Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

An Effective Diagnostic System for Identifying Covid-19 and Pneumonia Diseases Using Machine Learning Algorithms

Ramya Perumal[1] A. C. Kaladevi[2]
[1]Assistant Professor, Dept. of CSE, Sona College of Technology, Salem
[1]ramyaperumal@sonatech.ac.in
[2]Professor, Dept. of CSE, Sona College of Technology, Salem
[2]kaladeviac@sonatech.ac.in

Abstract

COVID-19, a pandemic disease causes a devastating effect on humans. The heavy loss of human lives triggers us to build an effective automated diagnostic system for identifying the disease with good accuracy. People get affected by this disease are increasing day by day to a greater extend. The availability of test kits is limited in numbers and its performance is not satisfactory. Hence the medical practitioners highly rely on radiological imaging. There is an immediate requirement for processing these enormous images that are generated daily to test the patient is infected with the disease or not. COVID-19 has similar characteristics with its related diseases such as viral pneumonia and bacterial pneumonia. Appropriately diagnosing the classes of diseases based on its severity is highly important in the medical domain. Our proposed system uses machine learning algorithms such Linear Support Vector Machine classifier and Logistic Regression classifier and provides remarkable results. The proposed system has experimented with 250 chest X-Rays from each classes of diseases such as COVID-19, viral pneumonia, bacterial pneumonia, and normal healthy subjects. The results are evaluated with performance measures such as Accuracy, Sensitivity, Specificity, F1-score, Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), and Error rate. Logistic Regression gives accuracy 93.4%, Sensitivity 94.1%, Specificity 92.9%, False Positive Rate 7.1%, False Negative Rate 5.9%., Positive Predictive Value 91.4%, Negative Predictive Value 95.1% and Error rate 6.6% which is comparatively better than Linear Support Vector Machine classifier that shows Accuracy 91.6%, Sensitivity 93.1%, Specificity 90.5%, False Positive Rate 9.5%, False Negative Rate 6.9%, Positive Predictive Value 87.1% and Negative Predictive Value 95% and Error Rate 8.5%.

Keywords: COVID-19, Machine Learning algorithm, Radiological imaging, Viral Pneumonia, Bacterial Pneumonia

I. INTRODUCTION

The first occurrence of Corona Virus is from Wuhan, China on December 31, 2019, and it gets progressed across the countries. According to the WHO report, there are nearly 100 million confirmed cases and 2 million death cases of COVID-19 around the world. It is named COVID-19 on February 11, 2020[7]. The spread of the disease happens through sputum produced by the patient and having close contact with them [5, 11]. The National Health Organization constantly insisting us to wear the mask and maintain social distancing to prevent the disease [5]. The Medical Practitioners detect the disease from the patient by inserting the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

test kit through his/her nasopharyngeal swab and obtain his/her biological material like blood or sputum then it is further tested by using real-time Reverse Transcription-Polymerase Chain Reaction (rRT-PCR). This conventional method has its drawbacks like the performance is highly dependent on sample preparation and clean environmental set-up and lacks in its availability. Even countries like Turkey and remote villages are lacking in medical facilities find it difficult to detect the COVID-19 disease [8, 10]. Hence the Medical Practitioners depend on radiological imaging such as CT-Scans and X-Rays. Among them, X-Rays are simple, low-cost and patient has undergone this treatment exposed to less radiation [6-8]. Processing the highly generated images and diagnosing various features from the images to identify diseases such as COVID-19, viral pneumonia and bacterial pneumonia is extremely tedious for the medical practitioner. It consumes time and requires much effort from them. To resolve this issue, we built an automated detection system that is highly efficient in producing precise results.

The evolution of the Machine Learning algorithm shows outstanding performance particularly in Image Processing, Object Detection, and Computer Vision. It outperforms in detecting diseases like brain tumors, cancer, and heart disease [10]. Our proposed system uses the machine learning algorithm where the system is feed with labeled chest X-Rays and it learns the data from the features in the images. The system is inputted with unseen Chest X-Rays, and then it categorizes the images into their classes of diseases. It simulates the working principle of the human brain. How the brain learns things by experience, the same way the system also learns from the inputted data.

II. RELATED WORKS

The existing systems mostly use advanced technology and a hybrid model for the detection of COVID-19 and its related disease. Advanced technology such as Deep Neural Network is used which constructs neural networks and adjusting the network parameters to minimizing the errors in detecting the classes of diseases. Particularly, Deep Convolution Neural Network is widely used for images. It consists of the convolution layer, pooling layer, and fully connected layer. During the convolution process, the filter is superimposed on the images in identifying the abnormalities present in the feature of an image. The filter size is small compared to the image size. The filter slices an image and extracts the features from it. Various activation functions are available to trigger the network. During the Pooling process, it reduces the number of features by highlighting the important features from an image. Finally, fully connected layers are used to flatten the output and use the soft-max function to find the maximum probability of the data points belonging to the classes of diseases. Also, existing works use Transfer Learning, a technique where images of a certain class of disease are few [6]. For example, Initially, COVID-19 images are few compared to other related diseases like pneumonia. The system which is used for identifying one particular disease is also used to identify another disease. The idea behind the concept of Transfer Learning is that it reuses the network for categorizing various diseases instead of building the network from its scratch [5-

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

10].Also the existing work uses hybrid model in categorizing the diseases but ours uses single classifier model for diagnosing the diseases [11-19]

III. PROPOSED SYSTEM

The working principle of the proposed system is that the chest X-Rays of 4 classes of diseases are given as input to the system. It consists of 3 modules. They are Pre-processing, Feature Extraction, and Machine Learning Classifier model. The block diagram of the proposed automated diagnostic system is shown in Fig.1.

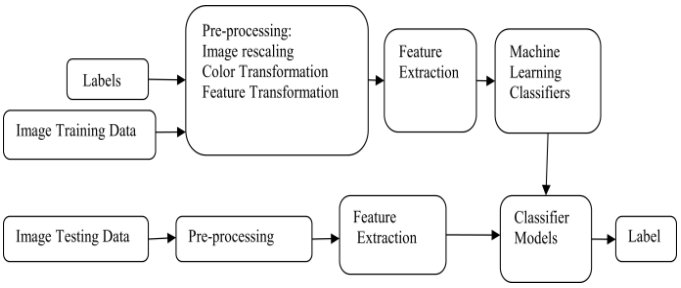

Fig.1. Block diagram of X-Rays diagnostic system

A. Pre-processing

It is the foremost step that should be performed before processing the images. The images are rescaled uniformly to the size of 224*224. It is of standard size for an image which greatly helps to have better image visualization. Images are rotated horizontally for further analysis. Then, the color image is transformed into a grey-scale image. The notable findings present in the features of an image are revealed because of color conversion. Finally, images are transformed into a pixel matrix for ease of analysis.

B. Feature Extraction

The proposed system uses KAZE an automated multi-scale 2D feature detection and description algorithm in nonlinear scale-spaces. The standard method for feature extraction on images is performed by using Gaussian filtering; a non-linear diffusion filtering method. It has disadvantages like reducing localization accuracy and distinctiveness, fail to detect natural boundaries of objects, treating equally the noise and details of an image, and has a blurring effect in Region Of Interest (ROI). To overcome these disadvantages, KAZE is used. It has advantages like blurring locally adaptive to the image data, reducing noise but retaining sharp boundaries, and gaining high localization accuracy and distinctiveness.

KAZE is the Japanese word that means wind. In nature, the wind is described as the flow of air on a larger scale and normally this flow is ruled by the non-linear process. The Same way the non-linear diffusion processes in the image domain. Non-linear diffusion approaches describe the evolution of luminance of an image through increasing scale levels as the divergence of certain flow function that controls diffusion process. These approaches are described by the non-linear partial differential equation. It is defined by

$$\partial L / \partial T = div(c(x, y, t). \nabla L) \quad (1)$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

Where div and ∇ are divergence and gradient operator. c is conductivity function. It is defined by

$$c(x,y,t)=g\ (|\nabla L\sigma\ (x,y,t)|) \quad (2)$$

where the luminance function ∇Lσ is the gradient of a Gaussian smoothed version of the original image L.

KAZE is the novel method that detects features in 2-dimensional and its Region of Interest (RoI) that exhibits maxima of the scale normalized determinant of hessian response through non-linear scale space. It includes the following steps,

The first step is the computation of non-linear scale-space that discretizing the scale space in logarithmic steps. Then, we need to transform the set of discrete scale levels in pixel units to time units by using the mapping function,

$$ti = (1/2)\ \sigma 2\ i,\ i= \{0...N\}, \quad (3)$$

second step is feature detection by using KAZE, we detect Feature of Interest FoI and calculate the response of scale normalized determinant of Hessian at multiple scale levels. The set of normalized differential operators are done to scale.

$$LHessian = \sigma^2 LxxLyy - L2xy \quad (4)$$

(Lxx, Lyy) are the second-order horizontal and vertical derivatives respectively and Lxy is second-order cross derivative. Feature detector computes maxima in scale and spatial location by analyzing the filtered images in non-linear scale-space at different scale levels.

The final step is finding dominant orientation and building descriptor to obtain rotation-invariant descriptor for that is necessary to estimate the dominant orientation in a local neighborhood centered at the keypoint location [20]

C. Machine Learning Classifier Model

The proposed system uses Machine Learning algorithms such as Linear Support Vector classifier and Logistic Regression classifier. It consists of two phases. They are the training phase and testing phase. During the training phase, the system learns the characteristics of the images and identifies abnormalities present in the features of an image. At the testing phase, the unseen images are categorized into their classes of diseases.

1. Linear Support Vector Machine classifier

Linear SVM is the supervised technique that is particularly used for classification. It draws a hyperplane to partition the data. But finding the optimal hyperplane is an important factor in this classifier model. The optimal hyperplane is the one that maximally separates the data points from its margin. The equation of the hyperplane is given by,

$$Y= wixi+b \quad (5)$$

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

Where Y is the output variable which is of categorical type, b is the bias parameter, xi is the input vector and wi is the weight vector.

If Y<1, then the data point belongs to the negative class. If Y>=1, then the data point belongs to the positive class

2. Logistic Regression Classifier

It is the supervised technique that is widely used for classification and regression. The classification is used to categorize the data points and regression is used to predict the numerical data point based on the attribute values. It uses the sigmoid function to compute the probability of the data points of different classes.

The hypothesis function of the classifier is given by,

$$Z = WX + B \quad (6)$$

$$h\Theta(x) = Sigmoid\ (Z)$$

$$Sigmoid\ (Z) = 1/\ (1+e^{-Z}) \quad (7)$$

If Z value goes ∞, then Y(Predicted) =1. Then the data point belongs to class 1

If Z value goes -∞, then Y(Predicted) =0. Then the data point belongs to class 0.

IV. DATASET COLLECTION

The proposed system has experimented with two data repositories; GitHub repository from which 250 COVID-19 images are obtained i.e. shared by Dr.Joseph Cohen at the University of Montreal. Another data repository is kaggle from which 250 bacterial pneumonia images, 250 viral pneumonia images, and 250 normal healthy subject images are obtained[3][4]. The image dataset comprises images of four class diseases that are split into an 80:20 ratio [7]. This train-test-split ratio is done to evaluate the proposed system how far it is good at finding the disease of an image. Our dataset consists of 250 images of each class disease that are randomly shuffled. In total,1000 images are experimented out of which 800 images are used as training data and 200 images are used as testing data.

V. EXPERIMENTAL SETUP

The proposed system is implemented in Googlecolab equipped with GPU for experimenting with the image dataset of different classes of disease. All computations are performed on Intel Core (TM) i5-2450M CPU @2.50GHz with 64bit windows 10 as the operating system. All the experiments are performed using the Python software package.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

VI. PERFORMANCE MEASURE

The proposed system is evaluated by using performance measures such as accuracy, sensitivity, specificity, False Positive Rate, False Negative Rate, True Positive Rate, True Negative Rate, and F1-score.

True Positive (TP) -Number of the positive cases that are correctly predicted by the classifier model

True Negative (TN) -Number of negative cases that are correctly predicted by the classifier model

False Positive (FP) -Number of cases that are predicted as positive by the classifier model but are negative in reality

False Negative (FN) -Number of cases that are predicted as negative by the classifier model but are positive in reality.

Accuracy is defined as the ratio of the total number of cases that includes both positive and negative cases which are correctly identified by the classifier model to the total number of cases.

Sensitivity is defined as the ratio of the total number of positive cases that are correctly predicted by the classifier model to the total number of positive cases in reality. It is also known as Recall or True Positive Rate.

Sensitivity=TP/ (TP+FN)

Specificity is defined as the ratio of the total number of negative cases that are correctly identified by the classifier model to the total number of negative cases in reality.

Specificity=TN/ (TN+FP)

Positive Predictive Value (PPV) denotes how many test positives are true positives. It is also known as precision.

PPV=TP/(TP+FP)

Negative Predictive Value (NPV) denotes how many test negatives are true negatives.

NPV=TN/(TN+FN)

F1-score is the harmonic average of precision and recall. It is given by,

F1-score= (2*Precision*Recall) / (Precision + Recall)

False Positive Rate (FPR) is defined as the ability of the classifier that mistakenly classifies the normal case as the pathological case.

FPR = FP / (FP + TN)

False Negative Rate (FNR) is defined as the ability of the classifier that mistakenly classifies the pathological case as the normal case.

FNR = FN / (FN + TP)

Error rate=1-Accuracy

VII. RESULT

The proposed system provides remarkable results in detecting the different classes of diseases such as COVID-19, Bacterial Pneumonia, Viral Pneumonia, and Normal Healthy subjects. The following table shows the performance measures of the proposed system.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

Table.1. Performance of the Proposed System

Model	Accuracy	Sensitivity	Specificity	FPR	FNR	PPV	NPV	Error Rate
LR Model	93.4%	94.1%	92.9%	7.1%	5.9%	91.4%	95.1%	6.6%
Linear SVC model	91.6%	93.1%	90.4%	9.5%	6.9%	87.1%	95%	8.5%

The performance of the proposed system is shown in Fig.2. The confusion matrix of the Linear Support Vector Classifier (SVC) and Logistic Regression are shown in Fig.3 and Fig.4 respectively. The Linear SVC classifier takes 11.55 minutes to converge the data to the network. The Logistic Regression takes only 55 seconds to converge data to the network

.

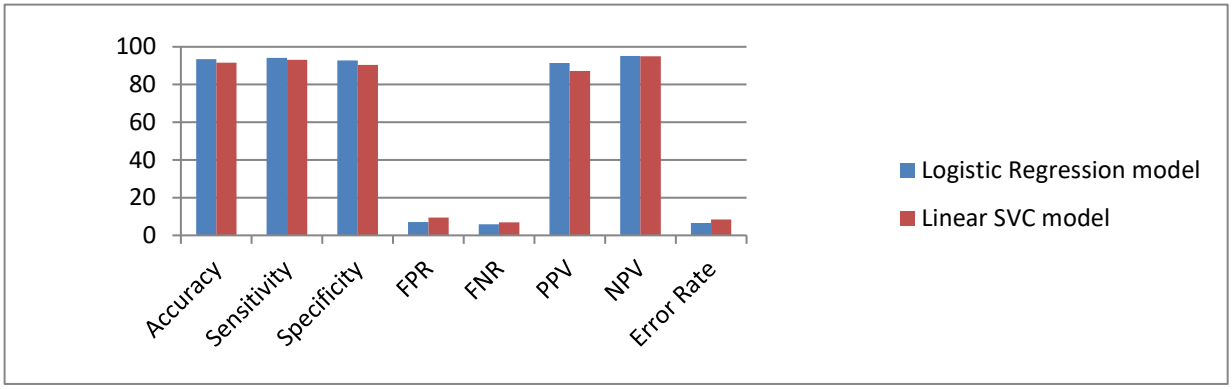

Fig.2. Performance measure of the proposed system

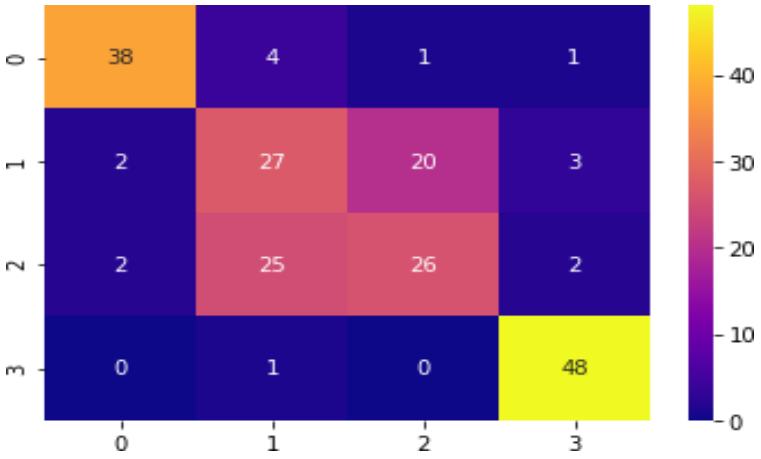

Fig.3. Confusion matrix of the Linear SVC classifier

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

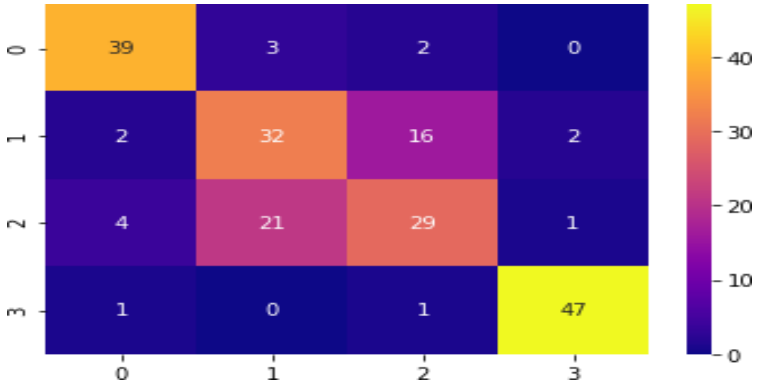

Fig. 4. Confusion matrix of the Logistic Regression classifier

VIII. CONCLUSION

In the current scenario, it is a timely requirement for conducting the preliminary test to check the person is infected with COVID-19, Bacterial Pneumonia, Viral Pneumonia, and Normal Healthy subjects. Our proposed system acts as an effective system for diagnosing the classes of diseases with precise results. The proposed system uses two machine learning classifiers such as Logistic Regression classifier and Linear Support Vector Classifier. Logistic Regression gives superior result with accuracy 93.4%,Sensitivity 94.1%,Specificity 92.9%,False Positive Rate 7.1%,False Negative Rate 5.9%.,Positive Predictive Value 91.4%, Negative Predictive Value 95.1%, and Error rate 6.9% compared to Linear Support Vector Classifier.

In the future, we plan to extend our work using a greater scale of data with explainable features to distinguish the COVID-19 from other pneumonia disease [10]. We also plan to investigate the data with different Deep Neural Network models to perform the comparative analysis on detecting COVID-19 and related pneumonia diseases [9].

Reference

[1] Tom M Mitchell Machine learning Publisher Mc Graw-Hill Science or Engineering
[2] https://www.who.int,/emergencies/diseases/novel-coronavirus- 2019
[3] Image dataset www.github.com
[4] Pneumonia and normal Image dataset www.kaggle.com
[5] Fatima M. Salman1, Samy S. Abu-Naser1, Eman Alajrami2, Bassem S. Abu-Nasser1, Belal A. M. Ashqar1 "COVID-19 Detection using Artificial Intelligence", Vol 4, Issue:3 March 2020
[6] Ioannis D. Apostolopoulos1 · Tzani A. Mpesiana2," Covid-19: automatic detection from X-ray images utilizing transfer learning with convolution neural networks" https://doi.org/10.1007/s13246- 020-00865-4
[7] Ezz El-Din Hemdan1, Marwa A. Shouman1, and Mohamed Esmail Karar 2," COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images"
[8] Talha Burak Alakus, Ibrahim Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection July 2020
[9] Md Zabirul Islam, Md Milon Islam, Amanullah Asraf, "A combined deep CNN-LSTM network for the detection of novel corona virus (COVID-19) using X-Ray images", August 2020

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 9078 - 9086
Received 16 February 2021; Accepted 08 March 2021.

[10] Tulin Ozturka, Muhammed Talob, Eylul Azra Yildirimc, Ulas Baran Baloglud, Ozal Yildirime, U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images"April 2020
[11] Samuel Lalmuanawma, Jamal Hussain, Lalrenfela, ChhakChhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review"Jan 2020
[12] Akib Mohi Ud Din Kanday, Sayed Tanseel Rabani, Qumaar Rayees Khan, "Machine learning-based approaches for detecting COVID- 19 using clinical text data"June 2020
[13] Haochen Yao, Nan Zhang, Ruochi Zhang, Meiyu Duan, Tianqi Xie,jiahui Pan, Ejun Peng, Juanjuan Huang, Yingli Zhang, Xiaoming Zhu, Hong Xu, Fenngfeng Zhou, Guoqing Wang, "Severity Detection for the Corona virus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests", July 2020
[14] Luca Brunese, Fabio Martinelli, Francesco Mercaldo, "Machine learning for corona virus COVID-19 detection from chest X-Ray"
[15] Sheetal Rajpal, Ankit Rajpal, Navin Lakhayani, Naveen Kumar, "COV-ELM classifier: An Extreme Learning Machine based identification of COVID-19 using Chest X-Ray Images"
[16] Armando Ugo Cavallo, Jacopo Troisi, Marco Forcina, Piervalerio Mari," Texture Analysis in the Evaluation of Covid-19 Pneumonia in Chest X-Ray Images: a Proof of Concept Study"
[17] Ali A. AL-Bakhrani, Husam H. Abdulmughni, Ahmed A. Hamoud, Soha Alrajjou, Ramesh Manza, Ratnadeep R. Deshmukh, "Machine Learning and Deep Learning to Do Early Predictions of COVID-19 Infection Using Chest X-Ray Images"
[18] Abolfazl Zarkari Khuzani, Morteza Heidari, S.Ali Shariati, "COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest x-ray images", October 2020
[19] Muhammad Imad1, Naveed Khan2, Farhat Ullah3, Muhammad Abul Hassan4, Adnan Hussain5 & Faiza 6, "COVID-19 Classification based on Chest X-Ray Images Using Machine Learning Techniques", October 2020
[20] Pablo Fern´andezAlcantarilla1, Adrien Bartoli1, and Andrew J. Davison2, "KAZE Features" A. Fitzgibbon et al. (Eds.): ECCV 2012, Part VI, LNCS 7577, pp. 214–227, 2012.
[21] Coronavirus Disease 2019 (COVID-19) Symptoms". Centers for Disease Control and Prevention. United States.
[22] Coronavirus (COVID-19) Mortality Rate". www.worldometers.info. 5 March 2020. Retrieved 23 March 2020.