, B. et al. (2019)**[4]**:** They used the Internet of Things (IoT) technology to monitor the acquired data, process the data, and predict the next data using a neural network. One algorithm proposed is a two-layer model prediction algorithm based on Long Short Term Memory Neural Network and Gated Recurrent Unit (LSTM & GRU). set a double-layer Recurrent Neural Network to predict the PM2.5 value.

This model is an improvement and enhancement of the existing prediction method Long Short Term Memory (LSTM). The experiment integrates data monitored by the IoT node and information released by the national environmental protection department. First, the data of 96 consecutive hours in four cities were selected as the experimental samples. The experimental results are close to the true value. Then, selected daily smog data from 2014/1/1 to 2018/1/1 as a train and test dataset. It contains smog data for 74 city sites. The first 70% of the data was used for training and the rest for testing. The results of this experiment show that the hybrid model can play a better prediction.

- **Ardabili, S. F. et al. (2020)** [5]**:** They compared machine learning and soft computing models to Susceptible-Exposed-Infectious-Recovered (SEIR) and susceptible-infected-recovered (SIR) models to predict the COVID virus outbreak. Two models out of a wide range of (ML) investigated showed promise. There has been strong generalization potential for longer-term predictions in the results of the adaptive network fuzzy (ANFIS) and multi-layered perceptron (MLPs) deference scheme. This study shows that (ML) is an efficient method to model an outbreak based on findings and the highly complex nature of the viral outbreak and its behavior variations from nation to nation. Work also shows that the integration of machine learning and SEIR models will achieve true innovation in outbreak prediction.
- **Tuli, S. et al. (2020) [4]:** To evaluate and forecast how the epidemic will grow to control the disease in the future, predict epidemic progression and develop strategies and policies to combat its spread and an improved mathematical model was used. To forecast the possible danger of the Coronavirus in countries around the world, an improved model based on machine learning (ML) has been used. They demonstrate that iterative weighting can be better suited to the

development of a prediction System utilizing a Generalized Inverse Weibull. This was used on a cloud computing platform to forecast the growth behavior of the epidemic more accurately and in real-time. A more accurate data-driven method can be very helpful for government and citizen proactiveness. Finally, they proposed a range of study possibilities and premises for further use.

- **Yadav, M. et al. (2020)** [6]**:** They suggested that different tasks connected to the current COVID-19 be investigated using a new SVM method. They also use the supported vectors in this work to obtain greater classification accuracy rather than just a regression line. The approach is evaluated and contrasted with other well-known regression models on the normal datasets available. The positive results reflect both success and accuracy.
- **Kasilingam, D. et al. (2020) [2]:** They used infrastructures, the environment, policies and independent variables associated with infections to predict early containment to create predictive supervised machine learning models. Data about infection with coronaviruses was used in 42 countries. Logistic regression findings indicate a positive significant relationship and signs of early containment between healthcare infrastructure and lock-down policies. Logistical regression, decision tree, random forest and SVM machine learning models are being built to demonstrate precision from 76.2% to 92.9% to forecast early symptoms of infection containment.
- **Lars, L. et al. (2020)** [7]**:** Used a successful sampling algorithm, they employed a system of temporary point processes and their model to quantify the impacts on the path of business restrictions, measures of social distancing and the disease of various tracing and testing methods. Based on this algorithm, Bayesian optimization was used to estimate the rate of transmission from infectious persons at the sites

they visited and their homes, as well as the reduction of mobility because of the social distancing from longitudinal case statistics.

- **Arun, S. et al. (2020)** [8]**:** Evaluated the effects of various testing and tracing techniques, social distance measures and business constraints, using an efficient sampling algorithm for the temporal points method. Based on this algorithm, Bayesian optimization was used to estimate the rate of transmission from infectious persons at the sites they visited and their homes, as well as the reduction of mobility because of the social distancing from longitudinal case data.
- **Barstugan, M. et al. (2020)** [9]**:** Detected on abdominal Computed Tomography (CT) images were carried out utilizing machine learning methods. Four separate datasets were generated for the detection of the Coronavirus by taking patches from the 150 CT images in the dimensions 16x16, 32x32, 48x48, 64x64. Patches were used to improve classification efficiency in the process of feature extraction. The function extraction approach was used with Discrete Wavelet Transform (DVT), Gray Level Size Zone Matrix (GLSZM), Grey Level Run Length Matrix (GLRLM), Local Directional Pattern (LDP) and Gray Level Co-occurrence Matrix (GLCM). The extracted features were graded by SVM. To assess classification results, F-score metrics, accuracy, specificity and sensitivity were used. With 10-fold cross-validation and the GLSZ M feature extraction process, the best classification precision was achieved as 99.68 %.
- **ArunKumar, K. et al. (2021)**[10]**:** They suggested state-of-art deep learning Recurrent Neural Networks (RNN) models to predict the country-wise cumulative confirmed cases, cumulative recovered cases and cumulative fatalities. The Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) cells along with Recurrent Neural Networks (RNN) were developed to predict the future trends of the

COVID-19 and used publicly available data from John Hopkins University's COVID-19 database and emphasize the importance of various factors such as age, preventive measures, and healthcare facilities, population density, etc. that play a vital role in the rapid spread of COVID-19 pandemic.

- **Koç, E. and Türkoğlu, M. (2021)** [11]**:** Suggested a deep learning solution focused on a deep long-term memory network to anticipate requests for health kits and the cases of Coronavirus outbreak numbers. A regression layer, a fully connected layer, a Multilayer LSTM network, a dropout layer and a normalizing layer is used for the proposed scheme. This model is used to predict the number of cases, the number of intensive care units and the amount of respiratory equipment during the coming days. A dataset with 77-days of Coronavirus data for sixty-eight days for training and nine days for testing was used to check the suggested method validity. The experimental outcomes showed the suggested MAPE values of (4,80 %, 3,29 %, 2,89 %), respectively (99,72 %, 99,85 %, 99,90 %) for the estimate of the cases, the respiratory kit number and intensive care beds.
- **Bodapati, S. et al. (2020)** [12]: Used proper learning models on time-series analysis, LSTMs and RNN was applied to predict future patterns such as the number of people recovered from COVID-19, virus deaths and reported positive viral cases number in the coming days. The data collection used was used for encouraging experimental outcomes.

1.3 Problem Statement

One of the important problems that emerged with the emergence of Covid-19 is the spread's speed and relationship with the people's movement and social distancing and find the relationship between time and people's

movement to determine crowded places to control and limit coronavirus spread speed by taking the necessary measures and choosing an appropriate forecast model for predicting social movement.

1.4 Aim of Thesis

This thesis aims to find the relationship between time and people's movement to controls coronavirus spread speed by predicting the people's movement and identifying crowded places (standard, high, low) to take the necessary measures by using recurrent neural network algorithms (RNN).

1.5 Outline of Thesis

Besides this chapter, the remaining parts of this thesis include the following chapters:

Chapter Two: Theoretical Background

In this chapter, the theoretical tools and techniques that were used in this thesis are presented.

Chapter Three: The Proposed Model

This chapter introduces the steps of the proposed prediction system, with its design and implementation.

Chapter Four: Experimental Results and Evaluation

This chapter presents the results of the proposed model implementation, analysis, testing and evaluates these results.

Chapter Five: Conclusions and Suggestions for Future Work

This chapter presents the conclusions of this work. Furthermore, it provides suggestions for future work.