# Predicting the Existence of COVID-19 using Machine Learning Based on Laboratory Findings

Hamza Turabieh
*College of computer and Information Technology*
*Taif University, saudi arabia*
h.turabieh@tu.edu.sa

Wahiba Ben Abdessalem Karaa
*High Institute of Management of Tunis*
*Tunis University. Tunisia*

2021 International Conference of Women in Data Science at Taif University (WiDSTaif) | 978-1-6654-4948-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/WIDSTAIF52235.2021.9430233

***Abstract*—Since December 2019, a new coronavirus disease (COVID-19) was detected in Wuhan, China, spread all over the world. Many research papers have been published to study this disease and help humans to overcome this pandemic. Here, we highlighted the prediction process of COVID-19 based on a combination between wrapper feature selected (FS) algorithm and four different classifiers, namely Convolutional Neural Network (CNN), decision trees (C4.5), nearest neighbors (kNN) and, Naïve Bayes (NB). A real dataset has been used in this paper generated by Hospital Israelita Albert Einstein at Sao Paulo, Brazil. The obtained results show an excellent performance of BGA with CNN compared to other methods with accuracy $76\%$**

***Keywords*—COVID-19, Classification, Feature Selection, Machine Learning.**

## I. Introduction

COVID-19 appears firstly in Wuhan, China, then spread in an exponential manner over all the world [1]. COVID-19 is a type of SARS-CoV2 virus and moves from animal to human [2]. COVID-19 has several symptoms (i.e., COVID-19 pneumonia, Acute Respiratory Distress Syndrome (ARDS), and acute respiratory failure) on humans that may cause death [3]. As a result, on March 11, 2020, World Health Organization (WHO) stated that COVID-19 is a global pandemic. Many countries took rigid procedures to stop the spread of COVID-19, such as close borders, schools, universities, and many other sectors.

The main transmitting methods of COVID-19 are breathing and close contact. Figure 1 presents these two main methods and how to control the speed of spread of COVID-19. Moreover, many governments have issued rigid regulations based on social distance and wearing masks in closed environment [4]. In general, these regulations help the governments sector to control the exponential growth of infected people. However, up to date, COVID-19 still present and a huge number of researchers investigate the best approaches to deal with this virus either preparing new vaccines or issuing new regulations.

Detecting the existence of COVID-19 is a challenging task. Machine learning provides excellent methods to detect it [5]. Moreover, Data Mining (DM) can be used to extract meaningful information from big data and perform complex tasks to discover hidden knowledge, especially for medical datasets [4]. In general, there are many methods that can be used to determine the existence of COVID-19, such as Support Vector Machine (SVM), Random Forest (RF), Apriori Algorithm (AA), Logistic Regression (LR), Ensemble Methods (EM), Naive Bayes (NB), k-nearest neighbors (kNN), Artificial Neural Networks (ANN), etc. [6].

In general, data preprocessing play a vital role in enhancing the overall performance of machine learning and DM methods. Improving data quality based on data re-balancing and feature selection will enhance the learning process and avoid overfitting problems [7–9]. An imbalanced data problem occurs when the number of samples for one class is larger than other classes. This will force the machine learning classifier biased toward one class more than the other one. In the dataset used here in this paper, the number of patients who did not have COVID-19 is higher than infected patients. This motivates us to solve this problem based on an oversampling method called Adaptive Synthetic Sampling (ADASYN) [10].

Feature selection (FS) is an important preprocessing set the can reduce the data dimensionality and remove irrelevant/redundant data from the original dataset. FS methods can be classified into two main groups: filter and wrapper methods. In simple, wrapper methods work much better than filter methods when the number of features is large. The main advantage of FS that it can reduce the learning process for machine learning. FS have been used successfully in many fields such as software fault prediction [11, 12], Intrusion detection [13], prediction of student performance [14] and others [15].In this paper, we adopted a wrapper feature selection method based called a binary genetic algorithm (BGA).

Up to date, a huge number of published papers employed machine learning to predict the existence of COVID-19. Many papers detect COVID-19 based on image processing. In contrast, few papers highlighted the detecting process based on blood reports. AlJame et al. [16] proposed two levels of prediction model to predict COVID-19: the first level consists of employing three machine learning classifiers which are: extra trees, random forest and logistic regression, while the second level is employing extreme gradient boosting (XGBoost) in order to enhance the overall performance. The proposed approach shows an excellent performance with accuracy $99.88\%$. Jiang et al. [17] evaluated six machine learning methods to diagnosis the existence of COVID-19. These methods are SVM, RF, KNN, LR LR, and two different decision trees (DT). The authors applied these methods over a real dataset obtained from Wenzhou Central Hospital and Cangnan People's Hospital in Wenzhou. The performance of SVM shows

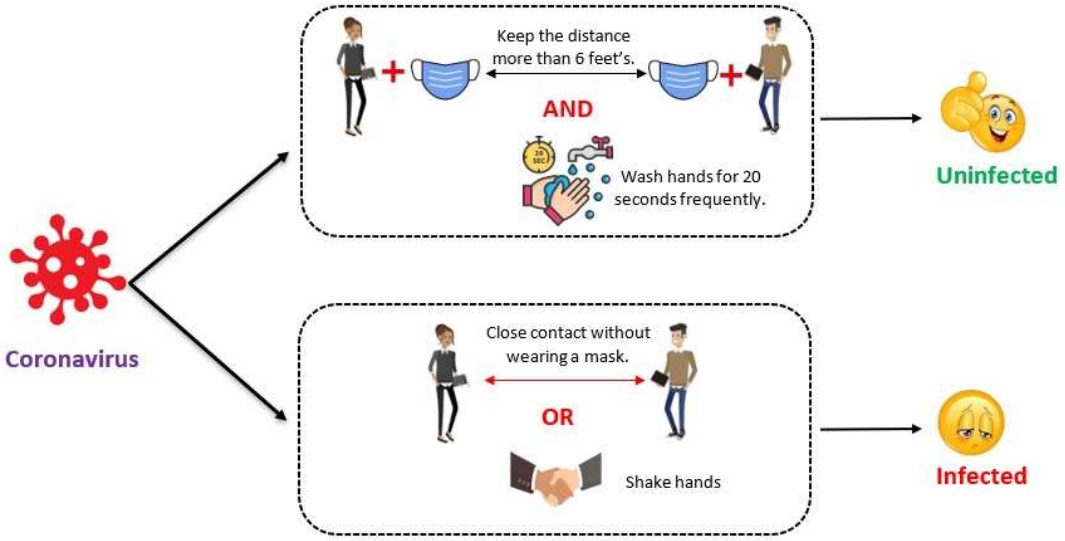


Fig. 1: Main transmission methods of COVID-19.

excellent performance compared to other methods with an accuracy 80%. Batista et al. [18] examine the performance of five machine learning methods, namely SVM, LR, RF, ANN, and gradient boosted trees (GBT). The authors examined the performance of these methods over a real dataset obtained from Hospital Israelita Albert Einstein at Sao Paulo, Brazil. The performance of SVM and RF outperform other methods with an AUC value equal $0.847$. Schwab et al. [19] examined five machine learning methods, namely RF, ANN, LR, SVM, and Gradient Boosting (XFB), to predict COVID-19 using the same dataset used in [18]. The performance of XGB showed an acceptable ratio compared to other methods with 66% AUC value. Barbosa et al. [20] proposed an intelligent method to diagnosis Covid-19 based on blood testing. The authors applied three algorithms to select the most valuable factors to detect Covid-19 based which are: particle swarm optimization, evolutionary algorithms, and manual selection. The obtained results show an excellent performance of detecting Covid-19. Alakus and Turkoglu [21] examined the performance of five deep learning methods (i.e., CNN, Long-Short Term Memory (LSTM), Recurrent Neural Networks (RNN), CNNLSTM, and CNNRNN) and one standard neural network (i.e., ANN) to detect COVID-19 based on laboratory findings. The performance of CNNLSTM with AUC $92.30\%$ outperforms other methods.

This paper is organized as follows: SectionII explores the proposed hybrid approach between BGA and different classification method. Section III highlights the COVID-19 dataset used in this work. Section IV explores the obtained results and their analysis. Finally, Section V presents the final findings and function works.

## II. Proposed Hybrid Approach

In this work, we investigated the performance of a different set of machine learning methods (i.e., kNN, CNN, NB, and C4.5) with wrapper feature selection (i.e., BGA). Figure 2 presents the flow chart of the proposed hybrid approach, where each classifier is trained on the training data, and simulated on testing data after training phase is finished . In this paper, two kinds of experiments were conducted: first, execute the proposed approach without feature selection, where we used the original dataset without any modifications. Second, execute the proposed approach with feature selection, where BGA is employed to reduce the data dimentionality of the original dataset.

### A. Machine learning classifiers

Up to date, there are several machine learning methods that can be employed as classification algorithms. In this work, we limited this research to use four classification methods, namely Convolutional Neural Network (CNN), decision trees (C4.5), Naïve Bayes (NB), and nearest neighbors (kNN). All these methods have been employed successfully in different domains. CNN is a robust machine learning classifier that is able to predict complex data based on deep learning concepts [22]. The kNN algorithm employs the concept of the similarity of a predetermined value to classify the dominant label (i.e., class) to the closest group [23]. In this work, we employed $k = 10$ for kNN method. C4.5 method is a type of decision tree (DT) classifier that uses the input data to build the DTs [24]. A cross-validation method is used to build all the machine learning classifiers with kfold=10. Interested readers about machine learning methods and their applications can read [22, 25–27].

### B. Binary Genetic Algorithm

Genetic Algorithms (GA) is an evolutionary search algorithm that shows an excellent performance in many domains [12, 28, 29]. In simple, GA has three main operations, which

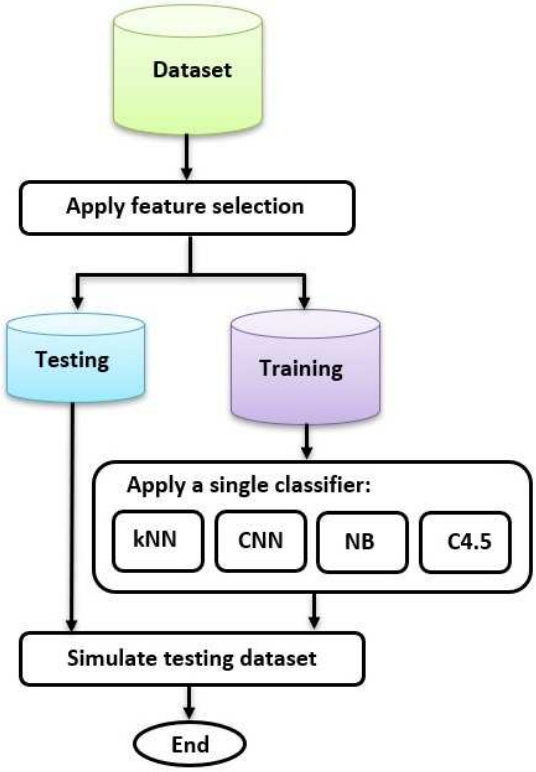


Fig. 2: A proposed methodology.

are selection, crossover, and mutation. These operations are performed on selected solutions (i.e., parents) from the population pool. GA is an iterative algorithm that is repeated until a stop condition is met (i.e., the optimal solution or the maximum number of iterations is reached). Figure 3 presents the standard pseudo-code for GA.

GA begins by generating a random set of solutions (population) based on a predetermined parameters called *population size*; each solution is evaluated based on a fitness function. Two solutions have to be selected based on a selection function. Then crossover operation has to be performed to generate new two solutions (i.e., offsprings), followed by the mutation process. Finally, the current population is updated based on the fitness value of the new solutions.

In this work, We employed GA as a binary search method, where each solution is presented as a binary vector as shown in Figure 4. Where 0 refers to a not selected feature, while 0 refers to a selected feature. Figure 5 presents a single generation for the proposed BGA. Here in this work, we employed ANN as an internal classifier to evaluate the selected solution based on the fitness function in Eq. (1). The variable $E$ refers to the overall error rate obtained from ANN, $\beta$ presents a predetermined fixed value ($\beta$ = 5), $|R|$ refers to the number of the selected features, $|N|$ refers to the total

Given:
-nP: base population size.
-nI: number of iterations.
-rC: rate of crossover.
-rM: rate of mutation.
Generate initial population of size nP.
Evaluate initial population according to the fitness function.
**While** $(current\_iteration \leq nI)$
//Breed $rC \times nP$ new solutions.
Select two parent solutions from current population.
Form offspring's solutions via crossover.
**IF**$(rand(0.0, 1.0) <$ rM)
Mutate the offspring's solutions.
**end IF**
Evaluate each child solution according to the fitness function.
Add offspring's to population.
//population size is now MaxPop=nP× (1+rC).
Remove the rC× nP least-fit solutions from population.
**end While**
*Output the global best solution*

Fig. 3: The original pseudo-code for the GA.

number of original features.

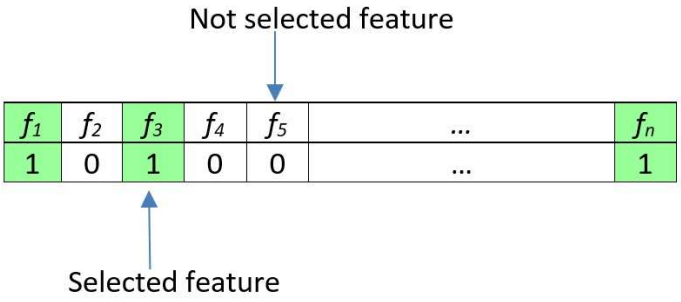


Fig. 4: Solution presentation.

$$Fitness = E * (1 + \beta * \frac{|R|}{|N|}) \quad (1)$$

## III. Dataset

A public dataset that was prepared by Hospital Israelita Albert Einstein at Sao Paulo Brazil is used in this paper [30]. This dataset simulates patient's data to detect SARS-CoV2. The dataset was collected at the beginning of 2020 based on a blood test for 5644 different patients with 111 features. To simplify the dataset, only 18 features have been selected. The dataset has no missing data, no information about patients such as age, gender, and race. Table II presents the selected features used in this paper.

## IV. Experimental Results

In this work, four different machine learning methods have been examined to detect the existence of COVID-19. This problem (i.e., feature selection) is considered a binary classification problem (i.e., Positive and Negative). The proposed

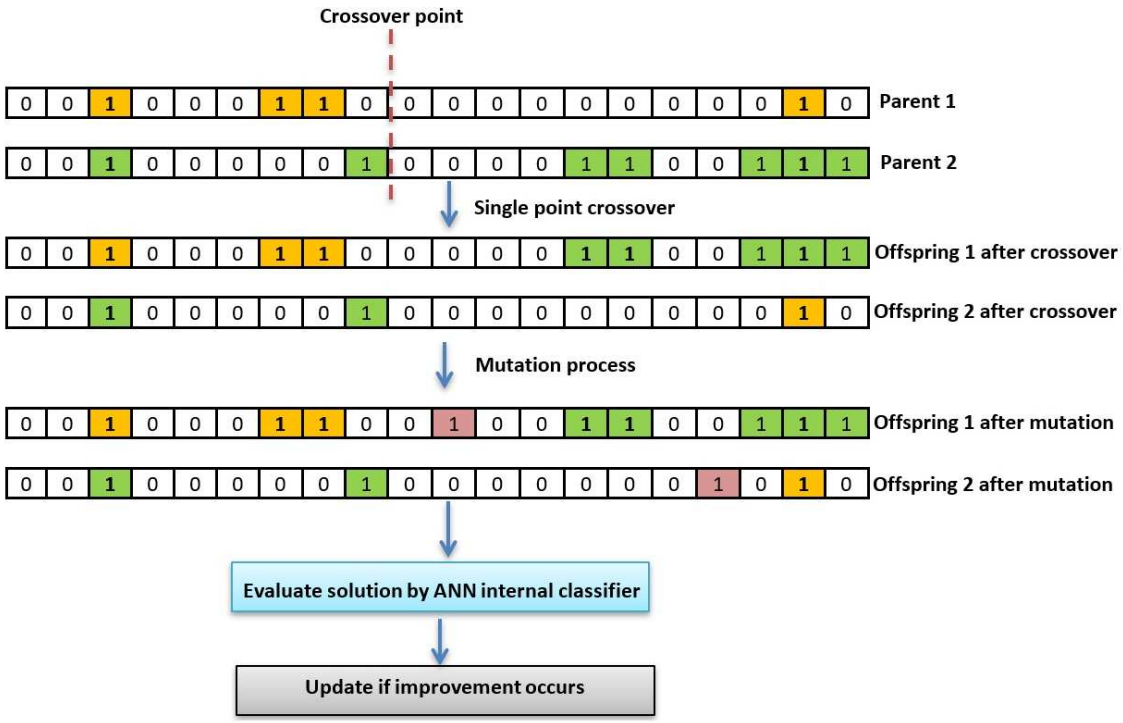


Fig. 5: An illustration of BGA for a single generation. [12].

TABLE I: Parameters setting for ANN.

| Parameters | Values |
|---|---|
| Number of neurons in input layer | Number of selected features |
| Number of neurons in hidden layer | 15 |
| Output layer | (Normal(0) or Have Covid-19 (1) ) |
| Training sample | 75% |
| Testing sample | 15% |
| Validation sample | 10% |
| Fitness function | Mean square error |

method has been implemented using MATLAB-R2019b. Table III reports the parameters setting for BGA. Moreover, we executed each classifier 21 times. To evaluate the proposed method, we employed four different measurements criteria that depends on the confusion matrix (See Figure 6). Equations (2), Eq.(3), (4), and (5) present accuracy, precision, recall, and F-measure criteria, respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F - Measure = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \tag{5}$$

The results of the first type of experiments (i.e., without feature selection) are presented in Table IV. It is clear that the

TABLE II: List of dataset features.

| | Features | Description |
|---|---|---|
| $f_1$ | Hematocrit | The proportion of RBCs in blood |
| $f_2$ | Hemoglobin | A protein found in RBCs that carries oxygen |
| $f_3$ | Platelets | Blood clotting |
| $f_4$ | Red blood Cells | Delivering oxygen to body cells and tissues |
| $f_5$ | Lymphocytes | Adaptive immunity |
| $f_6$ | Leukocytes | Nonspecific protection against infections |
| $f_7$ | Basophils | Pro-inflammatory |
| $f_8$ | Eosinophils | Parasitic infections and allergic reactions |
| $f_9$ | Monocytes | Phagocytosis of pathogens |
| $f_{10}$ | Serum Glucose | The major source of energy for cells and tissues |
| $f_{11}$ | Neutrophils | Protection against bacterial and fungal infections |
| $f_{12}$ | Urea | Monitoring kidney function |
| $f_{13}$ | C reative Protein | An inflammatory marker |
| $f_{14}$ | Creatinine | Monitoring kidney function |
| $f_{15}$ | Potassium | Cell membrane potential and other physiological functions |
| $f_{16}$ | Sodium | Cell membrane potential and other physiological functions |
| $f_{17}$ | Alanine transaminase | Monitoring liver function |
| $f_{18}$ | Aspartate transaminase | Monitoring liver function |

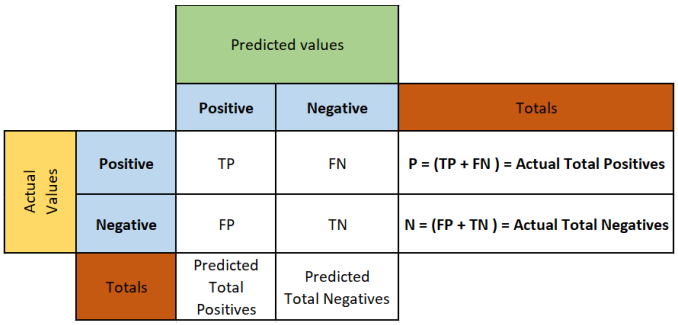


Fig. 6: The confusion matrix.

TABLE III: General parameters setting for Binary genetic algorithm.

| Parameters | Value |
|---|---|
| Length of the solutions | Number of total features |
| Population size | 150 |
| Population type | Bitstring |
| Maximum generations | 5000 |
| Selection types | Roulette wheel selection |
| Crossover rate | 80% |
| Crossover type | Uniform crossover |
| Mutation type | Multiple points |
| Mutation rate | 5% |

performance of CNN classifier us better that all other methods used here based on the accuracy value 76%. However, the reported results show that the performance of kNN was the worst one with an accuracy 0.63. Based on other measurement criteria, it is obvious here that CNN is a more robust model compared to other methods. Figure 7 presents the boxplot diagrams for all machine learning methods without feature selection. It is obvious that the performance of all methods is not stable.

TABLE IV: The obtained results without FS.

| | Accuracy | Precision | Recall | F1-measure |
|---|---|---|---|---|
| CNN | **76%** | 73% | 70% | 75% |
| kNN | 63% | 65% | 60% | 61% |
| NB | 66% | 63 | 62% | 65% |
| C4.5 | 73% | 70% | 72% | 69% |

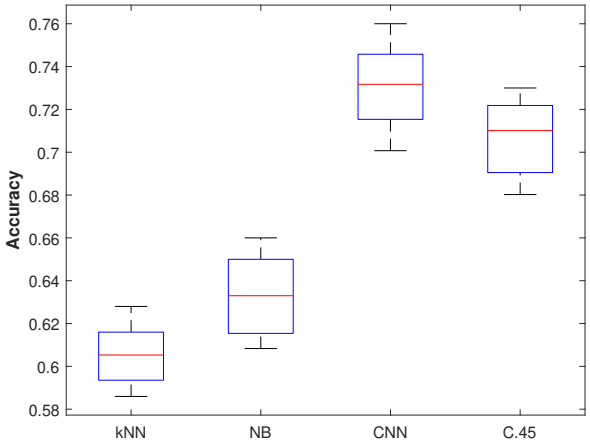


Fig. 7: Boxplots diagrams without feature selection for all classifier.

Table V reports the obtained results with a feature selection algorithm (i.e., BGA). Based on the reported results, the performances of all methods have been improved. For example, the performance of kNN is improved 5%, NB is improved 7%, CNN is improved 4%, and C.45 is improved 4%.Employing BGA as a feature selection method enhances the overall performance of all classifiers. Figure 8 explores the boxplot

diagrams for all classifiers with BGA. The performance of kNN and C.45 is not stable, while the performance of CNN is improved compared to Figure 7. Finally, it is clear that the performance of all classification methods with feature selection are improved with BGA.

TABLE V: The obtained results with FS.

| | Accuracy | Precision | Recall | F1-measure |
|---|---|---|---|---|
| CNN | **80%** | 76% | 74% | 78% |
| kNN | 68% | 66% | 63% | 65% |
| NB | 73% | 67% | 64% | 69% |
| C4.5 | 77% | 74% | 75% | 71% |

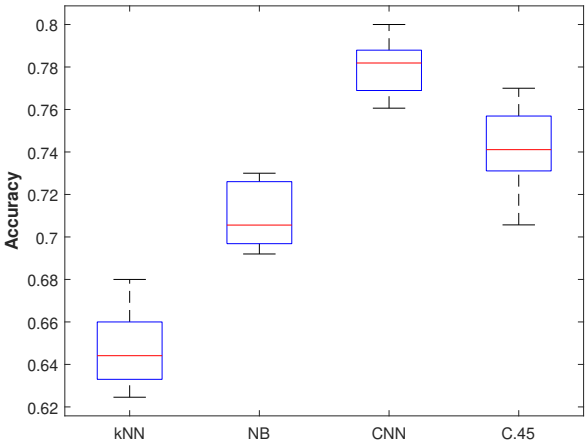


Fig. 8: Boxplots diagrams with feature selection for all classifier.

## V. Conclusion

In this work, we proposed a combination approach (i.e., hybrid) between one of the most well-known wrapper feature selection (i.e., GA) and a different set of classifiers to determine the existence of COVID-19 based on laboratory findings. We examined the performance of four classifiers (i.e., kNN, NB, CNN, and C.45) with a wrapper feature selection (i.e., BGA). We examined the proposed method over a real dataset obtained from Hospital Israelita Albert Einstein at Sao Paulo Brazil. The reported results show that CNN with BGA is able to predict the COVID-19 with accuracy $80\%$. In our future work, we will perform a deep analysis and examine different wrapper feature selections.

## References


[1] R. Ferrer, "Covid-19 pandemic: the greatest challenge in the history of critical care," *Medicina intensiva*, 2020.

[2] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.

[3] W. Kong and P. P. Agarwal, "Chest imaging appearance of covid-19 infection," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 1, p. e200028, 2020.

[4] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. Abo-Elsoud, "A new covid-19 patients detection strategy (cpds) based on hybrid feature selection and enhanced knn classifier," *Knowledge-Based Systems*, vol. 205, p. 106270, 2020.

[5] L. Muhammad, M. M. Islam, S. S. Usman, and S. I. Ayon, "Predictive data mining models for novel coronavirus (covid-19) infected patients' recovery," *SN Computer Science*, vol. 1, no. 4, pp. 1–7, 2020.

[6] G. Gayathri and S. Satapathy, "A survey on techniques for prediction of asthma," in *Smart Intelligent Computing and Applications*. Springer, 2020, pp. 751–758.

[7] T. Thaher and N. Arman, "Efficient multi-swarm binary harris hawks optimization as a feature selection approach for software fault prediction," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, pp. 249–254.

[8] I. Tumar, Y. Hassouneh, H. Turabieh, and T. Thaher, "Enhanced binary moth flame optimization as a feature selection algorithm to predict software fault prediction," *IEEE Access*, vol. 8, pp. 8041–8055, 2020.

[9] T. Thaher, M. Mafarja, B. Abdalhaq, and H. Chantar, "Wrapper-based feature selection for imbalanced data using binary queuing search algorithm," in *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, 2019, pp. 1–6.

[10] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328.

[11] I. Tumar, Y. Hassouneh, H. Turabieh, and T. Thaher, "Enhanced binary moth flame optimization as a feature selection algorithm to predict software fault prediction," *IEEE Access*, vol. 8, pp. 8041–8055, 2020.

[12] H. Turabieh, M. Mafarja, and X. Li, "Iterated feature selection algorithms with layered recurrent neural network for software fault prediction," *Expert Systems with Applications*, vol. 122, pp. 27 – 42, 2019.

[13] M. Mafarja, A. A. Heidari, M. Habib, H. Faris, T. Thaher, and I. Aljarah, "Augmented whale feature selection for iot attacks: Structure, analysis and applications," *Future Generation Computer Systems*, vol. 112, pp. 18–40, 2020.

[14] H. Turabieh, S. Azwari, M. Rokaya, W. Alosaimi, A. Alharbi, W. Alhakami, and M. Alnefaie, "Enhanced harris hawks optimization as a feature selection for the prediction of student performance," *Computing*, 01 2021.

[15] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[16] M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, "Ensemble learning model for diagnosing covid-19 from

routine blood tests," *Informatics in Medicine Unlocked*, vol. 21, p. 100449, 2020.

[17] X. Jiang, M. Coffee, A. Bari, J. Wang, X. Jiang, J. Huang, J. Shi, J. Dai, J. Cai, T. Zhang *et al.*, "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," *CMC: Computers, Materials & Continua*, vol. 63, pp. 537–51, 2020.

[18] A. F. de Moraes Batista, J. L. Miraglia, T. H. R. Donato, and A. D. P. Chiavegatto Filho, "Covid-19 diagnosis prediction in emergency care patients: a machine learning approach," *medRxiv*, 2020.

[19] P. Schwab, A. D. Schütte, B. Dietz, and S. Bauer, "Clinical predictive models for covid-19: Systematic study," 2020.

[20] V. A. de Freitas Barbosa, J. C. Gomes, M. A. de Santana, E. d. A. Jeniffer, R. G. de Souza, R. E. de Souza, and W. P. dos Santos, "Heg. ia: An intelligent system to support diagnosis of covid-19 based on blood tests," *Research on Biomedical Engineering*, pp. 1–18, 2021.

[21] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict covid-19 infection," *Chaos, Solitons & Fractals*, vol. 140, p. 110120, 2020.

[22] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11 – 26, 2017.

[23] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient knn classification with different numbers of nearest neighbors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774–1785, May 2018.

[24] L. A. Breslow and D. W. Aha, "Simplifying decision trees: A survey," *Knowl. Eng. Rev.*, vol. 12, no. 1, pp. 1–40, Jan. 1997. [Online]. Available: http://dx.doi.org/10.1017/S0269888997000015

[25] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1, pp. 105–139, Jul 1999. [Online]. Available: https://doi.org/10.1023/A:1007515423169

[26] C. C. Aggarwal and C. Zhai, *A Survey of Text Classification Algorithms*. Boston, MA: Springer US, 2012, pp. 163–222.

[27] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: a survey of some recent advances," *ESAIM: Probability and Statistics*, vol. 9, pp. 323–375, 2005.

[28] B. Qu, Y. Zhu, Y. Jiao, M. Wu, P. Suganthan, and J. Liang, "A survey on multi-objective evolutionary algorithms for the solution of the environmental/economic dispatch problems," *Swarm and Evolutionary Computation*, vol. 38, pp. 1 – 11, 2018.

[29] S. Mirjalili, *Genetic Algorithm*. Cham: Springer International Publishing, 2019, pp. 43–55.

models for coronavirus disease 2019," *arXiv preprint arXiv:2005.08302*, 2020.

[30] P. Schwab, A. D. Schütte, B. Dietz, and S. Bauer, "predcovid-19: A systematic study of clinical predictive