Ministry of Higher Education and
Scientific Research
University of Mosul
College of Computer Science and
Mathematics
Department of Computer Science

Classification for diagnosing COVID-19 based on Machine Learning Techniques

**A Thesis Submitted to the Council of the College of
Computer Science and Mathematics
University of Mosul
as a Partial Fulfillment of Requirements
for the Degree of Doctor of Philosophy in
Computer Science**

By

Ashraf Abdulmunim Abdulmajeed Althanoon

Supervised by

Asst. prof Nada Nimat Saleem

2023 A.D. **1444 A.H.**

Abstract

The COVID-19 virus is causing a massive outbreak in over 150 countries throughout the world, endangering the health and lives of countless people. The ability to identify infected persons early and place them in special care is a vital step in eliminating COVID-19. The application of intelligent technique using machine learning and deep learning to predict COVID -19 in patients would reduce the time needed to wait for medical test results and allow doctors to provide patients with the appropriate therapy. One of the methods used by medical personnel to diagnose Covid-19 patients is the use of CT-Scan and X-rays to detect the disease.

The proposed detection and classification approach for COVID-19 is covered in this thesis and is based on machine and deep learning techniques. Seven different intelligent models were developed and created, these approaches are: machine learning models (HOG-SVM), deep learning models (New Design for CNN), feature extraction models (Xception with SVM, KNN, and DNN), and hybrid deep learning models (CNN with LSTM). Finally, a hybrid created approach that combined the Dolphin Swarm with SVM and Xception was used. Each proposed model was built, analyzed. The results were presented as an explanation of the model's effectiveness tested to differentiate COVID-19 images based on X-ray images and CT scans. Each of these models has been tested in multiple proportions. The training data was divided into five different proportions, from 50% to 10%, to test the data. On each of these models, multiple proportions were evaluated. Five surrogate ratios, ranging from 50% to 10%, were used to stratify the training data.

In this study, dataset of chest X-rays and CT scans were build and obtained from a number of hospitals in Mosul in collaboration with the Department of Radiology and with the hospitals' health authorities' permission. Lang images were obtained for three groups of patients: those with proven COVID-19 infection, pneumonia, and normal incidence (normal lung). The dataset was 8112 (4482 for CT pictures of the lungs and 3630 for X-ray images of the lungs).

From the experimental, the DS-SVM model outperformed all other models in terms of CT scan data accuracy achieving 99.5%. Nevertheless, it fell short of the maximum accuracy in X-ray data (98.4%). While the CNN-LSTM model achieved an accuracy close to that of the DS-SVM model. Moreover, the accuracy was (96.3%) for X-rays and (96.8%) for CT scans

in the CNN model. Whereas, it was (96.7%) for CT scan data in the HOG-SVM model, with a 30% testing rate for test data. Compared to the X-Ray of the HOG-SVM model (95.6%). The XCP-KNN model had lesser accuracy, scoring 90,3% on x-ray data and 93.9% on CT scan data. In the XCP-DNN model, accuracy was (97.3%) for CT scan data, with a 20% testing rate for test data. compared to the X-Ray of the XCP-DNN model (96.8%). In CT data, the XCP-SVM and CNN-LSTM results are quite close, however, the XCP-SVM is less accurate in x-ray data (96.4%).

In conclusion, the researcher concluded that the DS-SVM model had excellent results compared to the other models proposed in the thesis in terms of accuracy. In second place, it was also found that the CNN-LSTM model achieved good results. The researcher recommends using the DS-SVM and CNN-LSTM models to diagnose COVID-19.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات
قسم علوم الحاسوب

تصنيف لتشخيص كوفيد-١٩ بالاعتماد على تقنيات التعلم الالي

اطروحة مقدمة
الى مجلس كلية علوم الحاسوب والرياضيات في جامعة الموصل
كجزء من متطلبات نيل شهادة دكتوراه فلسفة في
علوم الحاسوب

من قبل

أشرف عبد المنعم عبد المجيد الذنون

بإشراف
أ.م.د ندى نعمت سليم

١٤٤٤ هـ ٢٠٢٣ م

الخلاصة

يتسبب فيروس COVID-19 في انتشار واسع النطاق في أكثر من ١٥٠ دولة في جميع أنحاء العالم، مما يعرض صحة وحياة عدد لا يحصى من الناس للخطر. تعد القدرة على التعرف على الأشخاص المصابين مبكرًا ووضعهم في رعاية خاصة خطوة حيوية في القضاء على COVID-19. يساعد استخدام التقنيات الذكية باستخدام التعلم الآلي والتعلم العميق على التنبؤ بـ COVID-19 وكشفه لدى المرضى المصابين بشكل أسرع، حيث يوفر الوقت اللازم لانتظار نتائج الاختبارات الطبية والسماح للأطباء بتزويد المرضى بالعلاج المناسب بعد اكتشاف المرض. تتمثل إحدى الأساليب المتبعة والتي يستخدمها الطاقم الطبي لتشخيص مرضى COVID-19 في الاعتماد على صور الأشعة المقطعية والأشعة السينية للكشف عن المرض.

خلال هذه الاطروحة تم تقديم نهج للكشف والتصنيف المقترح لـ COVID-19 ويستند إلى تقنيات التعلم الآلي والعميق. حيث تم تطوير وإنشاء سبعة نماذج ذكية مختلفة، وهذه الأساليب المقدمة هي: استخدام نماذج التعلم الآلي (HOG-SVM)، ونماذج التعلم العميق (تصميم جديد لـ CNN) ، بالإضافة الى نماذج استخراج طرق التعلم العميق ودمجها مع التعلم الآلي (Xception مع SVM ، و KNN ، و DNN) ، والتعلم العميق المختلط موديلات (CNN مع LSTM). أخيرًا، تم استخدام نهج حديث هجين يجمع بين Dolphin Swarm وSVM وXception. تم بناء كل نموذج مقترح لكل من هذه الطرق وتحليله وعرضت النتائج كشرح لفعالية النموذج واختباره للتمييز بين صور COVID-19 استنادًا إلى صور الأشعة السينية والمسح المقطعي المحوسب. تم اختبار كل نموذج من هذه النماذج بنسب متعددة. تم تقسيم بيانات التدريب إلى خمسة نسب مختلفة، من ٥٠٪ إلى ١٠٪، لاختبار البيانات. في كل من هذه النماذج، تم تقييم النسب المتعددة. تم استخدام خمس نسب بديلة، تتراوح من ٥٠٪ إلى ١٠٪، لتقسيم بيانات التدريب إلى طبقات.

في هذه الدراسة تم الحصول على صور اشعة للصدر وصور الاشعة مقطعية من عدد من المستشفيات في مدينة الموصل بالتعاون مع قسم الاشعة وبموافقة الجهات الصحية بالمستشفيات. تم الحصول على صور لثلاث مجموعات من المرضى: أولئك الذين ثبتت إصابتهم بفيروس COVID-19 والالتهاب الرئوي وحالات عدم الاصابة (الرئة الطبيعية)، وكانت مجموعة البيانات ٨١١٢ (٤٤٨٢ للصور المقطعية للرئتين و ٣٦٣٠ لصور الأشعة السينية للرئتين).

من خلال ملاحظة النماذج المقدمة في هذه الاطروحة، تفوق نموذج DS-SVM على جميع النماذج الأخرى من حيث دقة الكشف عن المرض وتصنيفه بالاعتماد على بيانات الأشعة المقطعية، محققًا ٩٩.٥٪. بينما كانت الدقة القصوى في بيانات الأشعة السينية (٩٨.٤٪). بينما حقق نموذج CNN-LSTM دقة قريبة من دقة نموذج DS-SVM. علاوة على ذلك، كانت الدقة (٩٦.٣٪) للأشعة السينية و (٩٧.٥٪) للأشعة المقطعية في نموذج CNN، بينما كانت (٩٦.٧٪) لبيانات الفحص المعتمدة على الأشعة المقطعية في نموذج HOG-SVM ، بمعدل اختبار ٣٠٪. لبيانات الاختبار. مقارنة بالأشعة السينية لطراز HOG-SVM (٩٥.٦٪). كان نموذج XCP-KNN أقل دقة ، حيث سجل ٩٠.٣٪ باستخدام بيانات الأشعة السينية و ٩٣.٩٪ على بيانات التصوير المقطعي المحوسب. في نموذج XCP-DNN ، كانت الدقة (٩٧.٣٪) لبيانات الفحص الخاصة بالأشعة المقطعية ، مع معدل اختبار ٢٠٪ لبيانات الاختبار. مقارنة بالأشعة السينية لنموذج XCP-DNN (٩٦.٨٪). في بيانات CT، تكون نتائج XCP-SVM و CNN-LSTM قريبة جدًا ، ومع ذلك، فإن XCP-SVM أقل دقة في بيانات الأشعة السينية (٩٦.٤٪).

وفي الختام، استنتج الباحث أن نتائج استخدام نموذج DS-SVM كانت ممتازة مقارنة بالنماذج الأخرى المقترحة من حيث الدقة. ينصح الباحث أيضًا باستخدام نماذج DS-SVM وCNN-LSTM بالإضافة نموذج CNN لتشخيص مرض COVID-19.