المؤتمر الوطني الخامس لكليات الحاسب الآلي بالجامعات السعودية
The Fifth National Conference for Computers Colleges
جامعة أم القرى UMM AL-QURA UNIVERSITY
National Conference Computers Colleges

A New Machine Learning Framework for Detecting COVID-19 From Clinical Data On Lung And Heart Function

Introduction

- Currently, the need for real-time COVID-19 detection methods with minimal tools and cost is an important challenge. The available methods are still difficult to apply, slow, costly, and their accuracy is low.
- In this work, a novel machine learning-based framework to predict COVID-19 is proposed, which is based on rapid inpatient clinical tests of lung and heart function. Compared with current cognition therapy techniques, the proposed framework can significantly improve the accuracy and time performance of COVID-19 diagnosis without any lab or equipment requirements.
- This work adopted five parameters of clinical testing: Respiration rate, Heart rate, systolic blood pressure, diastolic blood pressure, and mean arterial blood pressure. After obtaining results for these tests, a pre-trained intelligent model based on Random Forest Tree (RFT) machine learning algorithm is used for detection.
- This model was trained by about 13,558 records of the COVID- 19 testing dataset collected from King Faisal Specialist Hospital and Research Centre (KFSH&RC) in Saudi Arabia.
- Experiments have shown that the proposed framework performs highly in detecting COVID infections by 96.9%. Its results can be output in minutes, which supports clinical staff in screening COVID-19 patients from their inpatient clinical data.

The State-of-the-art

- The chest CT and X-ray data to propose COVID-19 detection models based on deep learning algorithms (Das et al., 2022; Kong and Cheng, 2022; Paul, 2022; Tan et al., 2020).
- These models are widely adopted and give high accuracy.
- However,
  - These models require CT or CXR images, which are only available from specific centers.
  - It is an expensive and slow process.
  - Obtaining CT or CXR images has increased the crowd which increases the spread of infection among people.
  - There is a shortage of CT and CXR equipment in developing countries and rural areas.
- Audio and Cough data were also used in AI to develop fast detection COVID-19 tools (Alkhodari and Khandoker, 2022; Deshpande et al., 2022; Lella and PJA, 2021).
- The most common limitation in Audio diagnostics is a small size and poor quality of voice training data.
  - The collection of audio data needs high effort and is expensive.
  - High-quality data needs to reduce background noise, and this data is based on the quality of the microphone and the difference between country tones.
- A limited number of machine learning models are based on clinical tests without using imaging data
  - Yang et al., (2020) proposed a ML model based on a gradient-boosting decision tree (GBDT) with 30 laboratory test data to predict infection status. This model result was 83.8%.
  - Wu et al., (2020) used 49 clinical blood test data with the Random Forest (RF) algorithm, the accuracy was 91.7%.
  - Cabitza et al., (2021) based on a gradient enhancement decision tree (GBDT) with 30 laboratory test data to predict infection status, its accuracy scores range between 88%-93% depending on the algorithm used.
  - Banerjee et al., (2020) used logistic regression with 24 parameters of complete blood count test data, the accuracy scored 95%.
- The previous tests need about 24 hours for test results. These methods used a large number of test parameters, and their accuracy is low.

The Proposed Method

- According to the above problems, a fast, simple, low-cost, and easy-to-use machine learning model for COVID-19 detection has been proposed.
- The proposed model is based on clinical data provided by quick and simple tests that are available in different laboratories and even at home.
- That includes five clinical data on lung and heart function
- (Respiratory rate, Heart rate, systemic blood pressure, collapse blood pressure, and arterial blood pressure)
- These features are used by Random Forest Tree (RFT) to develop the model.
- Relying on simple clinical data is the best option to avoid risks as well as save cost and time.

Data Source

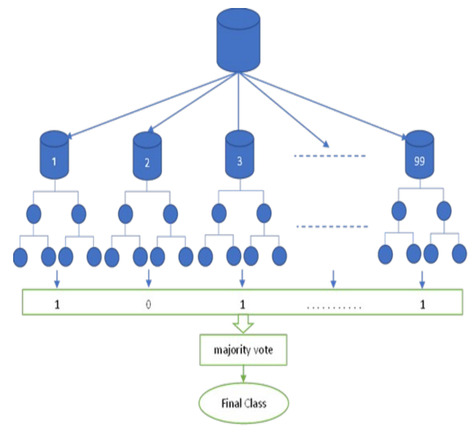
- Clinically conducted information on individuals who underwent Lung function tests and cardiac function tests was collected from King Faisal Specialist Hospital & Research Center (KFSH & RC).
- The primary dataset contains data from 20,511 individuals with 1991 patients infected with COVID which provides approximately 9.8% of the data set and 18,520 (approximately 90.2%) negative COVID tests from Saudi Arabia.

Data Preparing

- The data pre-processing stage aims to prepare the data for use in the prediction model.
- The data is prepared to be suitable for the adopted AI model.
- All missing data records are removed, 14,028 records containing complete data (12,725 (91.8%) negative and 1,303 (9.2%) positive).
- To improve the unbalanced data, under-sampling and over-sampling methods are applied, which are one of the most popular techniques to solve this case(Barandela et al., 2004; Junsomboon and Phienthrakul, 2017).
- Under-sampling is applied to negative class data by removing duplicate and extremely close records.
- Oversampling is applied to the positive class data by iterating each record with five very close values.
- The prepared data became 20,543 records divided into 12,725 (61.9%) negative and 7,818 (38.1%) positive test results.

Classification Model

- Random Forest Tree (RFT) is assumed a new supervised machine learning algorithm based on evaluating multiple decisions to produce the final decision (Breiman, 2001).
- RFT is one of the most accurate classification algorithms and works efficiently on large databases.
- This work consists of 99 sub-tree based on 99 subsets of the training dataset.
- The training dataset provides 66% of the prepared dataset (about 13,558 records are used for training).
- The results of all decision trees are evaluated by majority voting to determine the best overall result.



Experiments and Results

The method	*Accuracy*
SVC	63.5%
Bayes network	60.2%
Multilayer Network	62.1%
Quadratic Discriminant Analysis	60.1%
RFT	96.9%

	Positive	Negative
Positive	2700 (99.3%)	18 (0.7%)
Negative	224 (5.3%)	4042 (94.7%)

Accuracy	96.6%
Precision	92.33%
Recall	99.33%

	Processing time	Procedure	Equipment	Accuracy	Trained staff
Proposed	Fast(mints)	Simple	Simple	Excellent	Not required
Chest image	Slow (hours)	Complex	Complex	Excellent	Required
PCR	Slow (hours)	Average	Average	Very good	Required
Clinical tests	Slow (hours)	Complex	Complex	Very good	Required
Audio tests	Fast(mints)	Simple	Simple	Accepted	Not required

Conclusion

- In this work, a new machine learning-based framework for predicting COVID-19 based on inpatient clinical testing and a random forest tree (RFT) algorithm is proposed.
- The required tests are respiratory rate, heart rate, systolic blood pressure, diastolic blood pressure, and arterial blood pressure.
- A dataset of the COVID-19 tests collected from King Faisal Specialist Hospital & Research Center (KFSH & RC) in Saudi Arabia was used for model training.
- The results of the experiment showed that the proposed framework performs significantly in the detection of Covid infection by 96.9%.
- It can support clinical staff in screening COVID-19 patients from inpatient clinical data.

- Ahmad A. Alzahrani: College of Computers and Information Systems, Umm-AlQura University, aalzahrani@uqu.edu.sa
- Bilal Bataineh: College of Computers and Information Systems, Umm Al-Qura University, bmbataineh@uqu.edu.sa
- Hamoud H. Alshammari: College of Computer Sciences, Jouf University, hhalshammari@ju.edu.sa
- Saeed Kabrah: Faculty of Applied Medical Sciences, Umm Al-Qura University, smkabrah@uqu.edu.sa
- Mohammed Baghdadi: King Faisal Specialist Hospital and Research Centre (KFSH&RC), mbaghdadi@kfshrc.edu.sa