Classification for Covid-19 Diseases Based on Ensembled Models

Guanchen Xiao*

College of Arts and Science Boston University Boston Massachusetts United States

*Corresponding author email: gcxiao@bu.edu

Abstract. After Covid19 became a worldwide health issue, rapid diagnosis based on clinical symptoms from many diseases with similar symptoms to Covid19 became important to slow down the spread of the pandemic. This study attempts to find ways to classify and diagnose diseases with the help of computer technology quickly and accurately. In this study, the author developed an ensembled machine learning model to categorize four diseases using information on their distinct clinical signs. The authors used a Support Vector Machine (SVM) and Artificial Neural Network (ANN)-based ensemble model. To improve the accuracy of the classification result, this model adds the strong classifier SVM to the result of an intermediate hidden layer of the fundamental ANN deep learning model. The result of the study shows that the integrated model's prediction performance is significantly better than that of the original ANN model after the support of another strong classification algorithm. In conclusion, the effectiveness of the proposed method was proved for classifying the symptoms of patients with allergies, colds, flu, and Covid-19 in this study.

Keywords: Machine Learning, Convid-19, Deep Learning.

1. Introduction

Almost three years have passed since the World Health Organization declared Covid-19 a pandemic. Over the past two to three years, Covid-19 has significantly changed the globe. Global production and commerce have suffered greatly as a result of the epidemic, which presented a serious danger to the expansion of the global economy. The impact of the pandemic on the world is not limited to the economy. According to WHO data updated on September 7, 2022, Covid19 has caused more than 600 million confirmed diagnoses and more than 6.4 million deaths worldwide. Facing a pandemic that is still spreading today, an efficient and accurate diagnosis becomes especially important because it can effectively slow the spread of the virus and allow patients to receive more timely treatment [1].

An important method of diagnosing Covid19 is through Computed Tomography (CT). Doctors can confirm the diagnosis of the virus by looking at CT of the lungs and finding unique lung features in patients with a confirmed Covid19 diagnosis [2], reference number should be in order]. However, this approach is particularly ineffective since it necessitates both the patient's attendance at the hospital for a chest CT scan and the physician's analysis of each patient's CT data. Besides taking a chest CT, one of the most mainstream method for Covid19 detection is by Polymerase Chain Reaction (PCR). Although PCR no longer requires many experienced physicians to be involved in the diagnosis compared to CT, this method also has some disadvantages. In addition to the sample requirements and expensive equipment, the long reaction time required for PCR is an important issue [3]. The efficiency of PCR is relatively poor in situations where a large number of samples need to be tested in a short period of time.

Computer techniques as well as machine learning have become the new direction of researchers in the face of the problem of processing huge samples in a short period of time. By building machine learning models to assist in diagnosis, people can process patient data faster and draw conclusions. Soham et al. proposed the use of support vector machine to predict Covid19 [4]. They built SVM models and effectively distinguished between uninfected, mildly infected, and severely infected patients by training and testing the models. However, their study was limited to a small number of symptoms and a single virus. Narges and Elahe take a step further in ordinary machine learning. They

used an artificial neural network approach to build a deep learning model that effectively identifies Covid19 positive patients by collecting clinical symptoms [5]. Although they use a deep learning method, they still base their studies on clinical data from a single disease and a single model.

Table 1: The Sample Data in the Collected Dataset.

Feature name	Data-1	Data-2	Data-3
COUGH	0	0	1
MUSCLE_ACHES	1	1	1
TIREDNESS	0	0	1
SORE_THROAT	1	0	0
RUNNY_NOSE	0	0	0
STUFFY_NOSE	0	0	0
FEVER	0	0	0
NAUSEA	0	0	0
VOMITING	0	0	0
DIARRHEA	0	0	0
SHORTNESS_OF_BREATH	0	0	0
DIFFICULTY_BREATHING	1	1	1
LOSS_OF_TASTE	0	0	0
LOSS_OF_SMELL	1	1	1
ITCHY_NOSE	0	0	0
ITCHY_EYES	0	1	0
ITCHY_MOUTH	1	1	0
ITCHY_INNER_EAR	0	1	0
SNEEZING	1	1	1
PINK_EYE	0	0	1
TYPE	1	1	1

To solve the limitation mentioned above, the study will demonstrate the results of combining multiple machine learning models and the efficient differentiation of several similar diseases. This study will consider integrating models to combine the learning and prediction results of artificial neural network and support vector machine and draw conclusions. The study presented in this paper uses ensembled model, which effectively allows some of the errors in artificial neural network to be corrected by support vector machine algorithm. At the same time, the model considers larger number of possible features of Covid19 and more common features of the virus, thus allowing for a preliminary and effective diagnosis of the disease without going to the hospital.

2. Dataset and data pre-processing

Table 2: The Performance of Various Models.

Models	Accuracy
Support Vector Machine	92.891688%
Naïve Bayes	92.722978%
Random Forest	92.812956%
Artificial Neural Network	92.936412%
Ensembled Method (Artificial Neural Network + Support Vector Machine)	93.671265%

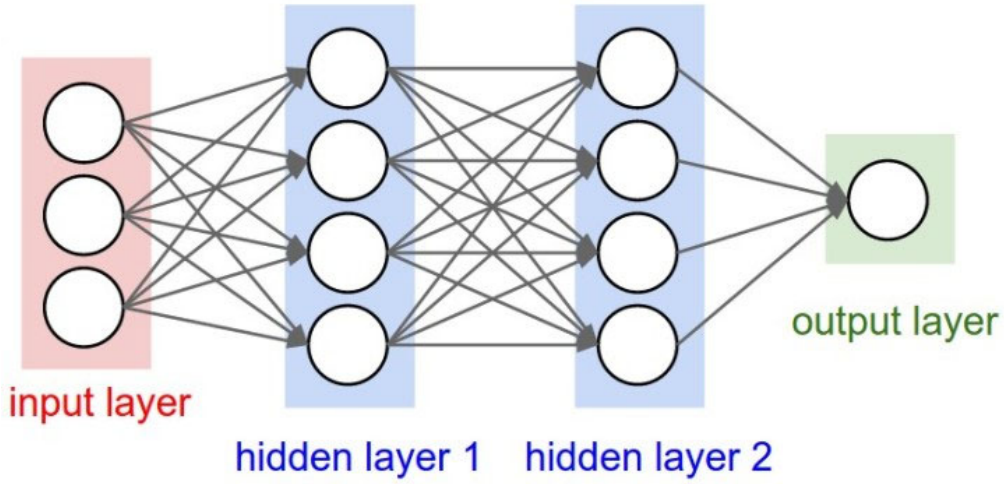


Figure 1: The Architecture of the ANN Model [7].

In this paper, a dataset on the Kaggle platform that recorded the symptoms of patients with allergies, colds, flu, and Covid-19 was used [6]. This dataset contains 44,452 patients with 20 different types of symptoms. Table 1 provides a part of sample data used in this study. All symptoms are represented by 1 and 0 for present and absent, and the last column represents in text form the patient's final diagnosed disease, which are allergy, cold, Covid-19, and flu. Diagnostic results are used as target variable in the model, so this study first processed the text content in the last column. This paper used the Label Encoder method to convert allergy, cold, Covid-19, and flu to 0, 1, 2, and 3, so the machine learning algorithms can be successfully applied the later.

3. Methods

Ensemble methods are used to build as well as train models in this study. An ensemble model combines the results of several machine learning algorithms. A single model has some probability of prediction error. The combination of multiple models can increase the percentage of successful predictions to some extent, because some prediction errors can be corrected in the integration step of this method. The ensemble method used in this study is stacking.

The dataset was split into a ratio of 80% training dataset and 20% test dataset. Four different machine learning algorithms will be trained based on the same training dataset, which are support vector machine, naïve Bayes, random forest, and neural network.

3.1 Support Vector Machine

Support vector machine (SVM) is a supervised machine learning method that can effectively classify examples into different categories. The SVM method looks for ways to separate the points of two categories in a plane with the maximum interval by training the model. The randomly selected training dataset belong to different categories, so they can be separated using the SVM model. Two parallel hyperplanes are selected at the boundary of the two groups of data points and the distance between them is calculated. The computer will continuously calculate and come up with the maximum distance between the two hyperplanes. A linear SVM model is used. In the SVM model used in this study, the kernel is linear, and the penalty parameter of the error term is 100.

3.2 Naïve Bayes Algorithm

Naïve Bayes algorithm is a classification method based on Bayes' Theorem. Different features in the dataset are considered independent. By applying naïve Bayes algorithm to training dataset, the computer calculates the conditional probability of each kind of diseases based on 20 independent features(symptoms). The one with highest conditional probability is output as the result.

3.3 Random Forest Method

The random forest approach derives classification results by integrating the learning results of many decision trees. Different decision trees produce different predictions for the outcome of each patient's diagnosis based on 20 features. After each decision tree reaches a conclusion, the conclusion with the most occurrences become the final output. After several testing and tuning processes, the parameters of the random forest model are as follows: number of estimators = 68, maximum depth = 20, minimum samples leaf = 20, minimum samples split = 10.

3.4 Artificial Neural Network

Artificial neural network (ANN) shown in Figure 1 is a deep learning method that acts similar to human brain, which is widely used in many studies [8-10]. The patient's symptoms as well as the diagnostic result will be used as data input and will be computed in the hidden layer through several algorithms to find similar features, learn the similarities and patterns, then output. Like the human brain, this algorithm inputs and learns the similarities of each disease's symptoms and thus differentiates between different viruses. In this study, the neural network model has four hidden layers with 100, 60, 40, and 25 neurons. The output layer had 4 units with activation function "Softmax". The whole model will be trained 5 times.

3.5 Ensembled Model

Ensembled model increase the accuracy of predictions by combining and learning from multiple models. The ensembled learning model used in this paper combines artificial neural network and support vector machine model. Based on the artificial neural network model mentioned in 4), the authors extract a hidden layer from the trained ANN model and add a support vector machine model to the results of the operations in this layer. After the optimization of the neural network model by SVM model, the accuracy of prediction can be improved. After tunning several parameters, when the number of neurons in the four hidden layers of ANN is 100, 80, 40, and 25, respectively, a significant improvement in the prediction accuracy is observed by extracting the hidden layer with 80 neurons and adding an SVM model after it.

4. Result and Discussion

By adjusting the parameters, each model has an accuracy of over 90% for disease prediction. The experimental result of each model can be found in Table 2. Support Vector Machine has the highest accuracy among the three machine learning models, and Naïve Bayes has the lowest accuracy. As a

deep learning model, the Artificial Neural Network has a higher accuracy rate of about 92.94%, which is also the most accuracy model among the non-ensembled methods.

By adjusting the parameters, each model has an accuracy of over 90% for disease prediction. Support Vector Machine has the highest accuracy among the three machine learning models, and Naïve Bayes has the lowest accuracy. As a deep learning model, the Artificial Neural Network has a higher accuracy rate of about 92.94%, which is also the most accuracy model among the non-ensembled methods.

The ensembled model proposed by the author combines Artificial Neural Network and Support Vector Machine, which has significantly higher accuracy than other models. In particular, the four hidden layers of Artificial Neural Network have 100, 80, 40, and 25 neurons, respectively, and the author add a linear Support Vector Machine model to the second hidden layer. By adding this strong classifier, the prediction accuracy of the model is improved by 0.74%.

The reason why the classification accuracy of the ensembled model is higher than that of a single artificial neural network is because the support vector machine is a strong classifier for the classification of clustered points. By calculating the maximum distance between two sets of clustered points, the SVM can better classify the data based on the training result of ANN. The reason is that after the classification and learning through ANN, the dataset has clearer margin, and less noise, SVM can perform better.

5. Conclusion

In this study, the author built an ensembled machine learning model to classify four diseases by learning various clinical symptoms of four different diseases. The authors applied an ensembled model consisting of ANN and SVM. This model added the strong classifier SVM to the deep learning results to make the classification results more accurate. The prediction result of the integrated model is 0.74% higher than the original ANN model.

However, the model proposed by the authors has not been tested against datasets with a wider variety of diseases. In the meantime, there are still ways to improve the existing model. In the future, the author plan to add additional classification models to the existing ensembled model and try to generate more accurate classification results. In addition, the interaction of different machine learning models in the integrated model is yet to be investigated.

References

[1] Angel N, Patrick S. 2022. Coronavirus Disease 2019 (COVID-19) and Diagnostic Error. Patient Safety Network. https://psnet.ahrq.gov/primer/coronavirus-disease-2019-covid-19-and-diagnostic-error

[2] Chung M. et al. et al. 2020. CT Imaging Features of 2019 Novel Coronavirus (2019-nCoV). Radiology. 295(1):202-207. doi:10.1148/radiol.2020200230

[3] Rong G. et al. 2021. COVID-19 Diagnostic Methods and Detection Techniques: A Review. Reference Module in Biomedical Sciences. B978-0-12-822548-6.00080-7. doi:10.1016/B978-0-12-822548-6.00080-7

[4] Soham G. et al. 2021. 18 - A novel approach to predict COVID-19 using support vector machine. Data Science for COVID-19. Academic Press. Pages 351-364. ISBN 9780128245361,https://doi.org/10.1016/B978-0-12-824536-1.00014-9.

[5] Roustaei N, et al. 2021. COVID-19 screening: use of an artificial neural network. BMJ Supportive & Palliative Care. doi: 10.1136/bmjspcare-2021-003391

[6] Walter Conway. 2021. COVID, FLU, COLD Symptoms. Kaggle. https://www.kaggle.com/datasets/walterconway/covid-flu-cold-symptoms

[7] Diginaltrends. 2019. What is an artificial neural network? Here's everything you need to know. https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/

[8] Yang T et al. 2022. Named Entity Recognition of Medical Text Based on the Deep Neural Network. Journal of Healthcare Engineering, 2022.

[9] Yu Q et al. 2022. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72, 103323.

[10] Song S et al. 2022. A new convolutional neural network architecture for automatic segmentation of overlapping human chromosomes. Neural Processing Letters, 54(1), 285-301.