A Data Mining Based Two-Staged Approach for COVID-19 Diagnosis

Dina Abdelfattah Gouda
Department of Computer Science
Faculty of Computers and Information, Menoufia University,
Egypt.
dinagouda1@gmail.com

Nader Mahmoud
Department of Computer Science
Faculty of Computers and Information, Menoufia University,
Egypt.
nader.mahmoud@ci.menofia.edu.eg

***Abstract*— COVID-19, because of the plague that threatens the entire planet, people continue to live in dread and instability. COVID-19 is considered one of the fastest pandemics that affected the whole world. As an initial step, PCR analysis is becoming more and more popular, although it has issues with accuracy. Using Convolutional neural network (CNN), it takes a while to identify a few computers tomography (CT) images despite its great accuracy in proper classification Additionally, classification using data mining techniques on (CT) images has been researched, and they achieve high accuracy in classification, but it is time-consuming. While scanning (CT) images, Patients are in danger when utilizing imaging equipment. Even after a thorough cleaning, there is a high likelihood that the virus will persist on the surface of the scanning chamber.**

This paper presents a COVID-19 patient diagnosis approach that apply to find a quickly and highly effective classification using a blood test. The suggested approach has two main phases: a feature selection phase and covid-19 prediction phase. The feature selection is applied by using The Chi-Square feature selection technique. This methodology, a filter feature selection technique, can swiftly identify the most useful subset of features. Then, an improved support vector machine is used to deliver an immediate and accurate diagnosis (RBF SVM). The basic concept behind the suggested RBF SVM is choosing the value of hyperparameters, C and Y, in the supporting vector machine (SVM) formula before computing the SVM. The results demonstrate that the suggested method provides an accuracy of up to 98.8%.

Keywords*— *COVID-19, Computerized Tomography, Chest X- ray, CNN, RBF SVM.

I. Introduction

The World Health Organization (WHO) confirmed coronavirus illness in China in December 2019[1]. The epidemic has spread fast, and due to that, the world has lost millions of people since the appearance of this epidemic until now.

Despite its appearance in three years only, this epidemic is considered the most effective epidemic facing the world. Immediate and accurate diagnosis is the most incredible way to lower the risks of this epidemic and the likelihood that it will occur, despite the medical team's and researchers' best efforts to find a cure.

In the beginning, the whole world is heading to Real-Time Reverse Transcription Polymerase Chain Reaction (RT-PCR) [2], which is the most common used traditional method for diagnosing COVID-19. Nevertheless, these conventional methods take a lot of time and are not very accurate.

On the other hand, medical images (e.g. (CT) images and X-rays) play an important role in COVID-19 detection [3]. However, a normal chest CT scan is performed on the great majority of COVID-19 patients. Furthermore, using imaging technology for COVID-19 patients puts doctors and other patients at danger. Even after a thorough cleaning, there is a great likelihood that the virus will remain on the surface of the scanning chamber. A CT scan is also fairly pricey, and it can be necessary to perform it on the same patient multiple times in order to secure his recovery.

Because each patient has unique characteristics, blood testing has been extensively researched as an alternative method of identifying COVID-19 patients. Blood testing has been associated with significantly reduced lymphocyte stiffness, increased monocyte cell size, the manifestation of smaller and less foldable erythrocytes, and the existence of large, foldable, activated neutrophils. As a result, blood analysis was used in this study to limit the hazards related to the use of alternative methods, to lessen direct patient contact with medical staff, and to lessen the chance of COVID-19 prevalence as a result of direct patient-staff contact.

In this paper, we present an approach for COVID-19 detection using blood tests. The presented approach consists of two stages. Firstly, the feature selection stage uses the Chi-Square Feature Selection method. We chose Chi-Square as our main Feature Selection due to its high performance, as reported in other research works [4]. Secondly, we apply the RBF SVM to the set of the selected features. We choose the SVM classifier to improve it because it has a high performance in other research works.

The remnant of the paper is presented as follows; sec. II presents a literature review of relevant works. Sec. III presents a methodology overview. Experimental results are discussed in Sec. IV. Conclusion and future work are presented in Sec. V.

II. Related Works

Reverse Transcription-Polymerase Chain Reaction (RT-PCR) is used to initially screen for the presence of the coronavirus infection since COVID-19 is thought to be the most common disease in the world. The RT-PCR test, however, requires a lot of time and has a significant false-negative rate [5]. In [6], the author used Distance Biased Naïve Bayes (DBNB) using laboratory tests. The DBNB consists of two stages to classify COVID-19. In order to select more beneficial characteristics from the dataset, the author first employed advanced Particle Swarm Optimization (APSO), which combines filter and wrapper methods. Second, DBNB is employed to get around the problem with traditional NB. The detection accuracy of DBNB, which consists of two modules, was 94.2%. The first module was used to weigh the chosen features, and the second was used to make a final choice based on the distance between the center of the target class and the input patient that needs to be classified.

In [7], a support system for COVID-19 diagnosis using blood tests named Heg. IA. In the preprocessing stage, the dataset has a numeric form. Then, Particle Swarm Optimization (PSO) and Evolutionary Search (ES) are applied as feature selection techniques. To detect COVID-19 patients using several classification techniques, but Bayes net method achieves high accuracy, equals 95%. Using a blood sample, a COVID-19 Diagnostic Technique (CDT) for COVID-19 detection was introduced in [8]. The dataset has 100 features, and CDT used Genetic methods (GA) and relief methods to choose the best accurate feature. The dataset was then classified using a random forest classifier. The result shows how accurate the model is. A novel Corona Patients Detection Strategy (CPDS) was proposed to check COVID-19 patients, as described in [9]. Two steps were used to implement CPDS. The first is a Hybrid Feature Selection Methodology (HFSM) for extracting the best feature from the dataset. The second stage is an Enhanced K-Nearest Neighbor (EKNN) implemented as a classifier for detecting COVID-19 patients. In [10] Improved K-nearest neighbor (IKNN) on a blood sample dataset was applied. The IKNN consists of two stages. Firstly, the author used chi-square feature selection (CSFS) to choose the most effective features. Secondly, applied the IKNN to predict. It yielded a detection accuracy of: 96.3%.

To summarize, many researchers head to utilize blood tests and avoid using (CT) [11-13] images for the above-mentioned reasons and work on how to improve the traditional techniques of data mining to achieve a high accuracy [14]. It has been experimentally validated that SVM yielded superior detection accuracy. So, this paper introduces the implementation of the improved version of SVM, which is RBF SVM, on blood test datasets. Thus, the proposed approach consists of two- stages.

1- Feature selection: The Chi-Square Feature Selection method is used to choose the features during the feature selection step. In actuality, this is a method of filter feature selection that can quickly select the most useful subset of features.

2-Covid-19 prediction stage: RBF SVM is a powerful technique which is the improved version of SVM and it depends on how to select the best values of parameters of RBF SVM formula [15], specially the values of C and Y:

$Min \frac{1}{2}\ ||w||^2 + c\ \sum_{i=1}^{n} \varepsilon_i$

Subject to $y_i(w\, x_i + b) \geq 1 - \varepsilon_i$

Where

$||w||^2$ = *the normal vector*

ε = *the distance to the correct margin with* $\varepsilon_i \geq 0$, $i = 1,..n$

C= a regularization parameter

y_i= *the i-th target value* (1)

III. Methodology

The ability to detect pandemics is improving, especially when making snap judgments regarding risky viral diseases like COVID-19. COVID-19 patients need to be recognized as soon as possible because their social interactions are increasing the number of infected persons [16–18].

Fig.1 shows the main stages of the suggested approach. It starts with a feature selection stage, followed by the classification stage. Feature selection is first introduced in Sec. A, and Sec. B presented the COVID-19 Prediction Stage.

A. Feature Selection Phase

In the feature selection phase, the Chi-Squared Feature Selection method is applied to measure the most effective

Features [19–21], The degree of independence between two variables belonging to the same class category is determined by the Chi-Squared Feature Selection method utilizing (2).

$$x_2^c = sum\left(\frac{(O_i - E_i)^2}{E_i}\right)$$

Where

C = degree of freedom

O=observed value(s)

E= expected value(s) (2)

Chi-Square measures how expected count E and observed count O deviate from each other. The higher the Chi-Square value of the feature, it can be selected for model training. When the observed count is close to the expected count, thus we will have a smaller Chi-Square value.

B. COVID-19 Prediction Stage

In the COVID-19 prediction stage, RBF SVM is chosen as our main classification method. RBF SVM is a new model of SVM classifier that can improve the efficiency of the classification, which can be implemented in various applications, including COVID-19 patient prediction. SVM is a well-known classifier that proves its classification efficiency in different applications [22]. SVM makes a decision based on formula (1). This formula mainly depends on two variables, where the quality of prediction results is highly dependent on these two parameters, C and Y.

In the literature, there have been numerous suggestions for how to find the best set of hyperparameters for a certain dataset, including general optimization techniques [23] as grid search, random search, simulated annealing, and Bayesian optimization.

This step is performed using the Random Search algorithm over pairs of C and Y and testing each pair using some cross-validation procedure. Examples of cross-validation procedures are Nested cross-validation, single cross-validation, and others. A nested cross-validation methodology is usually recommended to assess the performance of the models.

Each data set is separated into K1 partitions in the nested cross-validation process, and then each of these divisions is further divided into K2 partitions. The average validation accuracy of every conceivable combination of C and hyper-parameter values is evaluated using the inner partition (fitness value). Utilizing the best individual return from the inner loop tuning procedure, the test accuracy for the data in the outer loop is evaluated.

This approach has a significant computational cost because each technique is used K1 x K2 times to assess performance, but it minimizes the bias of the data when inducing models. One data set can be evaluated by 100 models if the data is divided into 10-folds and looped twice.

It is possible that the nested cross-validation method, especially with hyper-parameter adjustment, won't work in practical projects. So, this is why we employ the single cross-validation approach.

By dividing the data into training, validation, and test partitions, the single cross-validation method is applied. Fig. 2 introduces the single cross-validation process. every time a tuning method is used. There are k stratified partitions created from the data set. SVM is trained using k - 2 partitions (**training folds**), and the method finds a solution for each candidate. The remaining partition is divided into a (**test fold**) and a (**validation fold)**, respectively, to test the model.

The model created by integrating the training partitions is used to evaluate the test and validation accuracies, and the values of the hyperparameters are discovered using an optimization technique. In a single cross-validation, this process is done for all k permutations.

The final result is the person with the highest validation accuracy (along with the values of their hyper-parameters), and their technical performance is their average test accuracy. In this study, we applied a single cross-validation experimental approach.

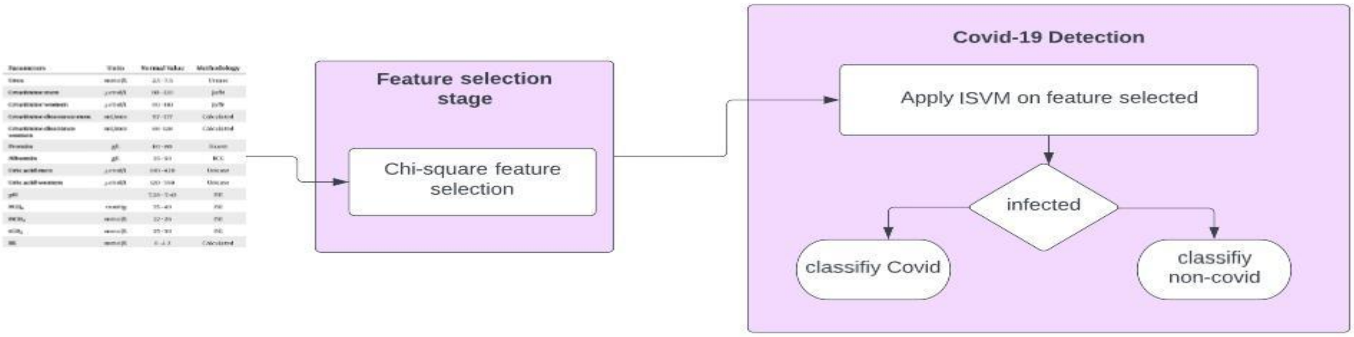


Fig.1: the main phases of the proposed approach

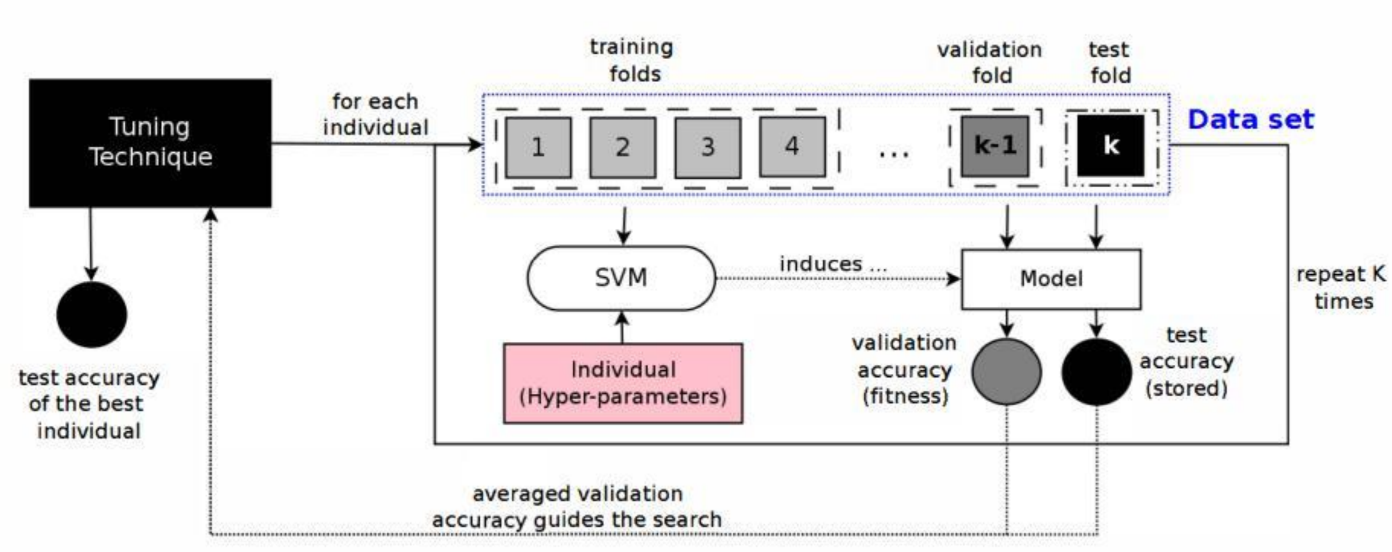


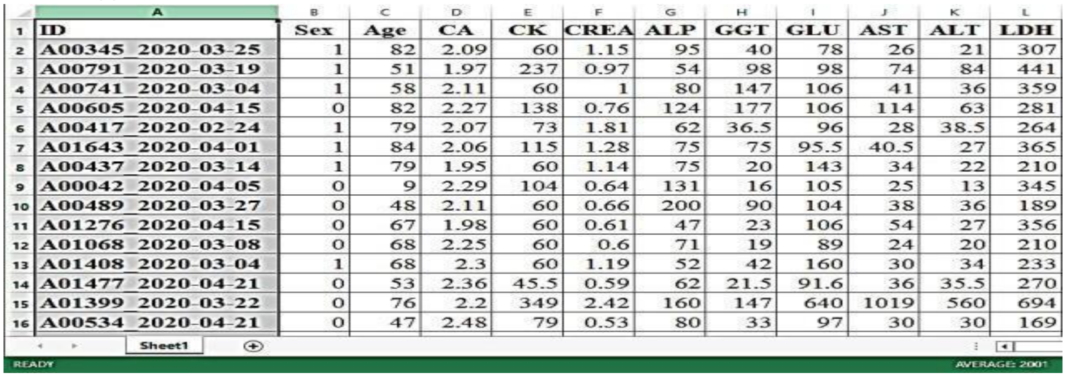
Fig.2: Single cross-validation experimental methodology for hyper-parameter tuning.

IV. Experimental Results and Discussion

This section introduces a quantitative evaluation of the suggested approach. Firstly, the datasets are presented in Sec. A. We present an evaluation of the common diagnostic strategies in Sec. B. Sec. C presents the computational time of each step of the suggested approach. Finally, Sec. D presents a results comparison.

A. Experimental Setup and dataset

In this paper, Dataset [21] is implemented to test the performance of the suggested approach. This dataset contains a group of “COVID” and “Non- COVID” patient laboratory data. This dataset consists of 5645 patients and 112 features. In the feature selection stage, the Chi-Squared feature selection approach is used, producing a list of the top 49 features. We divided the dataset into 30% for testing and 70% for training. There are 1693 test patients and 3952 training patients total. A sample of an image from our dataset is shown in Fig. 3. On a PC with an AMD CPU and 10 Cores, 4C+6G, 1.8 GHz, and 16.0 GB RAM, all the trials were carried out.



	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	Sex	Age	CA	CK	CREA	ALP	GGT	GLU	AST	ALT	LDH
2	A00345_2020-03-25	1	82	2.09	60	1.15	95	40	78	26	21	307
3	A00791_2020-03-19	1	51	1.97	237	0.97	54	98	98	74	84	441
4	A00741_2020-03-04	1	58	2.11	60	1	80	147	106	41	36	359
5	A00605_2020-04-15	0	82	2.27	138	0.76	124	177	106	114	63	281
6	A00417_2020-02-24	1	79	2.07	73	1.81	62	36.5	96	28	38.5	264
7	A01643_2020-04-01	1	84	2.06	115	1.28	75	75	95.5	40.5	27	365
8	A00437_2020-03-14	1	79	1.95	60	1.14	75	20	143	34	22	210
9	A00042_2020-04-05	0	9	2.29	104	0.64	131	16	105	25	13	345
10	A00489_2020-03-27	0	48	2.11	60	0.66	200	90	104	38	36	189
11	A01276_2020-04-15	0	67	1.98	60	0.61	47	23	106	54	27	356
12	A01068_2020-03-08	0	68	2.25	60	0.6	71	19	89	24	20	210
13	A01408_2020-03-04	1	68	2.3	60	1.19	52	42	160	30	34	233
14	A01477_2020-04-21	0	53	2.36	45.5	0.59	62	21.5	91.6	36	35.5	270
15	A01399_2020-03-22	0	76	2.2	349	2.42	160	147	640	1019	560	694
16	A00534_2020-04-21	0	47	2.48	79	0.53	80	33	97	30	30	169

Fig.3. A snap shot from Dataset [21] (which CA stands for Hypercalcemia, CK stands for creatine kinase level, CREA stands for creatinine level, a ALP stands for Alkaline Phosphatase, GGT stands for Gamma-Glutamyl Transferase, GLU stands for Glucose, AST/ALT stands for aspartate aminotransferase and LDH stands for Lactate Dehydrogenase.

B. Benchmark Evaluation

In this section, we quantitatively estimated the proposed approach and alternatives. We apply the proposed approach to a dataset [24] and alternatives to the IKNN [10], CPDS [9], and DBNB [6]. Table I shows the results in which our proposed approach yielded 98.8% prediction accuracy, and our alternatives yielded 90.8%, 88.0%, and 81%, respectively. All the algorithms are applied in our dataset [24] to standardize the comparison factors, and we obtain the source code, flowchart, or pseudo code from [10, 9, 6].

It is clear from the results shown in the table that our proposed approach obtained the highest results. This is because it works well with a dataset that has a clear separation margin between classes, and this is already present in any blood test dataset. And to better evaluate the approach. We applied the proposed approach to different datasets. Table II shows the results.

C. Computational Time

Table III present the computational time of the proposed approach and the common diagnostic strategies. The computational time of the proposed approach has been estimated and compared to a well-known diagnostic approach on the selected dataset It can be seen from Table III that the proposed approach runs significantly faster than recent alternatives. The proposed approach yielded the lowest computation time, approx. 2 min, which is far less than IKNN that, yielded approx. 5min. The average computation time for each stage of the suggested method is shown in Table IV. We measured computational time for the main steps. In the first step, we select the most important features. This step takes 1.0 minutes. In the second step, we choose the best values for C and Y (hyper-parameter tuning step). This step takes 0.4 minutes. In the last step, we apply the COVID-19 detection using RBF SVM. This step takes 0.6 minutes.

D. Results Comparison

We present a quantitative evaluative comparison with the most pertinent research works in this part. We present in Table IIV a quantitative comparison between the proposed approach and pertinent comparable research works, where the error percentage is calculated by subtracting the accuracy from 100, the sensitivity is determined by the ratio of true positives to false positives, the precision is determined by the ratio of true positives to false positives, and the accuracy is determined by adding the number of correctly classified cases.

V. CONCLUSION

An technique based on RBF SVM has been suggested in this paper for COVID-19 classification in the dataset of blood tests. This method is divided into two parts, the first of which uses the Chi-Squared Feature Selection technique. The COVID-19 Prediction Stage is then applied using RBF SVM. The sensitivity problems with (CT) pictures can be handled by this paradigm. A significant blood test dataset was used to evaluate the model and compare it to other models. According to experimental findings, the models were able to reach high accuracy in the allotted time. We intend to use several categorization algorithms after using the suggested approach to further increase accuracy.

TABLE I: RBF SVM and conventional diagnostic techniques are compared

Techniques	***IKNN[10]***	***DBNB[6]***	***CPDS[9]***	***Proposed Approach***
Accuracy	90.8%	88.0%	81%	**98.8%**

TABLE II: A COMPARISON BETWEEN APPLING PROPOSED APPROACH ON DIFFERENT DATASETS.

Dataset	***Dataset[25]***	***Dataset[26]***	***Dataset[27]***
Accuracy	98.2%	98.9%	98.5%

TABLE III: COMPUTATIONAL TIME BETWEEN RBF SVM AND THE COMMON DIAGNOSTIC STRATEGIES

Techniques	***IKNN[10]***	***DBNB[6]***	***CPDS[9]***	***Proposed Approach***
Time in minutes	5.00	6.00	7.00	**2.00**

TABLE IV: CALCULATION TIME FOR EACH STEP OF THE SUGGESTED METHOD

Step	***Feature classification***	***Hyper-parameter tuning***	***COVID-19 detection***
Time in minutes	1.0	0.6	0.4

TABLE IIV: A FULL COMPARISON BETWEEN RBF SVM AND THE COMMON DIAGNOSTIC STRATEGIES

Techniques	***IKNN[10]***	***DBNB[6]***	***CPDS[9]***	***RBF SVM***
Accuracy	90.8%	88.0%	81%	**98.8%**
Error	9.2%	12%	19%	**1.2%**
Sensitivity	96.2%	90.8%	88.4%	**98.2%**
Precision	90.1%	88.2%	85.2%	**95.7%**

REFERENCES

[1] WHO. coronavirus disease (COVID-19) pandemic [online]

From: https: //www.who.int/ [Accessed: 06-May-2023].

[2] Willard M. Freeman, Stephen J. Walker et al. “Quantitative RT-PCR: Pitfalls and Potential”. BIOTECHNIQUESVOL. 26, NO. 1, (2018)

[3] Fan, Xiaole et al. “COVID-19 CT image recognition algorithm based on transformer and CNN.” Displays vol. 72 (2022)

[4] Ratul, Ishrak Jahan, Ummay Habiba Wani, Mirza Muntasir Nishat et al." Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers by Performing Chi-squared Test and Hyper-parameter Optimization: A Retrospective Analysis." arXiv preprint (2022).

[5] D. Ferrari, A. Motta, M. Strollo, G. Banfi, and M. Locatelli, “Routine

blood tests as a potential diagnostic tool for COVID-19,” Clin. Chem.

Lab. Med., vol. 58, no. 7, pp. 1095–1099, (2020)

[6] W. Shaban, A. Rabie, A. Saleh, and M. Abo-Elsoud, “Accurate

Detection of COVID-19 Patients Based on Distance Biased Naıve Bayes (DBNB) Classification Strategy,” Pattern Recognition, Elsevier, vol.119, pp.108110, (2021)

[7] V. A. de Freitas Barbosa et al., “Heg. IA: An intelligent system to

support diagnosis of COVID-19 based on blood tests,” Res. Biomed. Eng., pp. 1–18, (2021)

[8] R. I. Doewes, R. Nair, and T. Sharma, “Diagnosis of COVID-19

through blood sample using ensemble genetic algorithms and machine

learning classifier,” World J. Eng., (2021).

[9] W. Shaban, A. Rabie, A. Saleh, et al., “A new COVID-19 Patients

Detection Strategy (CPDS) based on hybrid feature selection and

enhanced KNN classifier,” Knowl. Based Syst. Elsevier, vol.205, pp.1– 18, (2020)

[10] Mohamed, Alaa Mostafa, Ahmed Saleh, Doaa A Altantawy, and Mohy Eldin Ahmed Abo-Elsoud. "COVID-19 Patients Diagnosis (CPD) Strategy Using Data Mining Techniques." MEJ. Mansoura Engineering Journal 47, no. 2 (2022)

[11] Abayomi-Alli OO, Damaševičius R et al. "An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples." Sensors (Basel) (2022)

[12] Yury V. Kistenev, Denis A. Vrazhnov et al."Predictive models for COVID-19 detection using routine blood tests and machine learning",Heliyon, Vol.8(2022)

[13] Cabitza, Federico, Campagner, Andrea et al. "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests" Clinical Chemistry and Laboratory Medicine (CCLM), vol. 59(2021)

[14] Goda, Dina Abdelftah, and Nader Mahmoud. "Improving COVID 19 Detection based on a hybrid data mining approach." IJCI. International Journal of Computers and Information 9, no. 2 (2022)

[15] Wainer, Jacques, and Pablo Fonseca. "How to tune the RBF SVM hyperparameters? An empirical evaluation of 18 search algorithms." Artificial elligence Review 54, no. 6 (2021)

[16] M. Allam, and M. Nandhini,” A Study on Optimization Techniques in Feature Selection for Medical Image Analysis,” International Journal on Computer Science and Engineering (IJCSE), Vol. 9, no.3, PP. 75-82, (2017).

[17] Y. Li, “Text feature selection algorithm based on Chi-square rank correlationfactorization,”JournalofInterdisciplinary Mathematics,

Taylor& Francis Group, Vol. 20, no.1, PP.153-160 , (2017)

[18] P. Samant and R. Agarwal, “Machine learning techniques for medical

diagnosis of diabetes using iris images,” Comput. Methods Programs

Biomed., vol. 157, pp. 121–128, (2018)

[19] Ratul, Ishrak Jahan, Ummay Habiba Wani, Mirza Muntasir Nishat et al." Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers by Performing Chi-squared Test and Hyper-parameter Optimization: A Retrospective Analysis." arXiv preprint (2022).

[20] Sharma, Sanur, and Anurag Jain. "Hybrid ensemble learning with feature selection for sentiment classification in social media." In Research Anthology on Applying Social Networking Strategies to Classrooms and Libraries. IGI Global, (2023).

[21] Sadeghi, Hamid, and Abolghasem-A. Raie. "Histnet: Histogram-based convolutional neural network with chi-squared deep metric learning for facial expression recognition." Information Sciences 608 (2022).

[22] Iwendi, Celestine, Kainaat Mahboob, Zarnab Khalid, Abdul Rehman Javed, Muhammad Rizwan, and Uttam Ghosh. "Classification of COVID-19 individuals using adaptive neuro-fuzzy inference system." Multimedia Systems (2021)

[23] Anyanwu, Goodness Oluchi, Cosmas Ifeanyi Nwakanma, Jae-Min Lee, and Dong-Seong Kim. "Optimization of RBF-SVM Kernel using Grid Search Algorithm for DDoS Attack Detection in SDN-based VANET." IEEE Internet of Things Journal (2022).

[24] Kaggle COVID dataset [online] Available at
https://www.kaggle.com/datasets/einsteindata4u/covid19
[Accessed: 01-MAR-2023].

[25] Kagle COVID dataset [online] Available at

https://www.kaggle.com/datasets/plarmuseau/forecast-covid-death

[Accessed: 25-JUN-2023].

[26] Kagle COVID dataset [online] Available at

https://www.kaggle.com/datasets/meirnizri/covid19-dataset

[Accessed: 25-JUN-2023].

[27] Kagle COVID dataset [online] Available at

https://www.kaggle.com/datasets/tawsifurrahman/covid19-complete-blood-count-clinical-database
[Accessed: 25-JUN-2023].