Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Fast COVID-19 versus H1N1 screening using Optimized Parallel Inception

Alireza Tavakolian a, Farshid Hajati b, Alireza Rezaee a,*, Amirhossein Olihai Fasakhodi a, Shahatat Uddin c

a Department of Mechatronics Engineering, Faculty of New Sciences and Technologies, University of Tehran, Tehran 1439957131, Iran
b College of Engineering and Science, Victoria University Sydney, Sydney, NSW 2000, Australia
*c School of Project Management, Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia

A B S T R A C T

COVID-19 and swine-origin influenza A (H1N1) are both pandemics that sparked significant concern worldwide. Since these two diseases have common symptoms, a fast COVID-19 versus H1N1 screening helps better manage patients at healthcare facilities. We present a novel deep model, called Optimized Parallel Inception, for fast screening of COVID-19 and H1N1 patients. We also present a Semi-supervised Generative Adversarial Network (SGAN) to address the problem related to the smaller size of the COVID-19 and H1N1 research data. To evaluate the proposed models, we have merged two separate COVID-19 and H1N1 data from different sources to build a new dataset. The created dataset includes 4,383 positive COVID-19 cases, 989 positive H1N1 cases, and 1,059 negative cases. We applied SGAN on this dataset to remove issues related to unequal class densities. The experimental results show that the proposed model’s screening accuracy is 99.2% and 99.6% for COVID-19 and H1N1, respectively. According to our analysis, the most significant symptoms and underlying chronic diseases for COVID-19 versus H1N1 screening are dry cough, breathing problems, diabetes, and gastrointestinal.

1. Introduction

According to recent regulations by the international virus taxonomy committee, coronaviruses are non-segmented positive-sense Ribonucleic acid (RNA) viruses that belong to the family of Coronaviridae, the order Nidovirales, and the genus Coronavirus (Zhou, Yang, Huang, Jiang, & Du, 2019). One of the variants of coronaviruses is COVID-19. There are more than hundreds of COVID-19 strains which cause illness in animals. In a process called ‘spillover event’, COVID-19 has jumped from unknown animal sources to human. The case-fatality rate of COVID-19 varies from 10.8% in European countries like Italy at their first pandemic wave to 0.7% in Germany. But, the global fatality rate during the pandemic was from 0.3% to 3% (Vaillant et al., 2009). For diagnosing H1N1 with flu-like symptoms, routine investigations such as haematological, microbiological, biochemical, and radiologic tests are performed. Due to the body immune system reaction, common symptoms of H1N1 include high fever, coryza, and myalgia. In severe cases, viral pneumonia, superimposed bacterial pneumonia, and hemorrhagic bronchitis have been reported (Jilani, Jamil, & Siddiqui, 2020). Other symptoms of the H1N1 virus are similar to common seasonal flu-like symptoms such as sore throat, fatigue, running nose, cough, and headache. The most common COVID-19’s symptoms are cough, weakness, myalgia, fever, headache, impaired sense of smell, impaired sense of taste, sore throat, runny nose, and nasal congestion (Vaillant et al., 2009). The common symptoms of both COVID-19 and H1N1...
are similar, which makes their screening task challenging. Also, the peak season of these viruses may overlap (Foust, Winant, Chu, Das, Phillips, & Lee, 2020). So, a fast screening of patients regarding these two viruses using an invasive procedure can help healthcare systems respond better.

This research proposes a novel deep model, called Optimized Parallel Inception (OPI), for fast screening of COVID-19 and H1N1 patients. We evaluate the proposed model by measuring accuracy, precision, recall, F1-score, and Area Under Receiver Operating Characteristic (AUROC). We have built a dataset by merging two publicly available data of COVID-19 and H1N1 to conduct the experiments. Also, we identify the possible chronic disease predictors for COVID-19 and H1N1 using the experiments’ outcome. Finally, to address the lack of data, a Semi-supervised Generative Adversarial Network (SGAN) is deployed on 400 samples (less than 10% of the built dataset). The result shows the effectiveness of the SGAN for screening H1N1 and COVID-19 patients even with a small training dataset.

1.1. Related work

Machine learning algorithms have been dominant tools for rapid and accurate diagnosis in most diseases (Florini, Hajati, Barla, & Girosi, 2019; Rafiei, Rezaee, Hajati, Gheisari, & Golzan, 2021). Machine learning algorithms can be sub categorized into supervised, semi-supervised, and unsupervised. (Esen, Inalli, Sengür, & Esen, 2008a; Esen, Inalli, Sengür, & Esen, 2008b, 2008c). If the labels of training samples are known, the task will be categorized as supervised learning. (Esen, Esen, & Oızsoyak, 2017, Esen, Ozgen, Esen, & Sengür, 2009; Esen et al., 2009). Since we have the training labels of the COVID-19 patients in the screening task, we can consider it as the supervised learning. Many studies were conducted to diagnose COVID-19 patients and extract the most critical features in predicting the COVID-19 virus behaviour to reduce infection. Khanday et al. (2020) used clinical text data to classify COVID-19 patients using classical and ensemble machine learning methods. Their models grouped 212 patients into five classes: COVID-19, Acute Respiratory Distress Syndrome (ARDS), COVID-19 and ARDS, Severe Acute Respiratory Syndrome (SARS), and wholesome patients. Using the Term Frequency–Inverse Document Frequency (TF-IDF) technique, they first extracted the most correlated features. Then, they used multinomial naive Bayesian (Corbett-Davies & Goel, 2018) and logistic regression to achieve an accuracy of 96.2%. Wang and Wong (2020) used 13,975 chest radiography images to develop a Convolutional Neural Network (CNN) model to diagnose COVID-19 in patients. They compared their model with both Residual Networks-50 (ResNet50) and Visual Geometry Group-16 (VGG16) (Gikunda & Jouandeau, 2019). Their proposed model could classify patient having an accuracy of 93.3%. Yan et al. (2020) developed five machine learning algorithms, including logistic regression, support vector machine, gradient boosted decision tree, k-nearest neighbour, and neural network for prediction of critical COVID-19 using immune-inflammatory features at admission in Tongji Hospital, Wuhan. They studied the electronic records of 2,799 patients and tested the models on 29 patients. Finally, they extracted three significant features to distinguish critical patients.

Jiang et al. (2020) proposed a data-driven Artificial Intelligence (AI)-based algorithm to identify high-risk patients, those with ARDS. Fever and cough were the most common symptoms in their study population (53 individuals in total). They have concluded that elevated haemoglobin, body aches, and alanine aminotransferase (liver enzyme) are the most predictive features to recognize those prone to ARDS. Their model’s accuracy varied from 70% to 80% using decision trees, random forests, and support vector machine algorithms. Zoabi and Shomron (2020) used a gradient-boosting algorithm to build a model based on a decision tree for diagnosing COVID-19 patients. They have trained the model with 51,831 tested individuals, including 3,624 positive cases. Their test set consists of 47,401 samples, including 3,624 positive cases. They have used three features (gender, age, and close contact with a positive COVID-19 case) and five clinical symptoms (fever, cough, shortness of breath, headache, and sore throat). Their model is a binary classifier that predicts if the tested person infected by COVID-19 or not. They reported an AUROC of 0.90.

Batista et al. (2020) developed a binary classification model with various classic machine learning algorithms to diagnose COVID-19 emergency care patients. In this study, models were trained on 235 adults, including 101 positive COVID-19 cases. The support vector machine obtained the best result algorithm with an accuracy of 85%, the AUROC greater than 0.84, and the sensitivity of 0.68. They also extracted the most predictive features, which were lymphocytes, leucocytes, and eosinophils. Shen et al. (2020) tried to classify pneumonia and COVID-19 by comparative analysis on patient data and distinguishing features of each disease group. However, due to the similarity of patients’ clinical characteristics, they tried to find a suitable discriminator. They found that increased monocyte percentage, C-reactive protein, and decreased eosinophil were more common in COVID-19 patients compared to H1N1 patient. However, they were not very robust features to make a reliable classification. Finally, they reported that computed tomography (CT) scanning with nucleic acid detection is an effective and accurate method for detecting COVID-19. Despite of using medical imaging for detecting COVID-19 patients, using various machine learning models will help to investigate more aspects of the problem. Using of these models will improve the result (Esen, Inalli, Sengür, & Esen, 2008d; Esen, Inalli, Sengür, & Esen, 2008b, 2008c). Most developed models for diagnosing the COVID-19 virus have used a binary classifier to detect positive and negative cases. All the data-driven models have been developed on classical machine learning algorithms with limited number of parameters which give a better performance on small datasets. (Belkin, Hsu, Ma, & Mandal, 2019). Due to the COVID-19 pandemic and the limitations of medical services, CT imaging is not available to all patients (Shah, Keniya, Shridharani, Punjab, Shah, & Mehendale, 2021). In addition, infection in the early stages of the COVID-19 is not evident in the CT images. To address these limitations, we proposed a deep learning method for fast screening COVID-19 versus H1N1 using clinical symptoms that provide a rapid, accurate result. Since the proposed deep learning models find complex and nonlinear patterns, they can discriminate between COVID-19 and H1N1 with similar symptoms. Also, deep learning models can generalize to larger datasets.

2. Data

Currently, there is no publicly available dataset including both COVID-19 and H1N1 cases to evaluate the screening models. To address this limitation, we have merged two sets of publicly available data on COVID-19 and H1N1. This dataset can be used in COVID-19 and H1N1 screening researches.

2.1. COVID-19 data

For COVID-19, we use the COVID-19 symptom checker data (Bilal & Hungund, 2020). The cleaned data contains of 5,435 patients including 4,383 positive cases. This data contains information about symptoms of COVID-19 such as ‘Breathing Problem’, ‘Fever’, ‘Dry Cough’, ‘Headache’, ‘Sore Throat’, ‘Running Nose’, and ‘Fatigue’. Also, the dataset contains the history of patients’ chronic diseases such as ‘Asthma’, ‘Chronic Lung Disease’, ‘Heart Disease’, ‘Diabetes’, ‘Hypertension’, and ‘Gastrointestinal’. Moreover, patients’ behavioural information such as ‘Abroad Travel’, ‘Contact with COVID-19 Patients’, ‘Attended Large Gathering’, ‘Visited Public Exposed Places’, and ‘Family Working in Public Exposed Places’ are recorded in this dataset. The attributes’ values have been recorded as either ‘Yes’ or ‘No’ (see Fig. 1).

The percentage of each characteristic’s value in the whole dataset (both positive and negative cases) is listed in Table 1. The table shows
that 'Dry Cough' and 'Fever' are the most frequent symptoms in the dataset. We also, have shown the percentage of the chronic diseases for the positive cases only. Among patients with positive COVID-19 test, the hypertension is the most frequent chronic condition.

### 2.2. H1N1 data

The H1N1 data were obtained from the NIAID Influenza Research Database (IRD) (Zhang, et al., 2017). The dataset contains 996 patients including 989 positive H1N1 cases. The dataset includes attributes such as 'Collector Institution', 'Host Identifier', 'Collection Year', 'Country', 'Symptoms', 'Subject Age', and 'Temperature'. Information about age, gender, year of data collection is illustrated in Fig. 2. One of the major aspects of the data which is shown in Fig. 2(a) is the average age of infected patients which is 21 and 11 years for females and males, respectively. This information confirms that the 2009 H1N1 pandemic mostly affected younger individuals. Also, Fig. 2(b) shows that the first wave of H1N1 occurred from 2008 to 2009, while the second wave was from 2013 to 2015.

The dataset’s attributes that are underlying diseases and symptoms of H1N1 are shown in Table 2. In H1N1, the most visible symptoms comprise of ‘Fever’, ‘Chills’, ‘Cough’, ‘Sore Throat’, ‘Congested Eyes’, ‘Myalgia’, ‘Shortness of Breath’, ‘Weight Loss’, ‘Sneezing’, ‘Headache’, ‘Runny Nose’, ‘Coughing’, ‘Dizziness’, ‘Abdominal Pain’, ‘Decreased Appetite’, and ‘Fatigue’ (Vaillant et al., 2009). Also, this data contains the percentage of chronic diseases in the study population. Information about the chronic disease has shown in Fig. 3. The figure shows that the asthma and H1N1 virus have a high correlation. Also, 37% of H1N1 patients have asthma.

![Fig. 1. Percentage of each chronic disease in the individuals with COVID-19.](image1)

| Fig. 1. Percentage of each chronic disease in the individuals with COVID-19. |
| --- |
| Gastrointestinal | 47% |
| Hypertension | 52% |
| Diabetes | 49% |
| Heart Disease | 47% |
| Chronic Lung Disease | 46% |
| Asthma | 48% |

![Fig. 2. Descriptive statistics; (a) Age versus gender, (b) Age versus year of collection.](image2)

![Table 1: The COVID-19 data’s attributes.](image3)

| Attributes                          | Values (%) |
|------------------------------------|------------|
| Dry Cough                          | Yes: 79.3, No: 20.7 |
| Breathing Problem                  | Yes: 66.6, No: 33.4 |
| Fatigue                            | Yes: 51.9, No: 48.1 |
| Headache                           | Yes: 50.3, No: 49.7 |
| Running Nose                       | Yes: 54.3, No: 45.7 |
| Sore Throat                        | Yes: 72.7, No: 27.3 |
| Fever                              | Yes: 78.6, No: 21.4 |
| Gastrointestinal                   | Yes: 46.9, No: 53.1 |
| Asthma                             | Yes: 46.3, No: 53.7 |
| Chronic Lung Disease               | Yes: 47.3, No: 52.8 |
| Diabetes                           | Yes: 47.6, No: 52.4 |
| Heart Disease                      | Yes: 46.4, No: 53.6 |
| Hypertension                       | Yes: 49.1, No: 50.9 |
| Abroad travel                      | Yes: 45.1, No: 54.9 |
| Attended Large Gathering           | Yes: 46.2, No: 53.8 |
| Contact with COVID Patient         | Yes: 50.2, No: 49.8 |
| Family Working in Public Exposed Places | Yes: 41.6, No: 58.4 |
| Visited Public Exposed Places      | Yes: 48.1, No: 51.9 |
The idea behind random oversampling is to use oversampling. One of the first and simplest methods is random over sampling methods. Modified SMOTE (MSMOTE) is a modified version of SMOTE which divides samples of the minority class into three groups (safe, border, and latent noise instances) by calculating distances among all samples. The MSMOTE, unlike SMOTE, tries to first indicate noisy samples in the majority class. Then with defining three classes in the minority class, MSMOTE tries to generate new samples for minority class instances that are not classified as latent noise. Thus, result of using MSMOTE for generating new samples in minority class will lead to more similar samples (label wise) in the minority class than SMOTE. So, Using of MSMOTE rather than MSMOTE will increase balance between minority and majority class more robust. The result of using MSMOTE sampling and oversampling on COV-H1N1 dataset has shown in Fig. 5. As can be seen, the use of sampling methods has changed the density of values in each class. Among the applied methods, the MSMOTE with a safe border strategy at the minority class showed the best result. So, in this research, we use MSMOTE for the balancing the COV-H1N1 dataset. After we reach to desirable dataset, we divide the dataset into three sets: train, validation, and test sets.

### 3. Optimized Parallel Inception (OPI)

The inception model is a known deep CNN architecture that has an auxiliary path to increase computational efficiency. Here, we propose an Optimized Parallel Inception (OPI) model to screen the records of the created dataset. The structure of the proposed model is shown in Fig. 6. The proposed model, unlike conventional CNNs which extract information from two-dimensional images, use healthcare recording data. The structure of OPI consist of one main and two auxiliary paths. The first 8 layers of all paths are similar. These layers extract primary information from the input data. Inception layers with kernel sizes of 3 and 5 extract the relationship between co-occurring symptoms and comorbidities, while kernel sizes 7 and 9 focus on extracting relationship of other symptoms and underlying diseases. After these common feature extraction layers, Auxiliary Path 1 tries to classify the instances with the corresponding dense layer. Other two paths have more inception layers in their structure for extracting high-level features. After Inception layers in each path, different structure of fully connected layers has been used for classification. In the main path, 19 different layers including dropout, dense, convolutional, pooling, and inception are used. In the first auxiliary path, we have used a small window for the pooling layer and deployed a dropout layer with a higher probability to avoid overfitting problem. In the second auxiliary path, the opposite anatomy of the first auxiliary path has been developed. Also, we can use different activation functions like the Swish function and hyperbolic tangent function for different auxiliary paths. Experimental results show that the use of the Swish activation function for the shortest path helps OPI to reach more consistent performance. In the proposed model, the output of the main and two auxiliary paths enter a competitive layer. The strategy behind this layer is defined as:

- If the difference between the accuracy of the main and two auxiliary paths in the training phase is equal or more than 0.1%, the output of the path with the maximum accuracy will determine the model’s decision in the testing phase.
- If the difference between the accuracy of the main and two auxiliary paths in the training phase is less than 0.1%, the model’s decision will be determined using the average of the paths’ outputs in the testing phase.

| Attributes         | Values (%) | Missing (%) |
|--------------------|------------|-------------|
| Dry Cough          | Mild: 77, Severe: 3, No: 3.5 | 16.5        |
| Breathing Problem  | Mild: 8, No: 91 | 11          |
| Fatigue            | Mild: 78, Severe: 2, No: 11 | 9           |
| Headache           | Mild: 68, Severe: 6, No: 14 | 12          |
| Runny Nose         | Mild: 85, Severe: 0.2, No: 7 | 7.8         |
| Sore throat        | Mild: 86, No: 3 | 11          |
| Fever              | Mild: 79, Severe: 4, No: 9 | 8           |
| Gastrointestinal   | Mild: 98, No: 1 | 1           |
| Chills             | Mild: 88, No: 4 | 8           |
| Chest Pain         | Mild: 6, No: 90 | 4           |

#### 2.3. COV–H1N1 dataset

We have built the COV-H1N1 dataset by merging the COVID-19 and H1N1 data. Most machine learning algorithms only accept numerical data as input, so input data should be transformed into numerical features. First, filling the missing values is the priority. Based on categorical nature of the features such as symptoms, underlying disease and gender, missing values in each attributes are imputed with the most frequent value. In the COV-H1N1 dataset, the most percentage of missing values is 2.5% which belongs to ‘Dry Cough’. Other attributes have less than 2.5% (in total) missing values. Constant attributes are the type of attributes that contain only one single value. Constant attributes provide no useful information for the screening of the record. Therefore, we remove all the constant attributes from the dataset.

After cleaning the dataset, we apply three different encoding procedures. First, we used a label encoder that converts ‘No’ values to ‘0’ and ‘Yes’ values to ‘1’. Then, we apply the One-hot encoding and target encoding to the cleaned COV-H1N1 dataset. The use of one-hot and target encoders has shown promising results in CNNs. The one-hot encoder creates an orthogonal space for the values of attributes. Target encoder converts attributes’ values into numerical values according to the average of attributes’ value. Fig. 4 shows the correlation between the attributes in the COV-H1N1 dataset. Since the correlation between the attributes is within (−0.5,0.5), no feature elimination is required.

The classes of the COV-H1N1 dataset have various densities. Traditional classification algorithms such as K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and decision trees, which perform well in problems with balanced classes, do not necessarily achieve an acceptable performance in imbalanced class problems. One of the solution for reaching to a balanced dataset is to use oversampling. One of the first and simplest over sampling methods is random over sampling methods. The idea behind of the random over sampling is to generate instances in the minority class to reach equality in the class densities. Synthetic Minority Over-sampling Technique (SMOTE) is one of the most popular sampling methods. Many improved oversampling algorithms attempt to retain SMOTE’s advantages and reduce the shortcomings.
Fig. 4. Correlation between attributes of the COV-H1N1 dataset.

Fig. 5. Distribution of classes in the COV-H1N1 dataset; (a) original imbalanced dataset, (b) after applying MSMOTE, (c) after applying oversampling.
Using the above strategy, path or paths with the best individual performances will determine the choice of the model. This strategy is used due to the small difference among the accuracy of the paths.

The OPI uses a modified inception module which is shown in Fig. 7. For better performance, we use the inception layers before deep convolutional layers. We exploit the merit of using one-dimensional CNN with a high kernel size. These layers help us to reduce computational cost and the number of parameters, speeding up the training and improving the generalization. Also, because the healthcare data does not include spatial dimension, these layers help to capture the patterns along the depth dimension. Lastly, each pair of filters \((1 \times 1, 7 \times 7)\) and \([1 \times 1, 9 \times 9]\) acts like a single, powerful convolutional layer, capable of capturing more complex patterns (Gikunda & Jouandeau, 2019). Using the modified inception module, we achieve a deeper network without overfitting and the gradient vanishing would not affect the network.

Feature selection using Particle Swarm Optimization (PSO) is another aim of this research for COVID-19 versus H1N1 screening. The PSO is a population-based optimization technique inspired by the motion of bird flocks and schooling fish (Amoozegar & Minaei-Bidgoli, 2018). In the PSO, the system is initialized with a population of random solutions, and the search for the optimal solution is performed by updating generations. The PSO uses information of each individual and swarm’s search space in order to reach the best global minimum of the objective function (Amoozegar & Minaei-Bidgoli, 2018). One of the reasons for choosing the PSO for the feature selection task is the fast convergence speed.

Assume the location of the \(i\)th particle is \(P_i(t) = (p(i,1), p(i,2), \ldots, p(i, D))\), its velocity is \(V_i(t) = (v(i,1), v(i,2), \ldots, v(i, D))\), the optimal location found by this particle is \(L_i(t) = (l(i,1), l(i,2), \ldots, l(i, D))\), and the optimal location found by the swarm is \(G(t) = (g(t,1), g(t,2), \ldots, g(t, D))\). Then, location and velocity of each particle is updated as below.

\[
V_{ij}(t+1) = w \cdot V_{ij}(t) + r_1 \cdot c_1 \cdot (L_{ij}(t) - P_{ij}(t)) + r_2 \cdot c_2 \cdot (G(t) - P_{ij}(t)) \tag{1}
\]

\[
P_{ij}(t+1) = P_{ij}(t) + V_{ij}(t+1) \tag{2}
\]

where \(i\) is the iteration times, \(c_1\) and \(c_2\) are two acceleration coefficients, \(r_1\) and \(r_2\) are random values between 0 and 1, and \(w\) is an inertia weight of the particle on the fly velocity. The aim is to rank the features based on their significance for COVID-19 and H1N1 screening and select a proper feature set. The PSO algorithm defines the subsets of features randomly based on the optimization policy. For the evaluation of each particle, we use the OPI model. The objective function for evaluation of each particle (a subset of features) is determined by the measured accuracy of the OPI model. To increase the overall performance of the OPI, we use a \(P\) parameter which is defined as an average accuracy obtained for each path. For better discrimination of each particle, an \(E\) parameter which stands for the average mean squared error of three paths is also added to the objective function. So, the objective function, \(F\), is defined as

\[
F = a \cdot (1 - P) + (1 - a) \cdot \left(1 - \frac{N_s}{N_f}\right) + (1 - a) \cdot E \tag{3}
\]

when \(N_s\) is the number of selected features, \(N_f\) is the total number of features which is equal to 18 in the COV-H1N1 dataset, and \(a\) is a hyperparameter for specifying the contribution of accuracy and loss in the objective function. We have used a grid search strategy to find the optimum value for \(a\). Chosen range for \(a\) parameters is between 0.01 to 0.99. We have mentioned the result of this experiment in Fig. 8. Based this result, \(a\) has been specified as 0.9. Also, \(P\) and \(E\) are the aggregated accuracy and loss of OPI for each PSO subset. The optimization algorithm is presented below.

4. Generative model

To address the limitation in the available COVID-19 and H1N1 data, we present a semi-supervised data generator model based on
Fig. 7. The modified inception module.

Fig. 8. Relation between alpha values and ; (a) Cost function, (b) Average Accuracy.

Algorithm 1 Optimization Algorithm

\begin{verbatim}
1. The set of randomly selected features
2. The total number of features
3. Set the maximum iteration number

∀: Position;

\begin{align*}
W & \leftarrow \text{Objective Function}\\
L_i(t) & \leftarrow \text{in the swarm};
\end{align*}

while \( \text{iter} < \text{Maxiter} \) do

\text{for } i = 1: \text{Number of particle} do

Update \( V_i \);

\begin{align*}
V_{i,j}(t + 1) & = w \cdot V_{i,j}(t) + r_1 \cdot c_1 \cdot (L_{i,j}(t) - P_{i,j}(t)) + r_2 \cdot c_2 \cdot (G_{i,j}(t) - P_{i,j}(t))
\end{align*}

end

Update \( P_{i,j} \);

\begin{align*}
P_{i,j}(t + 1) & = P_{i,j}(t) + V_{i,j}(t + 1)
\end{align*}

end

end

end

\end{verbatim}

Table 3 shows the architecture of the proposed SGAN. The generator is a multi-layer perceptron that takes a noise vector and converts it to a one-dimensional vector in which length is equal to the number of features in the dataset. The discriminator is composed of a fully connected layer. Here, we use Leaky Rectified Linear Unit (Leaky ReLU) as the activation function. Using this activation function, the effect of the negative side of input layers also are considered for prediction. Also, we use batch normalization as a trainable layer which decreases any unwanted interdependence between parameters across layers and speeds up the training process and increases the robustness (Pan, et al., 2019). To avoid over-fitting, a dropout layer is also used.

5. Experimental results

For better evaluation of the proposed model, we use k-fold cross-validation. Here, we consider \( k = 10 \). A single fold acts as a validation set, while the remaining nine folds are used for training. Finally,
the results are averaged to represent a single estimation. We use the Linear Regression (LR), Random Forrest (RF), and Extreme Gradient Boosting (XGBoost) classifiers as the benchmarks (Corbett-Davies & Goel, 2018). For the evaluation, we use accuracy, precision, recall, F1-score, confusion matrix, and AUROC. For training of the model, we use Categorical Cross-Entropy (CCE) as the loss function and Adaptive Moment Estimation (Adam) as the optimizer.

First, we evaluated the proposed model without using an optimization block. The results of COVID-19 and H1N1 screening using different methods are shown in Table 4. The result shows the superiority of the proposed model in the screening of COVID-19 versus H1N1 compared to the benchmarks. Also, the accuracies of the top five folds are shown in Fig. 10. As can be seen, both auxiliary paths helped to increase the performance of main path. The average results of 10 folds cross-validations are converged to an accuracy of 98.88%.

Based on the results in Table 4, the difference between the accuracy and precision of the models is trivial. But based on the measured sensitivity and specificity, the proposed model performs better than others. Since the task in this research is screening of COVID-19 and H1N1 patients, the model’s ability to correctly detect positive patients (Sensitivity) and negative patients (Specificity) is more important.

For better observation of the performed screening task, we show the normalized confusion matrix of the proposed model and the benchmarks in Fig. 11. A key point of the proposed model’s confusion matrix is the high sensitivity (Recall) for both COVID-19 and H1N1. With such a high sensitivity, we may receive false alarms for COVID-19 or H1N1. However, the model can predict the positive COVID-19 and H1N1 cases with high confidence. The OPI perfectly detects the H1N1 and neither COVID-19 nor H1N1 cases, while the random forest algorithms have the best result for COVID-19 detection. For better observation of achieved result in each class, detailed information about precision, sensitivity, specificity and AUROC of each class has been shown in Table 5.

Based on Table 5 detection of H1N1 in patients with proposed OPI is more accurate than COVID-19. Result of achieved specificity in class “Neither COVID-19 Nor H1N1” and class “COVID-19” shows that the detection rate for true negative case in “COVID-19” class is higher than the class of “Neither COVID-19 Nor H1N1”. Although, based on achieved sensitivity, the detection rate for true positive case in “Neither COVID-19 Nor H1N1” is much higher than the class of “COVID-19”.

5.1. Optimization’s results

For achieving an acceptable performance using a limited number of features, we used a PSO-based optimization. We measured the performance of the proposed model by applying the optimization technique. With the increase in the number of features, the accuracy of the classifier increases and the loss of the classifier decreases. However, we can achieve an acceptable performance even with a limited number of features. For this purpose, we have conducted an experiment measuring the performance of the model using various subsets of the features.
Table 5  
Results of screening using the OPI for each class.  

| Class name                        | Precision | Sensitivity | Specificity | AUROC  |
|-----------------------------------|-----------|-------------|-------------|--------|
| Neither COVID-19 Nor H1N1         | 100%      | 100%        | 95.46%      | 0.998  |
| H1N1                              | 100%      | 100%        | 99.71%      | 0.998  |
| COVID-19                          | 96.48%    | 96.82%      | 100%        | 0.991  |

Fig. 12 shows the best results achieved in this experiment. Using the best seven subsets of features, we have calculated a significance score for each feature. Here, the significance score of a feature is the number of times the feature has been used in the best subsets. As it has shown in Fig. 12, one of the most significant feature to detect COVID-19 is the contact with other COVID-19 positive cases. ‘Dry Cough’ and ‘Sore Throat’ are also significant symptoms for the screening of COVID-19 versus H1N1. In addition, ‘Diabetes’ and ‘Gastrointestinal’ are the significant chronic diseases that can help the proposed model for COVID-19 versus H1N1 screening. However, using ‘Dry Cough’ and ‘Breathing Problems’ shows more promising results for reaching the highest accuracy.

5.2. Generative model’s results

Due to the lack of proper data for COVID-19 versus H1N1 screening in the healthcare systems, in this research, we have proposed a semi-supervised GAN (SGAN) model to tackle the issue. To evaluate the model, we randomly select less than 10% of the dataset (400 samples) and see if the model can accomplish the screening task appropriately. We have set a threshold for the accuracy, which is 99.2%, to stop the training. This threshold helps us to solve the problem of convergence which usually happens during GANs training procedure. After 3,396 iterations, the model’s accuracy reached the threshold and the training procedure finished. Using the proposed SGAN model, an accuracy of 99.7% achieved with only 400 samples. However, without using SGANs the accuracy of the proposed model could not reach 90% on the small dataset.
5.3. Comparison

For better comparison between the proposed OPI with similar work for screening COVID-19, we have compared the result of the proposed model with other research. The summary of this comparison has been shown in Table 6.

The comparison result shows that proposed OPI is superior than other models. Also, cough is the most repeated symptoms for detection of COVID-19. RF with proper hyper parameter tuning structure has shown promising result for COVID-19 detection.

6. Discussion

With the rage of COVID-19 at the end of 2019, detection of COVID-19 cases all around the world has gathered the attention of researchers. H1N1 is a branch of the influenza family that has similar symptoms to COVID-19. Peak prevalence of the H1N1 virus has been observed from
October to April. In this research, we proposed the Optimized Parallel Inception (OPI) model to screen COVID-19 versus H1N1. To evaluate the proposed model, we have built a dataset by merging two publicly available COVID-19 and H1N1 data. We proposed a procedure that processes the raw dataset in four steps: cleaning data, preprocessing, encoding, and balancing. The proposed model shows an accuracy of 98.88% for the screening task. Unlike existing procedures, we proposed a non-invasive screening method using symptoms, history of underlying disease, and social behaviour of each patient. The proposed procedure does not impose a cost on healthcare systems, decrease contact between positive cases and medical staff, and has no side effect.

Further investigation of related symptoms and each virus was conducted. For COVID-19, the most related symptoms to positive COVID-19 are ‘dry cough’ and ‘breathing problem’. The most related underlying disease to positive COVID-19 is ‘hypertension’ and ‘heart disease’. For the H1N1 virus, most related symptoms consist of ‘sore throat’, ‘fever’, and ‘myalgia’. The most related underlying disease to positive H1N1 are ‘asthma’ and ‘diabetes’. When both datasets are combined into COV-H1N1 dataset, the experiments shown that ‘Diabetes’ and ‘Gastrointestinal’ are the most significant chronic disease factors for screening COVID-19 patients from H1N1. Also, using ‘dry cough’ and ‘breathing problems’ symptoms have shown promising results. The proposed model is useful to develop an expert system to fast screen patients with precise accuracy and break the sequence chain of coincident COVID-19 and H1N1 waves. For better observation, we used only 400 samples (less than 10% of the dataset) for the screening task using the proposed semi-supervised GAN (SGAN). Even with a lower number of instances, the SGAN successfully achieved 99.7% accuracy. So, we suggest the SGAN model for the case of insufficient H1N1 and COVID-19 samples. Based on Fig. 12, combining OPI with PSO has shown that small subset of features can be used for COVID-19 screening with an accuracy of 98%. So, proposed OPI with PSO and SGAN can work with small subset of features or instances and outperform similar models of COVID-19 and H1N1 screening. Although, based on Table 6, OPI is showing superiority compared to other machine learning models, the novelty of this research is to develop an expert hybrid system (OPI with PSO) to use optimum number of features and reach a better result compare to similar models.

The OPI screens COVID-19 and H1N1 patients using their symptoms, chronic diseases, and social behaviour. In case of asymmetric patients, patients without symptoms, we may still use the model with the chronic diseases and social behaviour only as the input feature. However, a drop in the performance is expected.

7. Conclusion

The fast screening of COVID-19 versus H1N1 is a challenging task that is essential in disease trend monitoring and pandemic management. With the rapid growth in the number of COVID-19 patients, we should equip our healthcare’s systems with expert systems to dealing with this pandemic wisely. In this research, we presented the Optimized Parallel Inception (OPI) model as a high-performing machine
learning model to screen COVID-19 and H1N1 cases. The proposed model is robust to missing values and makes precise predictions in the presence of imbalance and error in the recorded attributes. The proposed model is state-of-the-art for the detection of COVID-19, H1N1, and neither COVID-19 nor H1N1 cases. The unique 99.6% accuracy for separating H1N1 and neither COVID-19 nor H1N1 cases was achieved. Diabetes and gastrointestinal are the most significant chronic disease indicators of COVID-19 and H1N1. Among the symptoms, dry cough and breathing problems have shown the most effective screening results. Also, a semi-supervised GAN (SGAN) model was presented to deal with the problem of insufficient data reached 99.2% accuracy. Compared to existing works for detection of COVID-19, OPI has shown 2.68% improvement in accuracy using electronics healthcare dataset. The proposed models help the healthcare providers in pandemics by rapid screening and decreasing human interactions. With emerging new variants like omicron or any other contagious variant, the future of this research is to train the proposed model using more data. Also, we will explore how to add more diverse symptoms to the proposed screening systems.

CRediT authorship contribution statement

Aliereza Tavakolian: Conceived and designed the study, Data curation, Formal analysis. Farshid Hajati: Conceived and designed the study, Data curation, Formal analysis, Investigation, Supervision. Aliereza Rezaee: Conceived and designed the study, Data curation, Formal analysis, Supervision. Amirhossein Ollaei Fasakhodi: Conceived and designed the study, Data curation, Formal analysis. Shahadat Uddin: Conceived and designed the study, Investigation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

All the authors contributed to the original drafting of the work for important intellectual content and approved the final completed article.

References

Amouzegar, M., & Minaei-Bidgoli, B. (2018). Optimizing multi-objective PSO based feature selection method using a feature elitism mechanism. Expert Systems with Applications, 113, 499–514.
Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32), 15849–15854.
Bilal, H., & Hungund, K. (2020). COVID-19 symptoms checker. kaggle URL: https://www.kaggle.com/iamhungundii/covid19-symptoms-checker.
Canas, L. S., Sudre, C. H., Pujol, J. C., Polidori, L., Murray, B., Molteni, E., et al. (2021). Early detection of COVID-19 in the UK using self-reported symptoms: a large-scale, prospective, epidemiological surveillance study. The Lancet Digital Health, 3(9), e587–e598.
Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv Preprint:1808.00023.
de Moraes Batista, A. F., Miraglia, L. J., Donato, T. H. R., & Chiavegatto Filho, A. D. P. (2020). COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. Cold Spring Harbor Laboratory Press, MedRxiv.
Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008a). Artificial neural networks and adaptive neuro-fuzzy assessments for ground-coupled heat pump system. Energy and Buildings, 40(6), 1074–1083.
Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008b). Forecasting of a ground-coupled heat pump performance using neural networks with statistical data weighting pre-processing. International Journal of Thermal Sciences, 47(4), 446–449.
Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008c). Modeling a ground-coupled heat pump system by a support vector machine. Renewable Energy, 33(8), 1814–1823.
Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008d). Modelling a ground-coupled heat pump system using adaptive neuro-fuzzy inference systems. International Journal of Refrigeration, 31(1), 65–74.
Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008e). Performance prediction of a ground-coupled heat pump system using artificial neural networks. Expert Systems with Applications, 35(4), 1940–1948.
Esen, H., Inalli, M., Sengur, A., & Esen, M. (2008f). Predicting performance of a ground-source heat pump system using fuzzy weighted pre-processing-based ANFIS. Building and Environment, 43(12), 2178–2187.
Esen, H., Orgen, F., Esen, M., & Sengur, A. (2009). Modelling of a new solar air heater through least-squares support vector machines. Expert Systems with Applications, 36(7), 10673–10680.
Feng, W., Huang, W., & Ren, J. (2018). Class imbalance ensemble learning based on the margin theory. Applied Sciences, 8(5), 815.
Fiorini, S., Hajati, F., Barla, A., & Girosi, F. (2019). Predicting diabetes second-line therapy initiation in the Australian population via time span-guided neural attention network. PLoS One, 14(10), 1–17.
Foust, A. M., Winant, A. J., Chu, W. C., Das, K. M., Phillips, G. S., & Lee, E. Y. (2020). Pediatric SARS, H1N1, MERS, EVA1L, and now coronavirus disease (COVID-19) pneumonic: what radiologists need to know. American Journal of Roentgenology, 215(3), 736–744.
García, S., Luengo, J., & Herrera, F. (2015). Dealing with missing values. In Data preprocessing in data mining (pp. 59–105). Springer.
Ghazikhani, A., Yazdi, H. S., & Monsefi, R. (2012). Class imbalance handling using wrapper-based random oversampling. In 20th Iranian conference on electrical engineering (pp. 611–616). IEEE.
Gilkunda, P. K., & Jouandeau, N. (2019). State-of-the-art convolutional neural networks for smart farms: A review. In Intelligent computing-proceedings of the computing conference (pp. 763–775). Springer.
Harshanand, B., & Sangaijah, A. K. (2020). Comprehensive analysis of deep learning methodology in classification of leukocytes and enhancement using swish activation units. Mobile Networks and Applications, 25(6), 2302–2320.
Iwendi, C., Bashir, A. K., Penskar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., et al. (2020). COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in Public Health, 8, 357.
Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., et al. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials & Continua, 63(1), 537–551.
Jilani, T. N., Jamal, R. T., & Siddiqui, A. H. (2020). H1N1 influenza (swine flu). StatPearls Publishing, StatPearls [Internet].
Khaday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Din, M. M. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology, 12(3), 731–739.
Nguyen, J. L., Yang, W., Ito, K., Matte, T. D., Shanman, J., & Kinney, P. L. (2016). Seasonal influenza infections and cardiovascular disease mortality. Jama Cardiology, (3), 274–281.
Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., & Zheng, Y. (2019). Recent progress on generative adversarial networks (GANs): A survey. IEEE Access, 7, 36322–36333.
Patel, M., Dennis, A., Flutter, C., & Khan, Z. (2010). Pandemic (H1N1) 2009 influenza. British Journal of Anaesthesia, 104(2), 128–142.
Punshhit, V., Kudale, A., Sundaram, N., Joseph, S., Schaetti, C., & Weiss, M. G. (2018). Public health policy and experience of the 2009 H1N1 influenza pandemic in Pune, India. International Journal of Health Policy and Management, 7(2), 154.
Rafiei, A., Rezaee, A., Hajati, F., Gheissari, S., & Golzadeh, M. (2021). SSP: Early reporting of seizures using fully connected LSTM-CN model. Computers in Biology and Medicine, 128, Article 104110.
Rodríguez, P., Bautista, M. A., González, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. Image and Vision Computing, 75, 21–31.
Shi, M., Fuad, N., Jamal, N., & Ismail, M. (2020). kNN and SVM classification for EIG: a review. In IEEECE2019 (pp. 555–565). Springer.
Shah, V., Keniya, R., Shridharani, A., Punjabi, M., Shah, J., & Mehendale, N. (2021). Diagnosis of COVID-19 using CT scan images and deep learning techniques. Emergency Radiology, 28(3), 497–505.
Shen, C., Tian, M., Song, X., Zhang, G., Liang, J., Yu, H., et al. (2020). Comparative analysis of early-stage clinical features between COVID-19 and influenza A H1N1 virus pneumonia. Frontiers in Public Health, 8, 206.
Shi, F., Xia, L., Shan, F., Song, B., Wu, D., Wei, Y., et al. (2021). Large-scale screening to distinguish between COVID-19 and community-acquired pneumonia using infection size-aware classification. Physics in Medicine and Biology, 66(6), Article 065031.
Vaillant, L., La Ruche, G., Tarantola, A., Barboza, P., et al. (2009). Epidemiology of fatal cases associated with pandemic H1N1 influenza 2009. *Eurosurveillance*, 14(33), 19309.

Wang, L., & Wong, A. (2020). COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-Ray images. arXiv preprint arXiv:2003.09871.

Yan, L., Zhang, H.-T., Xiao, Y., Wang, M., Guo, Y., Sun, C., et al. (2020). Prediction of criticality in patients with severe Covid-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. Cold Spring Harbor Laboratory Press, MedRxiv.

Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., et al. (2017). Influenza research database: An integrated bioinformatics resource for influenza virus research. *Nucleic Acids Research*, 45(D1), D466–D474.

Zhou, Y., Yang, Y., Huang, J., Jiang, S., & Du, L. (2019). Advances in MERS-CoV vaccines and therapeutics based on the receptor-binding domain. *Viruses*, 11(1), 60.

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, 4(1), 1–5.

Zoabi, Y., & Shomron, N. (2020). COVID-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach. Cold Spring Harbor Laboratory Press, MedRxiv.