ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Diagnosis of COVID-19 Infection via Association Rules of Cough Encoding

Suhaila N. Mohammed[1]

ABSTRACT

COVID-19 has roused the scientific community, prompting calls for immediate solutions to avoid the infection or at least reduce the virus's spread. Despite the availability of several licensed vaccinations to boost human immunity against the disease, various mutated strains of the virus continue to emerge, posing a danger to the vaccine's efficacy against new mutations. As a result, the importance of the early detection of COVID-19 infection becomes evident. Cough is a prevalent symptom in all COVID-19 mutations. Unfortunately, coughing can be a symptom of various of diseases, including pneumonia and influenza. Thus, identifying the coughing behavior might help clinicians diagnose the COVID-19 infection earlier and distinguish coronavirus-induced from non-coronavirus-induced coughs. From this perspective, this research proposes a novel approach for diagnosing COVID-19 infection based on cough sound. The main contributions of this study are the encoding of cough behavior, the investigation of its unique characteristics, and the representation of these traits as association rules. These rules are generated and distinguished with the help of data mining and machine learning techniques. Experiments on the Virufy COVID-19 open cough dataset reveal that cough encoding can provide the desired accuracy (100%).

Article information:

Keywords: COVID-19, Cough Encoding, Spectral Features, Apriori Algorithm, Association Rules, SVM

Article history:
Received: April 15, 2022
Revised: May 21, 2022
Accepted: August 16, 2022
Published: February 25, 2023 (Online)

DOI: 10.37936/ecti-cit.2023171.248192

1. INTRODUCTION

The pandemic of Coronavirus Disease 2019 (COVID-19) is posing a significant threat to worldwide public health [1]. As of February 23, 2022, approximately 428M confirmed cases and 5.91M people have lost their lives worldwide due to this disaster [2]. Despite significant efforts, the lack of a definitive cure for this disease worries researchers and the medical world alike. Several strategies been developed to protect people from the pandemic. Social distance, frequent hand washing, using of facial masks, and avoiding touching the face as much as possible are among the measures taken [3].

According to the World Health Organization (WHO) [4], COVID-19 symptoms include a dry cough, headache, fever, fatigue, breathing difficulty, muscle pain, and sore throat. Dry cough is one of the more prevalent symptoms of respiratory tract infections, occurring in 68% to 83% of patients who come in for a medical check. However, these symptoms are easily mistaken for a cold or the flu [5].

If non-coronavirus-induced coughs are discriminated against coronavirus-induced coughs through machine learning algorithms, a cost-effective, easy-to-use, fast, and early diagnosis system can be offered. Such a system can be installed as a smartphone app or displayed to users in a web-based environment. Suspected candidates can record their cough sounds on their smartphones at any time to make a preliminary assessment of their situation. It can also relieve the load on healthcare staff by effectively minimizing hospital congestion [3]. Consequently, applying machine learning and signal processing techniques to diagnose COVID-19 from cough sound recordings has become one of the most popular and essential topics of recent studies [6].

Till now, different types of features have been extracted by researchers to represent cough behavior. Solak (2021) [6] described the cough sounds using Multifractal Detrended Fluctuation Analysis (MDFA), Lempel–Ziv Complexity (LZC), and en-

[1] The author is with Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq, E-mail: suhailan.mo@sc.uobaghdad.edu.iq

tropy measures. For classification, a Support Vector Machine (SVM) was utilized. As a result of his study, 95.8% accuracy was reached using the Virufy dataset. Tena et al. (2022) [7] extracted a set of time–frequency features from the Wigner Distribution (WD) and used a Random Forest (RF) model for distinguishing COVID-19 coughs. An accuracy of 94.81% was obtained using the Virufy dataset. Manshouri (2022) [3] used power spectral density based on short-time Fourier transform and Mel-Frequency Cepstral Coefficients (MFCCs) features. Regarding results evaluation, the classification accuracy was 95.86% on the Virfuy dataset using an SVM classifier.

Chowdhury et al. [8] proposed a multi-criteria decision-making (MCDM) method that combines ensemble technologies based on the entropy metric. In addition, the feature set is reduced through recursive feature elimination with cross-validation under different estimators. A precision of 89% was achieved by MCDM using the Virufy dataset.

Although prior studies attempted to identify SARS-CoV-2 virus infection using various temporal and spectral domain characteristics, they failed to consider the relationships that may exist among the sequence of features generated from the cough frames' series. To the best of our knowledge, this is the first work in this research area that mines the association rules between frames of cough sound. The main contributions of this work can be summarized as follows:

1. Using numerical cough encoding, cough sound characteristics are represented as discrete values rather than continuous values.
2. Instead of using each feature held in each sound frame independently, association rules are mined to discover the relationships between cough encoding sequences.
3. A binary feature vector is presented to emulate the final feature vector of the association rules. The vector represents the ability of each rule to be applied to the examined cough sound.

Besides this introductory section, the remaining portion of the paper is organized as follows: Section 2 provides a brief presentation of the basic concepts of the methods used in the proposed work. Section 3 gives a detailed description of the proposed system. Section 4 illustrates the experimental results achieved when applying the proposed approach to a real-life clinical cough dataset. Finally, work conclusions and ideas for future work are provided in section 5.

2. THEORETICAL BACKGROUND

Basic concepts for the data mining and machine learning methods used in this work are described below.

2.1 Association Rule Mining

Association rule mining is a technique for systematically discovering correlations and patterns in large structured databases, such as transactional databases or other data repositories. The primary principle behind this approach is to divide the problem down into two subtasks [9]:

1. First of all, the algorithm attempts to determine the frequent item sets within a given database using predetermined minimum support and confidence values. Mining frequent item sets leads to the finding of relationships and correlations among items in large transactional or relational databases [10].
2. The second subtask is to generate association rules between the frequent item sets that have been previously identified. An association rule is a logical implication of the form X→Y such that X is the antecedent, Y is the consequent of the rule, and X and Y are disjoint item sets. Different rule-interest measures try to quantify the dependence between X and Y [11]:

1. *Support:* It is the frequency (probability) of the entire rule $(X \rightarrow Y)$ in a given transaction (T) concerning the total number of records in a database (D). It can be defined by using equation (1).

$$Support(X \rightarrow Y) = \frac{|\{T \in D | X \bigcup Y \subseteq T\}|}{|D|} \quad (1)$$

2. *Confidence:* It is the conditional probability that a transaction (T) in a database (D) contains the consequent Y, given that it includes the antecedent X. It can be defined by using equation (2).

$$confidence(\mathrm{X} \rightarrow Y) = \frac{Support(X \bigcup Y)}{Support(X)} = \frac{|\{T \in D | X \bigcup Y \subseteq T\}|}{|\{T \in D | X \subseteq T\}|} \quad (2)$$

Apriori and Frequent Pattern (FP) growth are examples of algorithms for robust association rule finding. Apriori employs an iterative level-wise search approach to generate frequent item sets. First, the collection of frequent 1-itemsets is found by scanning the database to accumulate the count for each item and then collecting those items that satisfy the minimum support and minimum confidence conditions. The resulting set is denoted by L_1. Next, L_1 is used to find L_2, the collection of frequent 2-itemsets, which is used to find L_3, and so on, until no more frequent k-item sets can be found. The finding of each L_k requires one complete scan of the database [10].

2.2 Support Vector Machine

SVM is one of the most essential machine learning algorithms that can be used for classification tasks. The main point in this algorithm is finding the optimal hyper-plane that separates positive class samples (+1) from negative class samples (-1) [12]. The patterns that lie on the edges of the hyper-plane are called support vectors. The perpendicular distance

Fig.1: *The block diagram of the proposed method.*

between the line of margin and the edges of the hyperplane is known as the margin. One of the objectives of SVM is to maximize this margin for better classification. SVM kernels can be linear or non-linear, such as polynomial kernels, radial basis function kernels, etc. [13].

3. MATERIAL AND METHOD

3.1 Dataset Description

The Virufy dataset [14], the first free, publicly available dataset for COVID-19 coughs [15], has been used in this study for model validation purposes. The dataset consists of 121 segments of cough sounds belonging to 16 patients. Forty-eight of these segments are labeled as "*COVID-19*" coughs since they belong to 7 patients with a positive Polymerase Chain Reaction (PCR) test. The other 73 segments belong to 9 patients with negative PCR tests and are labeled as "*Non-COVID-19*" coughs. Each segment is recorded at a sampling frequency of 48 kHz and is approximately one second long.

3.2 The Proposed Method

The proposed method involves six main stages: sound framing and windowing; spectral feature extraction; spectral feature encoding; association rule generation; association rule matching; and classification. The block diagram of the proposed method is shown in Fig. 1.

3.2.1 Sound Framing and Windowing

First of all, the cough sounds are divided into a number of frames (*frmc*), each with a frame size (*frms*) equal to the total number of samples in the whole sound divided by *frmc*. The framing step is followed by a windowing process to avoid the ripples due to sound framing. A hamming window is utilized for the windowing task. It can be defined using equation (3) [16].

$$w[i]=\begin{cases} 0.54-0.46\ \cos\left(\frac{2\pi i}{frms}\right) & 0\le i\le frms \\ 0 & otherwise \end{cases} \quad (3)$$

Where i is the index of the i^{th} frame in the window and $w[i]$ is the result of the windowing process.

3.2.2 Spectral Feature Extraction

To extract spectral features from each frame $frm(x_i)$ in the set of frames $frmc(Frm(C_x) = \{frm(x_1), frm(x_2), \ldots, frm(x_{frmc})\})$ of the cough sound C_x, the frame must be first transformed from the time domain into the spectral domain by applying Short Term Fourier Transform (STFT) using the following equation [17]:

$$F(v_i) = \frac{1}{frms}\sum_{j=0}^{frms-1} frm(x_i(j))e^{-g2\pi\frac{(v_i(j)x_i(j))}{frms}} \quad (4)$$

Where $frm(x_i)$ is a given speech frame (i) with length frms, $F(v_i)$ is the result of applying the Fourier transform on $frm(x_i)$ and $g=\sqrt{-1}$.

Four different spectral features are then extracted from the spectrum of each sound frame [18, 19]:

A) Spectral Centroid

The spectral centroid (SC_i) is the center of the 'gravity' of the spectrum. The value of the spectral centroid of the i^{th} cough sound frame F_i with size $frms$ is defined as:

$$SC_i = \frac{\sum_{j=1}^{frms} jF_i(j)}{\sum_{j=1}^{frms} F_i(j)} \quad (5)$$

B) Spectral Flatness

Spectral flatness (SF_i) indicates whether the frame spectrum distribution is spiky or smooth. It is defined as follows:

$$SF_i = \frac{\sum_{j=1}^{frms} e^{\log F_i(j)}}{SC_i} \quad (6)$$

Where $frms$ is the size of the frame F_i and SC_i is

the spectral centroid of that frame.

C) Spectral Bandwidth

Spectral bandwidth (SB_i) is the difference between the upper (U_i) and lower frequencies (L_i) in the spectrum of the i^{th} frame. It can be defined as:

$$SB_i = U_i - L_i \tag{7}$$

D) Spectral Deviation

The spectral deviation (SD_i) measures the deviation degree of the frame F_i frequencies from the center of gravity SC_i. It is calculated as follows:

$$SD_i = \sqrt{\frac{\sum_{j=1}^{frms}(F_i(j) - SC_i)^2}{frms}} \tag{8}$$

The extracted spectral features are then normalized to be within the range [0-1]. Fig. 2 shows an example of the extracted spectral characteristics for a Non-COVID-19 cough sample with $frmc = 30$. As illustrated in the figure, the generated features have continuous values ranging from 0 to 1, making it impossible to create association rules if the traits remain in this continuous form.

3.2.3 Spectral Feature Encoding

The extracted spectral features of each frame must be transformed from continuous values to discrete values using feature encoding to successfully generate rules that express the associativity in the typical patterns of the sequence of cough frames. The encoding is done by assigning a number from 1 to 20 to each feature value. It's dependent on the feature value's range, as indicated in Table 1. The range map has a 0.15 interval. For the same cough sound sample as in Fig. 2, Fig. 3 provides an example of encoding the extracted spectral features. The extracted features in this figure have discrete values, making them suitable for mining association rules.

Table 1: *Feature encoding table.*

Interval	code	Interval	code	Interval	code
[0.0-0.05]	1	[0.05-0.1]	2	[0.1-0.15]	3
[0.15-0.2]	4	[0.2-0.25]	5	[0.25-0.3]	6
[0.3-0.3 5]	7	[0.35-0.4]	8	[0. 4-0.45]	9
[0.45-0.5]	10	[0.5-0.55]	11	[0.55-0.6]	12
[0.6-0.65]	13	[0.65-0.7]	14	[0.7-0.75]	15
[0.75-0.8]	16	[0.8-0.85]	17	[0.85-0.9]	18
[0.9-0.95]	19	[0.95-0.1]	20		

3.2.4 Association Rule Generation

The cough encoding database is now ready to be scanned for mining association rules. Each feature encoding is subjected to the Apriori algorithm twice, each time with different MinSup and MinCon values. It is used for the first time in the cough encoding of samples labeled with the COVID-19 class to uncover the association rules that characterize the expected behaviors in this class of coughs. The Non-COVID-19 class samples are used for the second time.

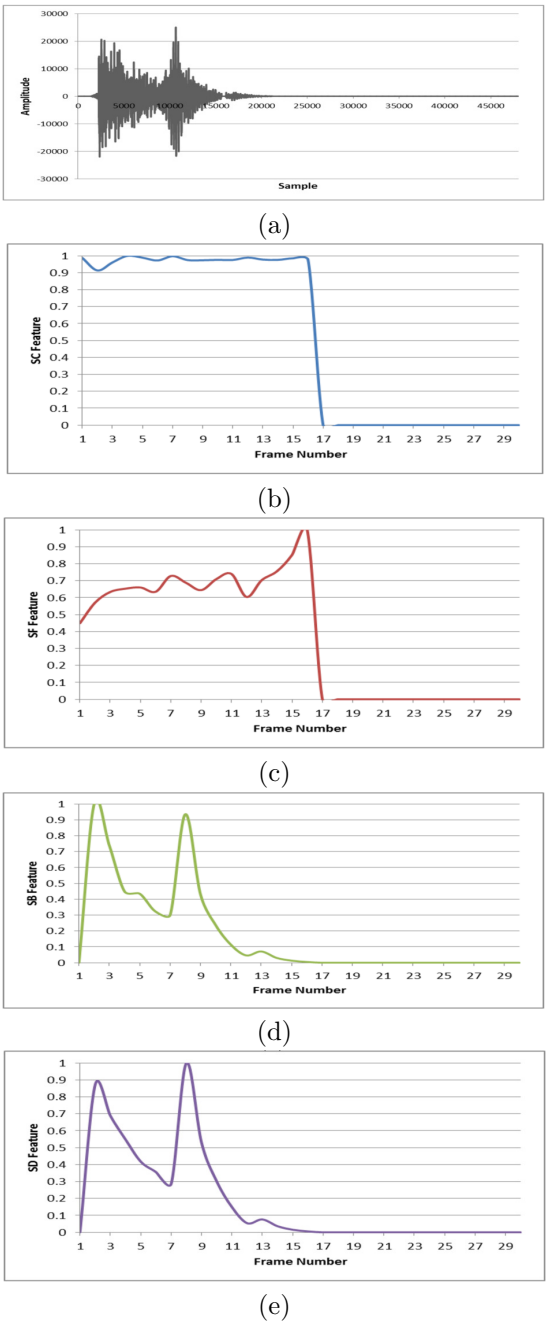
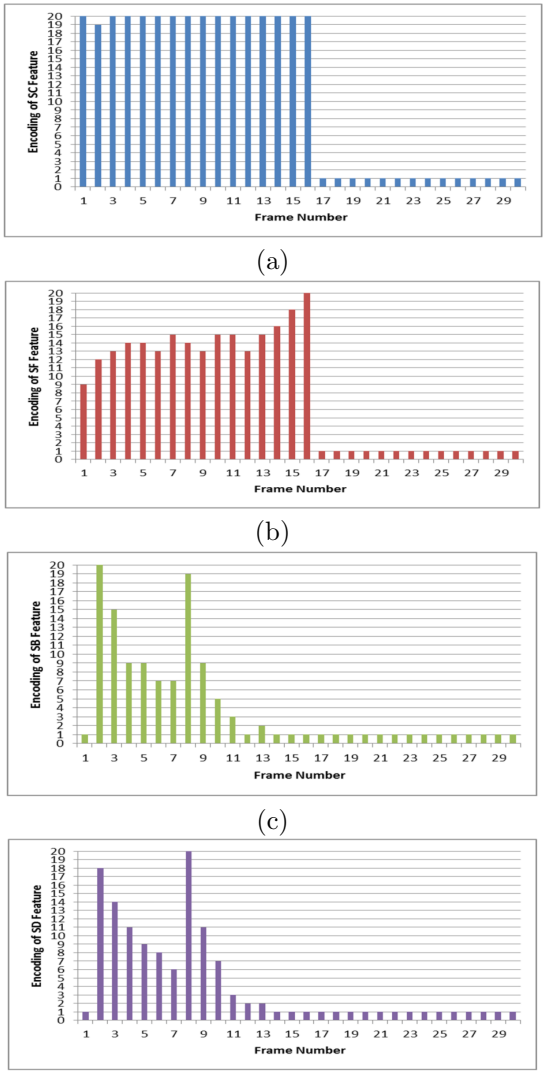


Fig.2: *An example of the extracted spectral features represented as charts, (a) the original cough sound, (b)SC feature, (c)SF feature, (d)SB feature, (e)SD feature.*

The result of this stage is a number of the COVID-19 association rules equal to $NR_{COVID-19}$ rules and Non-COVID-19 association rules equal to $NR_{non-COVID-19}$. Thus, the total length of the final feature vector is $NR_{COVID-19} + NR_{Non-COVID-19}$ which comprises rules-based features. It is worth noting that many transactions will reflect the same cough sound, which will affect classification accuracy since more data about cough behavior will be collected.



(a)

(b)

(c)

(d)

Fig.3: *An example of encoding the extracted spectral features represented as charts, (a)SC feature encoding, (b)SF feature encoding, (c)SB feature encoding, (d)SD feature encoding.*

3.2.5 Association Rule Matching

The binary SVM classifier is used in this work for feature vector classification. Unfortunately, SVM can work only with numerical feature vectors. Thus, the rules-based feature vector must be converted into a numeric feature vector. The binary feature vector is generated for this task through rule matching. Each rule is checked against the cough encoding of each sample to verify whether the rule can be applied to this cough encoding. If this is the case, it is encoded with a 1 in the binary feature vector; otherwise, it is encoded with a 0.

3.2.6 Classification

The final feature vector will have 1s and 0s indicating whether or not each rule applies to the encoding of the cough sound. Finally, the SVM is trained using the binary feature vectors from the training set samples, and the weights obtained are assessed to determine the classification accuracy.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The Virufy dataset is used for model construction and validation purposes. To avoid overfitting during the training process, 70% of samples within each class are used for system training and 30% for testing.

For measuring the effectiveness of the proposed system, the classification accuracy metric is used for this purpose [20]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \qquad (9)$$

TP and TN denote the number of cough segments correctly classified into COVID-19 or Non-COVID-19 classes, respectively, while FP and FN represent the number of sound segments incorrectly classified into the categories as mentioned earlier.

Several experiments were carried out to identify the optimum association rules that might offer the highest accuracy in the identification of COVID-19-infected patients' coughs. First, the efficacy of system parameters ($frmc, MinSup, MinCon$) is discussed, followed by the outcomes of association rules.

4.1 Effectiveness of System Parameters

This experiment aims at measuring the ability of the encoded spectral features to distinguish the infected cough sound with different parameter values. Five $frmc$ values are tried (5, 10, 15, 20, and 25) with different $MinSup$ (in the range [0-1]) and $MinCon$ (in the range [0-1]). Figures 4, 5, 6, and 7 depict the accuracy obtained by mining SC, SF, SB, and SD features, respectively. Tables 2, 3, 4, and 5 give data (minimum accuracy (Min), maximum accuracy (Max), mean of the accuracies (Mean), and standard deviation (STD)) concerning the optimal scenario that obtains the highest accuracy in each feature encoding for the various frame numbers.

As demonstrated in Figs. 4, 5, 6, and 7, when $frmc = 5$, the model performs poorly in all feature types, with the SF feature providing the accuracy (75.207%) when $MinSup = 0.1$ and $MinCon = 0.1$.

However, when $frmc = 10$, a significant improvement in classification accuracy is observed, with all feature types achieving the same best accuracy (90.083%) for $MinSup = 0.8, 0.8, 0.8, 0.8$ and $MinCon = 1, 1, 0.8, 0.9$ for SC, SF, SB, and SD features, respectively.

The model performs the best when frmc=15, with the SC feature providing a classification accuracy of 98.347% when $MinSup = 0.2$ and $MinCon = 0.8$ and the SF feature providing a classification accuracy of

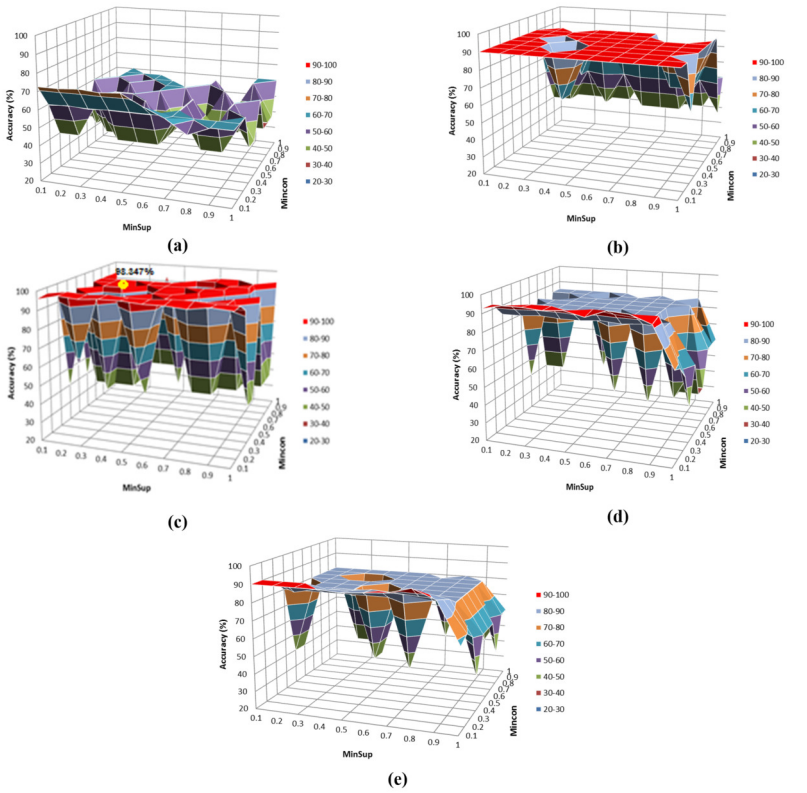


Fig.4: *The effects of various parameter configurations on classification accuracy when encoding SC feature: (a)Frmc=5, (b)Frmc=10, (c)Frmc=15, (d)Frmc=20, (d)Frmc=25.*

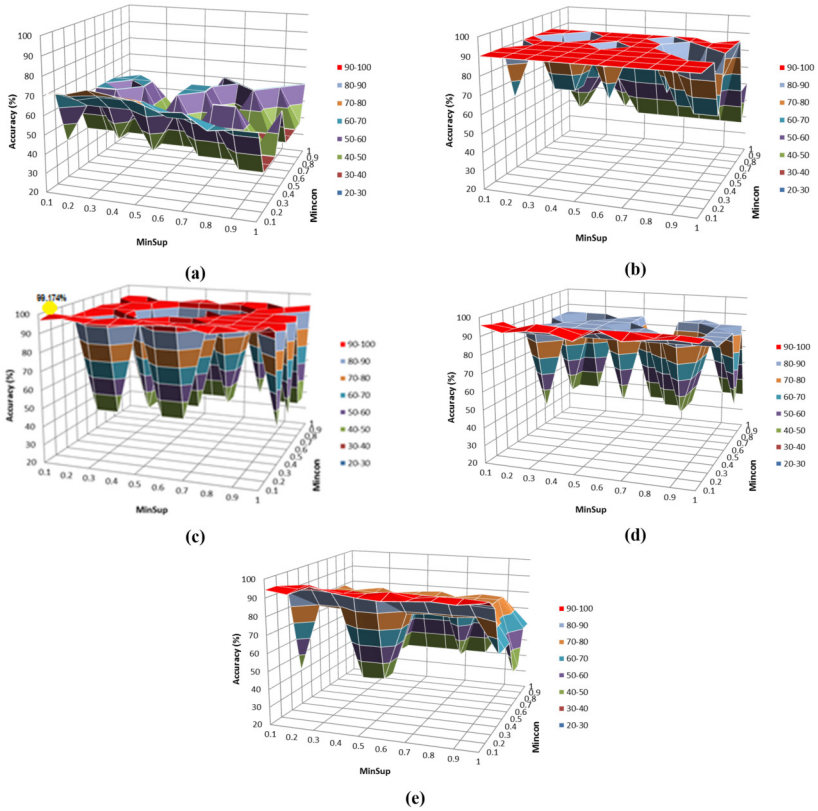


Fig.5: *The effects of various parameter configurations on classification accuracy when encoding SF feature: (a)Frmc=5, (b)Frmc=10, (c)Frmc=15, (d)Frmc=20, (d)Frmc=25.*

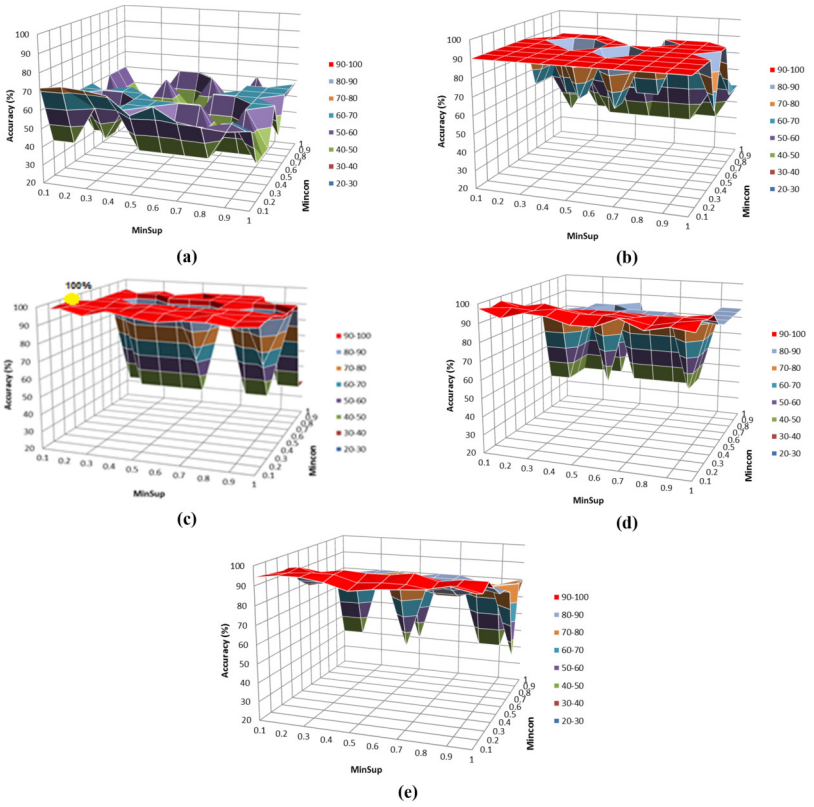


Fig.6: *The effects of various parameter configurations on classification accuracy when encoding SB feature: (a)Frmc=5, (b)Frmc=10, (c)Frmc=15, (d)Frmc=20, (d)Frmc=25.*

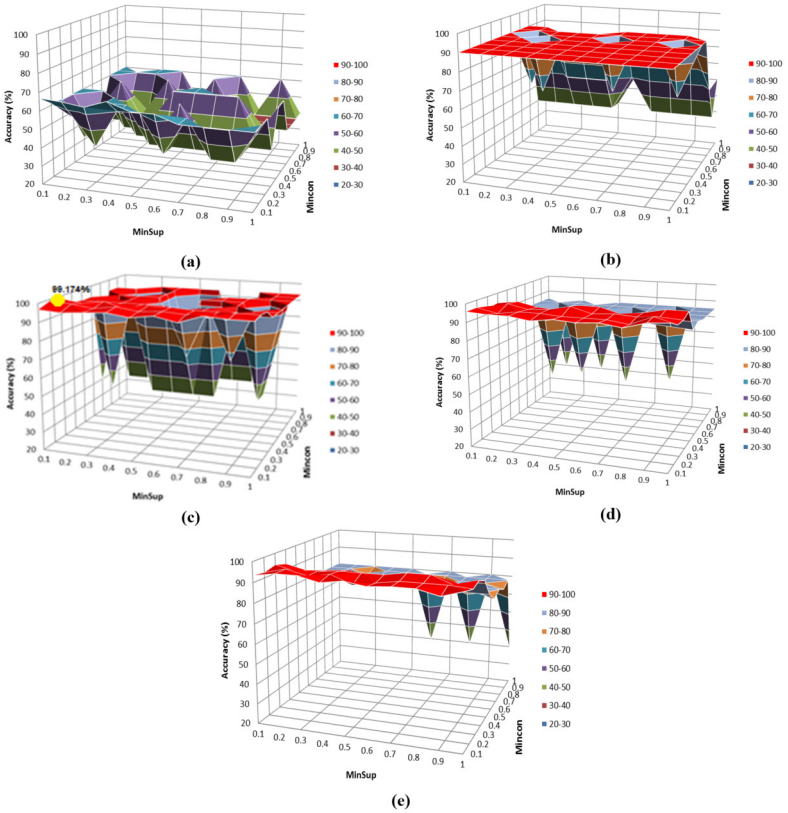


Fig.7: *The effects of various parameter configurations on classification accuracy when encoding SD feature: (a)Frmc=5, (b)Frmc=10, (c)Frmc=15, (d)Frmc=20, (d)Frmc=25.*

99.174% when $MinSup = 0.1$ and $MinCon = 0.1$. At the same time, SB provides the desired accuracy (100%) when $MinSup = 0.1$ and $MinCon = 0.4$, SD provides the maximum accuracy (99.174%) when $MinSup = 0.2$ and $MinCon = 0.1$.

On the other hand, when $MinSup < 0.5$, the model exhibits the best behavior where the best accuracies are reached. This is because more rules with small support values can be generated, and this can convey different patterns of behavior that occur in the coughs among other patients. Similarly, the value $MinCon$ tends to be almost negligible (in the best case) to allow more rules to be generated.

Table 2: *Some statistics about the accuracy achieved when mining SC encoding for different frame numbers.*

Frmc	Min	Max	Mean	STD
5	39.669	73.554	51.662	11.372
10	39.669	90.083	78.058	19.22
15	39.669	98.347	76.314	25.756
20	39.669	96.694	78.306	17.69
25	39.669	90.083	75.179	14.524

Table 3: *Some statistics about the accuracy achieved when mining SF encoding for different frame numbers.*

Frmc	Min	Max	Mean	STD
5	39.669	75.207	51.166	12.159
10	39.669	90.083	77.876	18.697
15	39.669	99.174	78.719	25.004
20	39.669	97.521	78.893	19.344
25	39.669	95.041	73.499	17.513

Table 4: *Some statistics about the accuracy achieved when mining SB encoding for different frame numbers.*

Frmc	Min	Max	Mean	STD
5	39.669	73.554	50.386	11.243
10	39.669	90.083	77.992	18.114
15	39.669	**100**	86.38	20.271
20	39.669	97.521	81.545	19.403
25	39.669	96.694	81.919	13.48

Table 5: *Some statistics about the accuracy achieved when mining SD encoding for different frame numbers.*

Frmc	Min	Max	Mean	STD
5	39.669	68.595	48.935	10.706
10	39.669	90.083	79.066	18.488
15	39.669	99.174	82.752	22.666
20	39.669	97.521	86.488	12.473
25	39.669	95.041	83.186	9.956

4.2 Association Rule Results

Table 6 presents the total number of generated association rules within each feature encoding for the parameters' configuration scenario that achieved the maximum accuracy. The SC and SD features failed to generate any COVID-19 class rules and instead relied on rules generated from non-COVID-19 cough samples. In contrast, the SB feature provides 65 rules that correlate the typical behavior amongst COVID-19 cough segments. SB feature is the most useful spectral feature in capturing COVID-19 cough proprieties. Tables 6, 7, 8, 9, and 10 provide samples of the generated association rules for SC, SF, SB, and SD features, respectively. The support (Sup) and confidence (Con) values for each generated rule are also included in the tables.

Table 6: *The number of association rules generated within each feature encoding.*

Feature Type	Non-COVID19	COVID-19	Total
SC	47	0	47
SF	292	16	47
SB	2382	65	308
SD	152	0	2447

Table 7: *Samples for the association rules were generated using SC encoding with Frmc=15, MinSup=0.2, and MinCon=0.8.*

Association Rules	Sup	Con	Class
{12,20}→{1}	0.2876	1	Non-COVID-19
{8,20}→{1}	0.2054	1	Non-COVID-19
{15}→{20}	0.1917	1	Non-COVID-19
{10}→{20}	0.2328	1	Non-COVID-19
{8}→{20}	0.2054	1	Non-COVID-19

4.3 Comparison With Other Works

Table 11 depicts a comparison made between the proposed work and previous studies. Only works based on the Virfuy datasets are used for evaluation purposes to ensure a fair comparison under the same conditions. The table also lists the method used and the classification accuracy achieved by each work. As indicated in the table, the accuracy attained by the proposed system outperforms the outcomes of previous results. This reflects the effectiveness of the association rules in capturing the correlations implied within the sound segment during patients' coughs.

Table 8: *Samples for the association rules were generated using SF encoding with Frmc=15, MinSup=0.1, and MinCon=0.1.*

Association Rules	Sup	Con	Class
{1,2,3,4}→{20}	0.1232	1	Non-COVID-19
{13,14,20}→{1}	0.1643	1	Non-COVID-19
{16,18}→{20}	0.1232	1	Non-COVID-19
{9}→{20}	0.2739	1	Non-COVID-19
{1}→{20}	1	1	Non-COVID-19
{1,19}→{20}	0.1041	1	COVID-19
{1}→{20}	0.16666	1	COVID-19
{2}→{1,20}	0.1041	1	COVID-19
{19}→{1}	0.1041	1	COVID-19
{2,20}→{1}	0.16666	1	COVID-19

Table 9: *Samples for the association rules were generated using SB encoding with Frmc=15, MinSup=0.1, and MinCon=0.4.*

Association Rules	Sup	Con	Class
{1,2,16,20}→{9}	0.1095	0.6666	Non-COVID-19
{1,16,20}→{2,9}	0.1095	0.4705	Non-COVID-19
{2,9,20}→{1,7}	0.1232	0.45	Non-COVID-19
{7,9}→{1,2,20}	0.1232	0.6	Non-COVID-19
{2,9}→{1,4,20}	0.1506	0.55	Non-COVID-19
{2}→{20}	0.1041	1	COVID-19
{20}→{5}	0.1041	0.625	COVID-19
{1,20}→{11}	0.1041	0.625	COVID-19
{18,20}→{1}	0.1041	1	COVID-19
{1,8}→{20}	0.1041	1	COVID-19

Table 10: *Samples for the association rules were generated using SD encoding with Frmc=15, MinSup=0.2, and MinCon=0.1.*

Association Rules	Sup	Con	Class
{6}→{20}	0.4794	1	Non-COVID-19
{1,2,6}→{20}	0.3013	1	Non-COVID-19
{1,3,6}→{20}	0.2054	1	Non-COVID-19
{16}→{1}	0.2191	1	Non-COVID-19
{12}→{20}	0.3287	1	Non-COVID-19

Table 11: *Comparison between the proposed system and previous works .*

Authors	Ref.	Method	Classification Accuracy
Solak (2021)	[6]	MDFA, LZC, entropy, and SVM	95.8%
Tena et al. (2022)	[7]	WD and RF	94.81%
Manshouri (2022)	[3]	MFCCs and SVM	95.86%
Chowdhury et al. (2022)	[8]	MCDM	89%
The proposed system	/	Association rules of SB feature encoding and SVM	100%

5. CONCLUSIONS

The long-term presence of the Corona pandemic has heightened the need for developing early detection techniques. Cough sounds were employed in this study to establish an effective COVID-19 infection detection system. The system's ability to differentiate between COVID-19-induced and non-COVID-19-induced coughs was enhanced by disclosing the correlations between cough sound frames. The proposed cough-based COVID-19 diagnosis system can be further expanded to be used as a mobile application. Thus, it can be installed on smartphones and used as an instance diagnosis tool by everyone anywhere. However, the system works on infection diagnosis without taking the patient's age into account. The acoustic parameters differ concerning the patient's age. Thus, the system can be improved by first identifying the patient's age group and then diagnosing the infection status based on that age. The same is true for the patient's gender (i.e., male or female).

AUTHOR CONTRIBUTIONS

Conceptualization; methodology; software.; validation; formal analysis; investigation; data curation; writing—original draft preparation; writing—review and editing; visualization; supervision; funding acquisition, S. N. Mohammed. All authors have read and agreed to the published version of the manuscript.

References

[1] S. Mohammed, F. Alkinani, and Y. Hassan, "Automatic Computer Aided Diagnostic for COVID-19 Based on Chest X-Ray Image and Particle Swarm Intelligence," *International Journal of Intelligent Engineering and Systems*, vol.13, No.5, pp. 63–73, 2020.

[2] COVID-19 Data Explorer ,visited: 23/2/2022 `https://ourworldindata.org/explorers/coronavirus-data-explorer`.

[3] M. Manshouri, "Identifying COVID-19 by Using Spectral Analysis of Cough Recordings: A Distinctive Classification Study," *Cognitive neurodynamics*, vol. 16, no. 1, pp. 239–253, 2022.

[4] World Wild Health Organization, Visited: 23/2/22022, `https://www.who.int/`

[5] M. Manshouri, "Diagnosis Of COVID-19 and Non-COVID-19 Patients by Classifying Only A Single Cough Sound," *Neural Comput & Applic*, vol. 33, pp. 17621–17632, 2021.

[6] F. Solak, "Identification of COVID-19 from Cough Sounds Using Non-Linear Analysis and Machine Learning," *European Journal of Science and Technology*, Special Issue 28, pp. 710-716, 2021.

[7] A. Tena, F. Clarià, and F. Solsona, "Automated Detection of COVID-19 Cough," *Biomedical Signal Processing and Control*, vol. 71, pp. 1-11, 2022.

[8] N. Chowdhury, M. Kabir, M, Rahman, and S. Islam, "Machine Learning for Detecting COVID-19 from Cough Sounds: An Ensemble-Based MCDM Method," *Comput Biol Med.*, vol. 145, 2022.

[9] P. Tan, V. Kumar, M. Steinbach, and A. Karpatne, *Introduction to Data Mining*, 2nd edition, Pearson Education, 2019.

[10] H. Jiawei, and K. Micheline, *Data Mining: Concepts and Techniques*, 2nd edition, Elsevier, ISBN: 978-1-55860-901-3, 2006.

[11] A. Ana, *Data Mining and Knowledge Discovery in Databases*, 4th edition, IGI Global, 2018.

[12] S. Mohammed, and H. Rada, "English Numbers Recognition Based On Sign Language Using Line-Slope Features And PSO-DBN Optimization Method," *Journal of Engineering Science and Technology*, vol. 15, no. 3, pp. 1855 - 1867, 2020.

[13] S. Snehal, and D. Navnath, "Survey: Support Vector Machine and Its Deviations in Classification Techniques," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 4, no. 12, pp. 993-997, 2014.

[14] C. Chaudhari, X. Jiang, A. Fakhry, A. Han, j. Xiao, S. Shen, and A. Khanzada, "Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough," *ArXiv, 2020. 2011.13320*, 2020.

[15] Virufy-covid, visited : 1/12/2021, https://github.com/virufy/virufy-covid

[16] S. Mohammed, and A. Hassan, "Automatic Voice Activity Detection Using Fuzzy-Neuro Classifier," *Journal of Engineering Science and Technology*, vol. 15, no. 5, pp. 2854 – 2870, 2020.

[17] Y. Hussain, and S. Mohammed, "Intelligent System for Parasitized Malaria Infection Detection Using Local Descriptors," *International Journal of Intelligent Engineering and Systems*, vol.14, no.1, pp. 296-305, 2021.

[18] T. Giannakopoulos, and A. Pikrakis, *Chapter 4 - Audio Features*, Editor(s): T. Giannakopoulos, A. Pikrakis, Introduction to Audio Analysis, Academic Press, pp. 59-103, 2014.

[19] M. Ben Nasr, S. Ben Jebara, S. Otis, B. Abdulrazak, and N. Mezghani, "Spectral-Based Approach for BCG Signal Content Classification," *Sensors*, vol. 21, no. 1020, 2021.

[20] S. Mohammed, A. Jabir, and Z. Abbas, "Spin-Image Descriptors for Text-Independent Speaker Recognition," in *Proc. of Saeed F., Mohammed F., Gazem N. (eds) Emerging Trends in Intelligent Computing and Informatics. IRICT 2019. Advances in Intelligent Systems and Computing*, vol. 1073, Springer, Cham, 2019.

Suhaila N. Mohammed received a BSc degree (2010) and an MSc degree (2016) in computer science from the University of Baghdad, Iraq. She received her Ph.D. degree (2020) in computer science from the University of Technology, Iraq. Currently, she is working as a teaching staff member in the computer science department at the college of science, University of Baghdad, Iraq. Her research interests include computer vision, artificial intelligence, data mining, multimedia, image processing, signal processing, machine learning, and soft computing. Email: suhailan.mo@sc.uobaghdad.edu.iq