Early Diagnosis Prediction From COVID-19 Symptoms Using ANN-Based Machine Learning Method

Charlyn V. Rosales, Bulacan State University, Philippines

The Southeast Asian Conference on Education 2024
Official Conference Proceedings

Abstract

Timely diagnosis of COVID-19 is crucial to mitigate the risk of virus transmission. Traditional diagnostic methods, such as medical laboratory and antigen tests, while effective, are not always easily accessible. This study proposes an innovative approach to detect COVID-19 promptly using Artificial Neural Networks (ANN), eliminating the need for laboratory tests. By analyzing an individual's current symptoms, the ANN serves as a powerful tool for early diagnosis. The dataset employed in this research was sourced from Kaggle, specifically the COVID-19 presence and symptoms dataset. To enhance data pre-processing and hyperparameter tuning, GridSearchCV was utilized, incorporating 10-fold cross-validation. The optimal configuration, derived from these procedures, facilitated the construction of an effective prediction model using ANN. The findings reveal that hidden layer sizes of (100,), (50, 100, 50), and (50, 50, 50), coupled with relu and tanh activation functions, adam solver, alpha values of 0.05 and 0.0001, and adaptive or constant learning rates, collectively achieved the highest algorithm performance. Employing this optimal configuration, the ANN-based prediction model demonstrated an impressive 98.84% accuracy, 98.69% specificity, 100% sensitivity, and a 98.84% ROC curve. This developed prediction model holds the potential to revolutionize COVID-19 detection by enabling real-time identification of the disease without the reliance on laboratory tests. Applications utilizing this model could significantly contribute to early intervention and prevention strategies, ultimately reducing the spread of the virus in the community.

Keywords: Algorithms, Artificial Neural Networks, COVID-19, Early Diagnosis, Machine Learning, Hyperparameter Optimization

iafor

The International Academic Forum
www.iafor.org

Introduction

COVID-19, caused by the SARS-CoV-2 virus, is a highly contagious disease with respiratory symptoms ranging from mild to severe, particularly in individuals with underlying comorbidities (Coronavirus, 2024). The global impact of the disease is evident, with 703,961,073 confirmed cases and 7,004,395 recorded deaths as of March 7, 2024 (Covid Live Update, 2024). Prevention measures recommended by the World Health Organization (WHO) include social isolation, mask-wearing, frequent cleaning, hand hygiene, and vaccination (Coronavirus, 2024).

Early detection is crucial to curbing the spread of COVID-19, as evidenced by the potential for severe symptoms and the risk of outbreaks. Traditional methods, such as COVID-19 RT-PCR testing, though effective, are often expensive and contribute to testing kit shortages and delays (Villavicencio et al., 2021). Addressing this challenge, deep learning models, specifically the Multilayer Perceptron (MLP) artificial neural network, offer promise for early detection applications.

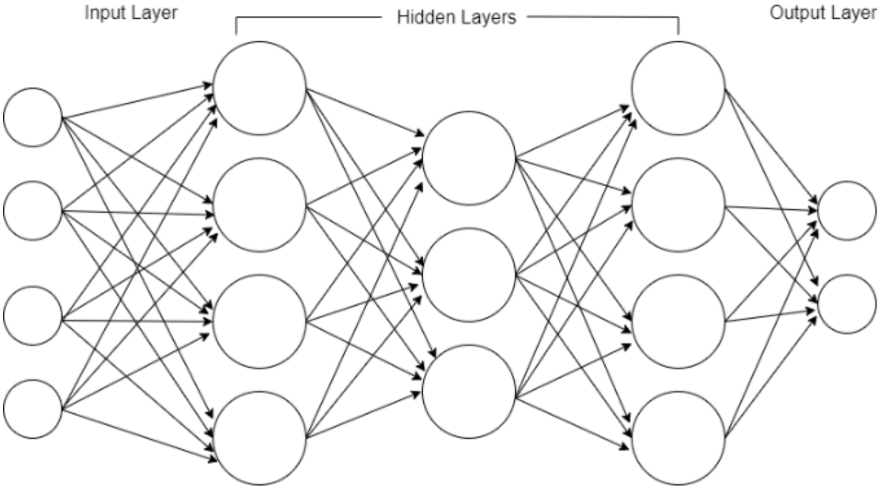

Figure 1: This is an image

In this study, an MLP architecture was employed due to its efficacy in both classification and regression tasks (Itano et al., 2008). Figure 1 illustrates a simplified representation of the MLP structure, featuring input, hidden, and output layers inspired by the human brain's interconnected network.

The depicted ANN structure involves four neurons in the input layer, representing predictors from the symptom dataset. Three hidden layers, with four, three, and four neurons respectively, process and classify the data to determine the potential presence of COVID-19 in an individual. Recognizing the challenges faced by the medical sector during the pandemic, the study aims to develop an ANN-based deep learning prediction model for real-time, early diagnosis of COVID-19. This model seeks to provide a rapid and cost-effective alternative to traditional laboratory tests, aiding in the prompt identification and management of COVID-19 cases.

Materials and Methods

The methodology of this study can be visualized in Figure 2.

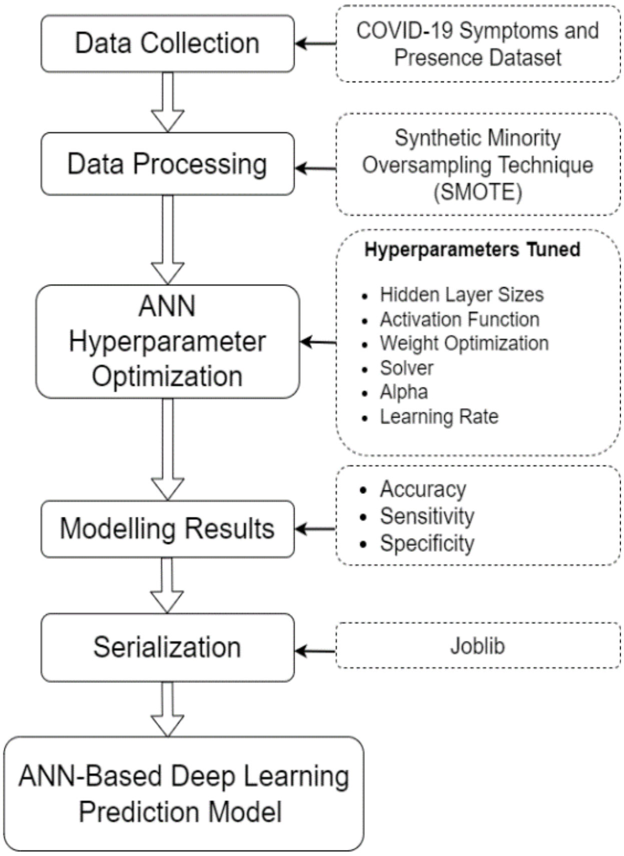

Figure 2: Methodology

Methodology details are illustrated in Figure 2, outlining the process of gathering data from the Kaggle dataset named COVID-19 symptoms and presence dataset. This publicly available dataset comprises 20 variables, with one target variable indicating COVID-19 positivity or negativity.

Acknowledging the 4:1 class imbalance in the dataset (Villavicencio et al., 2021), Synthetic Minority Oversampling Technique (SMOTE) was implemented to address this imbalance. SMOTE mitigates prediction bias by oversampling the minority class through the addition of synthetic samples based on randomly selected neighbors (Chawla et al., 2002). This adjustment ensures a more balanced representation of COVID-19 positive and negative instances, enhancing the classifier's ability to avoid bias in predictions.

Subsequent to data processing, the prepared dataset underwent the training phase of the prediction model. The optimized hyperparameters included hidden layer sizes, alpha, learning rate, activation function, weight optimization, and solver, seeking the optimal performance of the Artificial Neural Network (ANN) algorithm. Following hyperparameter optimization, the modeling process was executed, incorporating three statistical measures and scores to describe the model's performance.

Accuracy, defined as the ratio of correct predictions to the total number of predictions made by the model (Villavicencio et al., 2021), was one of the statistical measures considered. Sensitivity, or the true positive rate (TPR), represented the ratio of correct predictions of true samples over all positive samples in the dataset (Ul Haq et al., 2019). Additionally, specificity, or the true negative rate (TNR), denoted the ratio of correct negative sample predictions to all negative samples in the dataset (Ul Haq et al., 2019). These measures collectively provided a comprehensive evaluation of the ANN algorithm's predictive performance.

After completing the modeling phase, the model was serialized using the Joblib package. This tool facilitated the compilation of the model into a file object, making it applicable for integration into various machine learning-based applications. Consequently, the ANN-based Deep Learning prediction model has been successfully developed and is now poised for deployment.

Results and Discussion

The dataset initially presented a 4:1 class imbalance, consisting of 4383 positive and 1051 negative samples. Employing the SMOTE dataset balancing method increased the total samples to 8766, ensuring an equal distribution of 4383 instances for both COVID-19 positive and negative classes.

Subsequently, the balanced dataset was divided into training and testing sets using a 7:3 ratio, resulting in 6136 samples for training and 2630 samples for testing. This prepared dataset underwent hyperparameter tuning during the modeling phase to optimize the Artificial Neural Network's (ANN) configuration for maximum accuracy.

Hyperparameter tuning was executed using the GridSearchCV method from the sklearn model selection package, involving the modification of predefined hyperparameter values. The 10-fold cross-validation resampling approach was applied in all experiments. The outcome included the recorded hyperparameter values and the accuracy of the generated classifier, with the best ten results.

Among the various hyperparameter configurations, five setups emerged as top performers, achieving the highest accuracy of 98.84%. The optimal configuration for the ANN algorithm, as determined through hyperparameter tuning, is detailed below: hidden layer sizes of (100,), (50, 100, 50), and (50, 50, 50), relu and tanh activation functions, adam solver, alpha values of 0.05 and 0.0001, and constant and adaptive learning rates.

Following hyperparameter tuning, the prediction model was constructed, with the tuned hyperparameters. This refined model underwent testing on both the training and testing datasets, delivering robust performance with 98.84% accuracy, 100% sensitivity, and 98.79% specificity. These results establish a strong foundation for the development of machine learning-based applications designed to facilitate early COVID-19 diagnosis in individuals.

The results suggest that the ANN MLP is a viable algorithm for constructing a COVID-19 prediction model using symptoms as predictors. Subsequent to model development, the dump() function from the joblib package was employed to serialize the prediction model into a file object, facilitating seamless integration into various machine learning-based applications.

Conclusion

The primary objective of this research is to design an Artificial Neural Network-based Deep Learning Model for Early Diagnosis of COVID-19, utilizing symptoms as predictors and eliminating the need for laboratory tests. Effective hyperparameter tuning, coupled with 10-

fold cross-validation, was employed to identify the optimal configuration for maximizing algorithm performance. Through this process, the most successful classifier configuration featured hidden layer sizes of (100,), (50, 100, 50), and (50, 50, 50), with relu and tanh activation functions, adam solver, alpha values of 0.05 and 0.0001, and adaptive and constant learning rates.

The results showcase the developed prediction model achieving an accuracy rate of 98.84%, perfect sensitivity (100%), and a specificity score of 97.69%, utilizing the most effective configuration of ANN MLP. This study lays the groundwork for machine learning-based applications that serve as health monitoring and management tools, offering real-time insights into the potential presence of COVID-19 in individuals solely based on symptoms. By alleviating the reliance on laboratory tests, this approach reduces both the effort and expenses associated with medical testing, ultimately contributing to the prevention of potential COVID-19 outbreaks.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357.

Coronavirus. (2024). (World Health Organizations) Retrieved February 16, 2024, from https://www.who.int/health-topics/coronavirus

COVID Live Update. (2024, March 7). (Worldometers Info) Retrieved March 7, 2024, from https://www.worldometers.info/coronavirus/

Itano, F., de Abreu de Sousa, M., & Del-Moral-Hernandez, E. (2018). Extending MLP ANN hyper-parameters Optimization. *International Joint Conference on Neural Networks*. Brazil.

Ul Haq, A., Li, J. P., Memon, M. H., Khan, J., Malik, A., Ahmad, T., Shahid, M. (2019). Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson’s Disease Using Voice Recordings. *IEEE*, 7, 37718 - 37734.

Villavicencio, C. N., Macrohon, J. E., Inbaraj, X., Jeng, J.-H., & Hsieh, J.-G. (2021). COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA. *Algorithms*, 14(7).