ISSN: 0067-2904

A New COVID-19 Patient Detection Strategy Based on Hidden Naïve Bayes Classifier

Zainab Haider Ameen[1], Nadia F. AL-Bakri[1], Azhar F. Al-zubidi[1], Soukaena Hassan Hashim[2], Zahraa A. Jaaz[1]

[1]*Computer Science Department, College of Sciences, AL Nahrain University, Jadriya, Baghdad, Iraq*
[2]*Computer Sciences Department, University of Technology, Baghdad 10066, Iraq*

Received: 23/10/2023 Accepted: 10/6/2024 Published:30/**11**/2024

Abstract

COVID-19 is a universal infectious disease recognized first by people with influenza and bacterial pneumonia symptoms in Wuhan, Hubei Province. Currently, a new mutated disease has the same symptoms as COVID-19 and influenza and causes dangerous infections in the body. Due to the fact that these two diseases share some diagnostic features and symptoms in common with one another, healthcare workforces require aid and support in predicting patients' conditions. This was done by using machine learning methods in diagnosis. From this point, this paper proposes a diagnostic model to detect patients' symptoms and classify them into one of five disease groups, utilizing Neighborhood Component Analysis (NCA) as a feature selection method and the Hidden Naïve Bayes (HNB) method as a multiclass classifier. This paper suggests the model consists of two significant phases: the pre-processing phase (cleaning, normalization, and discretization) and the classification phase. Conducting the COVID-19 dataset, the experimental findings showed that the suggested multi-class model had 89% accuracy for disease diagnoses. Furthermore, according to the patient's symptoms, the proposed classification model led to a good diagnosis for the mutated COVID-19 disease.

Keywords: Classification, COVID-19, Feature Selection, Flu, Hidden Naïve Bayes.

استراتيجية جديدة للكشف عن مرضى كوفيد-19 بناءً على مُصنف . *Hidden Naïve Bayes*

زينب حيدر أمين[1], نادية فاضل البكري[*1], أزهار فليح الزبيدي[1], سكينة حسن هاشم[2], زهراء عبد الحسين جعاز[1]

[1], قسم الحاسوب, كلية العلوم, جامعة النهرين,الجادرية, بغداد, العراق

[2] قسم الحاسوب, كلية العلوم, الجامعة التكنولوجية, بغداد, العراق

الخلاصة

ان كوفيد-19 مرضا معديا عالميا تم التعرف عليه أولاً من قبل اشخاص مصابين بأعراض الأنفلونزا والالتهاب الرئوي الجرثومي في ووهان بمقاطعة هوبي. في الوقت الحالي، هناك مرض متحور جديد له نفس أعراض فيروس كورونا (COVID-19) والأنفلونزا ويسبب التهابات خطيرة على الصحة. ونظرًا لأن هذين

*Email: nadia.f.al-bakri@nahrainuniv.edu.iq

المرضين يشتركان في بعض السمات التشخيصية والأعراض المشتركة مع بعضهما البعض، فقد احتاجت القوى العاملة في مجال الرعاية الصحية إلى المساعدة والدعم في التنبؤ بظروف المرضى. وتمت هذه المساعدة باستعمال أساليب التعلم الآلي في التشخيص. من هذه النقطة، تقترح هذه الورقة نموذجًا تشخيصيًا للكشف عن أعراض المرضى وتصنيفهم إلى واحدة من خمس مجموعات مرضية باستعمال أختيار الميزة بطريقة NCA و بأستعمال المصنف متعدد الطبقات Hidden Naïve Bayes.

يتكون النموذج المقترح من مرحلتين مهمتين: مرحلة المعالجة المسبقة (التنظيف والتطبيع والتمييز) ومرحلة التصنيف. من خلال إجراء الاختبار على مجموعة بيانات كوفيد-19، أظهرت النتائج التجريبية أن النموذج متعدد الفئات المقترح يتمتع بدقة 89٪ للتنبؤ. علاوة على ذلك، ووفقًا لأعراض المريض، أدى نموذج التصنيف المقترح إلى تشخيص جيد لمرض كوفيد-19 المتحور.

1. Introduction

Data mining (DM) has gained popularity in the commercial and healthcare sectors as a synthetic intelligence tool. It is also known as fact gathering, data/pattern analysis, knowledge extraction, and discovery. The development in health care is due to the many varied uses of smart devices and the creation of intelligent systems for diagnosis in the early stages. Hospitals gather much data in the form of electronic health records (EHR), which contain patient information. Expert systems based on machine learning and decision support systems (DSSs) are currently needed in health and medical applications that rely on the early diagnosis of disorders. Most of these algorithms, including Hidden Naive Bayes (HNB), an extension of Naive Bayes, deal with classifications established based on the kind of processed data [1]. A severe respiratory syndrome, the coronavirus illness (COVID-19), which was initially identified in Chinese Wuhan City in 2019 and resulted in serious community health and social economic turmoil, is initiated by the SARS coronavirus. Some of the primary symptoms of COVID-19 are breathing difficulties, fever, coughing, and a sore throat [2]. Numerous classification algorithms have been used for illness datasets to diagnose chronic diseases, and the results are extremely promising [3]. Nave Bayes (NB) and Hidden Naïve Bayes (HNB) classifiers are probabilistic classifiers that predict a class based on membership probability. It is among the most effective classification techniques because it analyzes the correlation between the independent and dependent variables to calculate conditional probability. In the last two decades, several studies have been done to lessen the independence assumption of the NB classifier. The HNB classifier, based on creating an additional layer representing a hidden parent for each feature, was first developed in one of these investigations [1]. In terms of attribute interdependence, Compared to naive Bayes, hidden naive Bayes is a more accurate classification method. A Bayesian classifier called HNB avoids unsolvable complexity and considers the impact of all features [4].

In this work, an algorithm for classification HNB is utilized to predict whether a patient is thought to have one of the five mutation-related diseases (asymptomatic, mild COVID-19, severe COVID-19, flu, or normal), according to the patient's symptoms. The rest of this article is structured as follows: Section 2 describes the nature of the problem definition and prime objective. Then, the related work that has been highlighted is presented properly in Section 3. In Section 4, the machine learning definitions are summed up. Next, Section 5 explains feature selection categorizations. Then, Section 6 describes the proposed model. Section 7 presented a discussion and the results from the suggested model, while Section 8 concluded the discussion.

2. Problem Definition and Main Objective

Because of the wide spread of infected people with similar symptoms and diseases, early illness diagnosis is crucial for treating and managing the conditions. The healthcare system

may now be protected from overload by quickly identifying and isolating affected individuals, which will flatten the epidemic curve. The main task is to help healthcare organizations detect the disease cases of tested patients in their early stages to help reveal to what level the similarities between COVID-19 and flu infections are and to decrease the side effects of both. According to the fast mutation of COVID-19, accurate disease detection will be more helpful during the pandemic and assist in developing promising approaches and taking creative decisions.

3. Related Work

The medical community, particularly the World Health Organization, is under pressure as the disease spreads rapidly over the globe. As a result, extensive research is underway to determine how to diagnose people with a mutated low-risk COVID-19 infection based on symptoms similar to those of other diseases. Widespread interest has developed in studying many types of research related to coronaviruses with machine learning algorithms [5]. There are numerous related works about COVID-19 with machine learning. In this section, some of them are described as the following:

The research work in [4] developed the HNB algorithm to classify and forecast cardiac disease. The suggested method used discretization and IQR filters for improvement. The experimental results had the best accuracy of 100 percent compared to the NB classification model. It was determined that the suggested approach, which used the HNB model, helped create a reliable decision support system for disease diagnosis. This research [5] developed machine learning strategies using multilayer perceptron, Bayes network, naive Bayes, and locally weighted learning algorithms for classifying patient symptoms into H1N1 and COVID-19 classifications. Bayes network had an accuracy of 86.57 percent, while the accuracy of the NB was 82.34 percent. The multilayer perceptron accuracy was 99.31 percent, while the locally weighted learning method had an accuracy of 88.89 percent, and the random forest had an accuracy of 83.16 percent. Another research work in [6] consumed an epidemiology COVID-19 dataset from patients from South Korea. To forecast patient recovery, they applied decision trees, logistic regression, support vector machines, naive Bayes, random forests, and K-nearest neighbor techniques. The results indicate that the decision tree approach model has an overall accuracy of 99.85 percent in predicting whether or not infected people will recover from the COVID-19 pandemic. One more study on the COVID-19 dataset in India [7] explained the concept of SVM (support vector machine) and cross-validation. The results show that support vector machine SVM produced accurate results in classifying data on recovered and deceased patients, with an accuracy of 99 percent and a precision of 98 percent, while a recall of 95 percent in the performance measure. In additional research work in [8], Hidden Naive Bayes and Ada-Boost algorithms were used. They employed a discretization approach instead of replacing missing values. Also, the feature selection method is utilized to remove unnecessary features in each iteration of Ada-Boost. The suggested technique required less training time and provided a greater chance of scaling to data mining applications. In one more research work [9], reports for textual clinical were divided into four groups using ensemble learning techniques. Three feature extraction methods were used: (TF/IDF) term frequency/inverse document frequency, (BOW) bag of words, and report length. The results demonstrated that logistic regression and the multinomial Naïve Bayes algorithm outperformed all other algorithms, with a precision of 94 percent, recall of 96 percent, F1 score of 95 percent, and accuracy of 96.2 percent. At the same time, random forest and gradient boosting also performed well, with an accuracy of 94.3 percent and 94.3 percent, respectively. Furthermore, the study [10] outlined a novel coronavirus detection method using a Naïve Bayes classifier by calculating all individual probabilities that could be applied to the coronavirus attribute that is the target, including all

possible probabilities. As a result, three groups of coronaviruses were divided. Mammalian coronaviruses fall under groups 1 and 2, and avian coronaviruses fall under group 3. Also, different classical methods were used in coronavirus diagnosis, including RT-PCR and ORF1ab. The research [11] described and categorized pandemics using two algorithms: the fuzzy C-mean (FCM) clustering technique and the back propagation (BP) classification algorithm. Two phases made up the implementation: first, the FCM algorithm determines the type of virus. Then, use BP as a classifier to determine the pandemic class. And for optimization, the associated features that improve the model to enhance accuracy are used in information gain (IG). The system's accuracy was up to 0.79 percent using the back propagation method. In additional research work in [12], an improved algorithm—the double-hidden naive Bayes algorithm—is proposed to make full use of the dependency relationship between attributes. Experimental results show that this classification algorithm based on improved TF-IDF and double-hidden naive Bayes can improve the speed, accuracy, and recall rate of classification results.

4. Machine Learning Categorization

1. The main areas within machine learning methods are:
2.
3. Supervised Learning: takes train-set input variables with previously identified labels. A specific classifier is utilized for learning the map function from input to output, also known as the one-label classification. Some of the classification methods are: naïve Bayes, hidden naïve Bayes and support vector machines [13] and [14].
4. Unsupervised learning takes input variables from training datasets with no previously known labels, such as clustering methods. [15] [16].

4.1 Hidden Naïve Bayes Classifier

Over the past two decades, much research has been done on the objectivity theory of the Naïve Bayes NB classifier. As shown in Figure 1, the HNB classifier, proposed in one of the studies, is a novel classifier that is based on building a second layer that acts as a hidden parent for each attribute, which is represented as a dashed circle in the figure. The advantage of employing hidden parents (A_{hpi}) is the combination of the weighted inspirations from all other attributes (A_i) where i, j = 1, 2... N and ($j \neq i$). Prob(C) is the probability of class. The definition of the joint distribution is given by Eq. (1). The hidden parent is illustrated in Eq. (2), and the HNB classifier is formulated in Eq. (3) [14, 17].

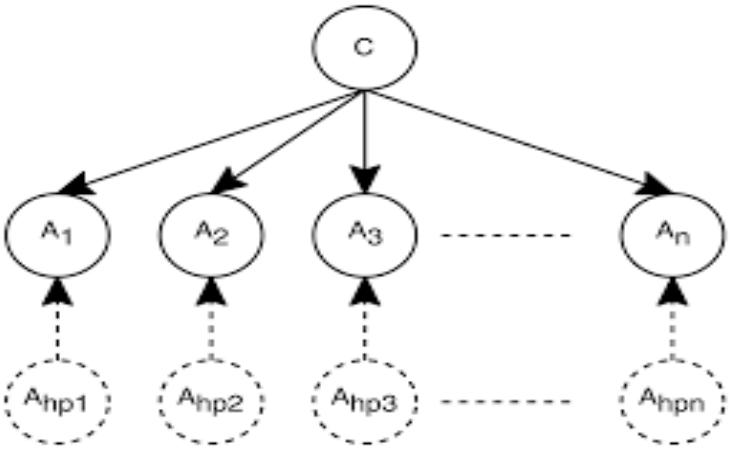

Figure 1: Hidden Naïve Bayes Structure [17]

$$Prob(A_{1,\dots,}A_N|\mathrm{C}) = Prob(\mathrm{C}) \prod_{i=1}^{N} Prob(A_i|A_{hpi,}\, C) \quad (1)$$

Where

$$Prob\big(A_i\big|A_{hpi}, C\big) = \sum_{j=1,j\neq i}^{n} W_{ij} * Prob(A_i|A_j, C) \quad (2)$$

And $\sum_{j=1,j\neq i}^{N} W_{ij} = 1$. the hidden parent A_{hpi} for A_i is a combination effect from all other weighted attributes. Then the classifier for the hidden Naïve Bayes on an event E= {$a_1 \dots a_n$} is shown as:

$$c(E) = \arg\max_{c\in C} Prob(c) \prod_{i=1}^{n} Prob(a_i|a_{hpi}, c) \quad (3)$$

Using conditional mutual information (CMI) between each pair of features (A_i) and (A_j), the weights W_{ij} are computed initially. The formulas Eq. (4), Eq. (5) depicts the CMI formulation [14, 17].

$$W_{ij} = \frac{CMI\ (A_i;A_j|C)}{\sum_{j=1,j\neq i}^{n} CMI(A_i;A_j|C)} \quad (4)$$

Where

$$CMI\big(A_i; A_j\big|C\big) = \sum_{a_i,a_j,c} Prob\big(a_i, a_j, c\big) log \frac{Prob(a_i,a_j|c)}{Prob(a_i|c)P(a_j|c)} \quad (5)$$

By estimating the parameters using the training data, it is simple to integrate the effects of all the other variables using CMI information. However, estimating the parameters in HNB using the training data is a key step in model learning [14].

5. Feature Selection

Before using a learning system, feature selection is a crucial data processing step. To increase the effectiveness of the classification and reduce the amount of data stored in memory, it is required to identify a subset of important features from the original features. It helps lessen the effects of the curse of dimensionality, reduces processing requirements, increases learning accuracy, and helps to pinpoint which characteristics could be relevant to a given issue [18].

5.1 Feature Selection Methods

They are classified into unsupervised, supervised, and semi-supervised according to the dataset (label or un-label) as follows: [18]

- Unsupervised: A feature selection method that depends on clustering quality metrics and might result in several equally valid feature subsets is a less confined search issue without class labels.
- Supervised: A feature selection method that can be divided into embedded models, filter models, and wrapper models. Firstly, the filter model separates selecting features from classifier learning, preventing one algorithm's bias from influencing the other's prejudice. It is based on the training data's general properties, including distance and correlation. The most representative filter model techniques are information gain, Fisher score, and relief approaches. Secondly, wrapper models employ a chosen learning algorithm's predicted accuracy to assess the features' quality. Unfortunately, the cost of implementing these algorithms for data with a lot of characteristics is high. Lastly, the embedded model fills the space left by the wrapper and filter models. As with the filter model, it begins by incorporating the statistical criteria to choose potential feature subsets with a specific cardinality. The subgroup with the best classification accuracy is then selected.
- Semi-supervised: A feature selection method used when both labeled and unlabeled data are used to provide relevant estimates.

5.2 Neighborhood Component Analysis (NCA)

This method is used to choose features to increase the predictive power of classification and regression algorithms. It's non-parametric and embedded method. Regularization of feature weight learning aims to minimize the mean (leave-one-out) loss classification over the train set of data by minimizing the objective function [8].

6. General Design of the Proposed System

The present medical analysis reveals identical symptoms between the flu and COVID-19 (mild or severe). Therefore, a machine learning algorithm is needed to classify patients according to their symptoms (asymptomatic information). In this section, the essential phases of the suggested model are presented. Utilizing the COVID-19 dataset, the proposed model will reveal whether the patient is affected by: 1-asymptomatic (class 1), 2-mild COVID-19 (class 2), 3-severe COVID-19 (class 3), 4-flu (class 4), and 5-normal (class 5). Accordingly, the model is composed of two phases: (1) The pre-processing phase. (2) Classification phase. Figure 2 shows our proposed model's general structure, which is made up of the following steps:

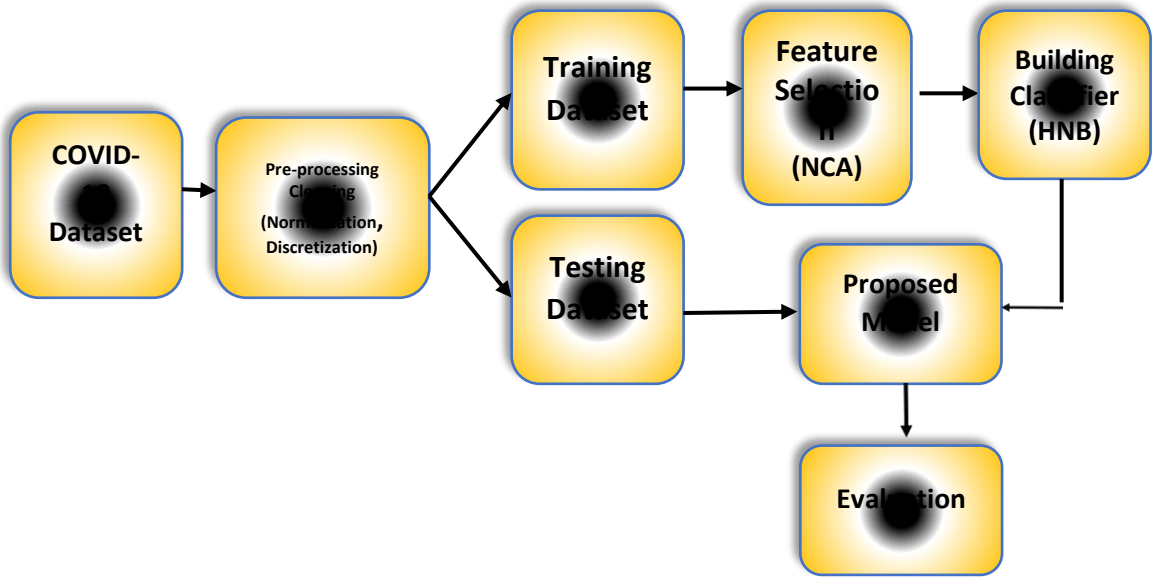

Figure 2: The Proposed Model Block Diagram

6.1 Pre-Processing Stage

During this phase, a transformation is performed to convert the incomplete and inadequately extracted raw data into quality data and provide a useful data pattern for the proposed models. The COVID-19 dataset was loaded and conducted from the website https://covid19-influenza-response.cells.ucsc.edu for 59,570 patient records with 26 data columns (not considering the label column) for patients as shown in Table 1 (a, b, and c). The features are: Cell Id, Sample ID, Patient ID, Disease group, Comorbidity, Hospital-day, WBC/micro-L, Neutrophil/micro-L (%), Lymphocyte/micro-L (%), Monocyte/micro-L (%), C-reactive protein (mg/ dl), Chest X-ray, Treatment, Respiratory rate (BPM), O2 Saturation, O2 Supplement, Temperature (Temp.), Systolic BP, Heart rate, Consciousness, Score for NEWS, Severity, Cell type, UMI number, Gene number, and Percentage of mitochondrial gene. Three essential pre-processing steps are performed. They are:

1. Cleaning data: in this step, data cleaning routines are done to clean the data by removing the columns that do not affect decision making, such as Patient Id and Sample Id. Also, columns with many no values (NAN), such as Lymphocyte/micro-L, Monocyte/micro-L, and C-reactive protein (mg/dL), are ignored.

Table 1: a- The First 10 Features of COVID-19 Dataset

Cell Id	Sample ID	Patient ID	Disease group	Comorbidity	Hospital day	WBC/ Micro-L	Neutrophil/ Micro-L (%)	Lymphocyte/ Micro-L (%)	Monocyte /micro-L (%)
AAACCCAAG GGCAATC-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACCCACA GCTGAAG-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACCCAGT CTTCGAA-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACCCAGT TCCGCTT-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACGAAAG GGAGGTG-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACGAACA AGGTTGG-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACGAAGT CTACAAC-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACGCTAG GCTTAAA-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)
AAACGCTC ATGACCCG-1	nCoV-1	C1	severe COVID-19	none	16	21540	19235 (89.3)	1055 (4.9)	1271 (5.9)

Table 1: b- The Next Features of COVID-19 Dataset

Cell Id	C-reactive protein	Chest X-ray	Treatment	Respiratory rate	O2 saturation	O2 Supplement	Temp.	Systolic BP
AAACCCAAG GGCAATC-1	7.58	Pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACCCACA GCTGAAG-1	7.58	Pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACCCAGT CTTCGAA-1	7.58	Pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACCCAGT TCCGCTT-1	7.58	Pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACGAAAG GGAGGTG-1	7.58	Pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACGAACA AGGTTGG-1	7.58	pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACGAAGT CTACAAC-1	7.58	pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACGCTAG GCTTAAA-1	7.58	pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92
AAACGCTC ATGACCCG-1	7.58	pneumonia	nucleoside/ritonavir, hydroxychloroquine, anticoagulant	24	90	ECMO + MV	37.6	92

Table 1: c- The Last Continuing Features of COVID-19 Dataset with Label Column

Cell Id	Heart rate	Consciousness	NEWS score	Severity	Cell-type	Number of UMI	Number of Gene	Percentage of mitochondrial gene	Disease condition
AAACCCAAGGGCAATC-1	122	Unresponsive	14	severe	B cell, IgG+	15679	3311	7.296384	COVID-19 (Severe)
AAACCCACAGCTGAAG-1	122	Unresponsive	14	severe	B cell, IgG-	3630	1146	13.16804	COVID-19 (Severe)
AAACCCAGTCTTCGAA-1	122	Unresponsive	14	severe	CD4, EM-like	6796	1785	3.546204	COVID-19 (Severe)
AAACCCAGTTCCGCTT-1	122	Unresponsive	14	severe	classical Monocyte	6803	2147	7.393797	COVID-19 (Severe)
AAACGAAAGGGAGGTG-1	122	Unresponsive	14	severe	Platelet	819	332	2.319902	COVID-19 (Severe)
AAACGAACAAGGTTGG-1	122	Unresponsive	14	severe	Platelet	1188	488	5.30303	COVID-19 (Severe)
AAACGAAGTCTACAAC-1	122	Unresponsive	14	severe	NK cell	5444	1958	5.180015	COVID-19 (Severe)
AAACGCTAGGCTTAAA-1	122	Unresponsive	14	severe	classical Monocyte	11649	2946	9.22826	COVID-19 (Severe)

2. Normalization: In this step, the data is normalized to provide better results for the next step. The scaled information falls within a smaller range [0.0 to 1.0]. The normalization equation given by Eq. (6) is applied to the continuous features as follows:

$$\text{Value X} = \frac{\text{ValueX in } F_i - \text{Minimum value for feature } F_i}{\text{Maximum value for feature } F_i - \text{Minimum value for feature } F_i} \quad (6)$$

Where X is value in feature F_i

3. Discretization: In this step, the continuous features of the dataset are converted to discrete forms with a finite number of values. The conversion is to increase the speed and accuracy of the model. Moreover, HNB cannot handle new untrained continuous values during the testing phase. The Entropy Minimization Discretization method is utilized for conversion. After our calculation using MATLAB, the normalized and discretized values for each attribute in the dataset are shown in Table 2. Also, Table 3 depicts our computed intervals for each attribute using the discretization process.

Table 2: Sample of COVID-19 Dataset after Pre-processing Stage

WBC Micro-L	Neutrophil	Chest X-ray	Treatment	Respiratory Rate (BPM)	O2 Saturation	O2 supplement	Temp.	Systolic BP	Heart rate (BPM)	Consciousness
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5
0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0.6-0.9	0	0	0.6-0.9	0	1	0.5

Table 3: The Discretized Features Formulation using Discretization Process

Feature No.	Feature name	Values range	Feature No.	Feature name	Values range	Feature No.	Feature name	Values range
1	Disease group	0 0.25 0.5 0.75 1	*7*	Treatment	0 0.1-0.5 0.6-0.9 1	*13*	Heart rate (BPM)	0 0.1-0.5 0.6-0.7 1
2	Comorbidity	0 0.1-0.5 0.6-0.9 1	*8*	Respiratory rate (BPM)	0 0.1-0.5 0.6-0.9 1	*14*	Consciousness	0 0.5 1
3	Hospital-day	0.1-0.8	*9*	O2 saturation	0 0.6-0.9 1	*15*	NEWS score	0 0.1-0.5 0.6-0.7 1
4	WBC/micro-L	0 0.1-0.5 0.6-0.9 1	*10*	O2 supplement	0 0.3-0.7 1	*16*	Severity	0 0.5 1
5	Neutrophil	0 0.1-0.5 0.6-0.9 1	*11*	Temperature	0 0.1-0.5 0.6-0.9 1	*17*	Cell-type	0 0.1-0.5 0.6-0.9 1
6	Chest X-ray	0 0.1-0.5 0.6-0.9 1	*12*	Systolic BP	0 0.1-0.5 0.6-0.9 1			

6.2 Classification Process

After the pre-processing steps, the data is suitable for applying the feature selection method to select the most relevant features that are most effective in the classification process. The dataset is partitioned into a 2/3 training set and a 1/3 testing set. Subsequently, the training set is forwarded to the Neighborhood Component Analysis (NCA) feature selection algorithm. Table 4 lists the 17 features out of 20 and their weights. Table 5 shows a sample of values for the 17 selected features

Table 4: The Resulted Selected Feature's Names and Their Weights using NCA Algorithm

Feature No.	Feature Name	Weight	Feature No.	Feature Name	Weight
1	Disease group	3.6454	*10*	O2 supplement	1.482561
2	Comorbidity	0.626398	*11*	Temperature	1.171667
3	Hospital-day	0.676937	*12*	Systolic BP	0.54072
4	WBC/micro-L	0.218827	*13*	Heart rate (BPM)	0.377473
5	Neutrophil	0.703579	*14*	Consciousness	0.650335
6	Chest X-ray	0.878778	*15*	NEWS score	0.49488
7	Treatment	0.937299	*16*	Severity	3.191692
8	Respiratory rate BPM	1.171097	*17*	Cell-type	0.036733
9	O2 saturation	0.256953			

Table 5: The COVID-19 Dataset after Feature Selection Process

Disease group	Comorbidity	Hospital-day	WBC/ micro-L	Neutrophil	Chest X-ray	Treatment	Respiratory rate	O2 saturation	O2 supplement	Temperature	Systolic BP	Heart rate (BPM)	Consciousness	NEWS score	Severity	Cell-type
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.07
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.14
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.79
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.79
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.64
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.64
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.64
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.29
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.79
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.14
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.29
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.36
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.79
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.64
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0-	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.43
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.36
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.07
0.75	0.9	1	0.96	0.92	0.8	0.57	0.69	0	0	0.7	0	1	0.5	1	1	0.14

The resulting output from the previous step is split into training and testing sets and forwarded to the Hidden Naïve Bayes (HNB) multiclass algorithm. Accordingly, the training records are used to build the HNB classifier, and the testing records are used to classify the test patients's data into the designated five classes. To implement HNB, a calculation is done to compute the probability for each class (label) of the dataset as given by Eq. (7):

$$\text{Probability (Class)} = \frac{\text{class Frequency in training set of covid}-19 \text{ dataset}}{\text{Total number of records in training set of covid}-19 \text{ dataset}} \quad (7)$$

Table 6 shows our calculations for each class's probability and their frequency. As it is shown, COVID-19 mild has the highest probability, which is considered the common class in the dataset.

Table 6: Frequency and Probability for each Class label

Class Type		**Frequency of Class in Training Dataset**	**Probability of Class in Training Dataset**
1	Asymptomatic	4425	0.10612
2	COVID-19(Mild)	16742	0.40149
3	COVID-19(Sever)	10296	0.24691
4	Flu	6226	0.14930
5	Normal	4011	0.09619
		Sum =41700	**Sum** =1

After that, the probability of each value in each interval of the 17 selected features for all classes is computed. We performed this by finding the frequency and the probability of each value in the interval for each of the selected features, as shown in Table 7.
Based on the previous resulted computation, a conditional mutual information CMI is computed, which is the sum of the conditional mutual information probability for each pair of features F_i, F_j in each class C for $i, j = 1\ to\ 17$ in the training set.

As a result, we find the conditional mutual information probability CMIP first for each interval, as shown in Table 8. This is obtained by computing the following requirements:
- Calculating the probability (F_i, F_j, C) by dividing Frequency of seemed $(F_i \& F_j \text{in } C)$ on total size of training records of the dataset.
- Calculating the Probability$(F_i, F_j|C)$ by dividing Frequency of seemed $(F_i \& F_j \text{in } C)$ on Frequency of class C.
- Calculating Probability$(F_i|C)$ by dividing Frequency of seemed $(F_i \text{ in } C)$ on Frequency of class C.
- Calculating Probability$(F_j|C)$ by dividing Frequency of seemed $(F_j \text{ in } C)$ on Frequency of class C.

Then, the CMI will be summed for all paired features to find the weight between each two features. Table 9 shows our calculated weight between the attribute comorbidity and the sixteen other attributes. As a final step, the hidden parent is computed using Eqs. (4) and (5).

Table 7: The Frequency and Probability values for each of the selected feature's Intervals for all Classes

Feature No.	Feature	Interval	Asymptomatic class	Probability	COVID-19 (Mild)class	Probability	COVID-19 (Sever) class	Probability	Flu class	Probability	Normal class	Probability
1	**Disease group**	0	0	0	0	0	0	0	0	0	17590	1
		0.25	0	0	16742	1	0	0	0	0	0	0
		0.5	4425	1	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	10296	1	0	0	0	0
		1	0	0	0	0	0	0	10519	1	0	0
2	**Comorbidity**	0	0	0	4526	0.270	0	0	0	0	0	0
		0.1-0.5	4425	1	12216	0.729	5832	0.566	4895	0.465	17590	1
		0.6-0.9	0	0	0	0	4464	0.433	4293	0.408	0	0
		1	0	0	0	0	0	0	1331	0.126	0	0
3	**Hospital-day**	0.1-0.8	4425	1	16742	1	10296	1	10519	1	17590	1
4	**WBC/micro-L**	0	0	0	0	0	0	0	0	0	5677	0.322
		0.1-0.5	4425	1	8504	0.507	3717	0.361	4212	0.400	3459	0.196
		0.6-0.9	0	0	7576	0.452	4860	0.472	6307	0.599	8454	0.480
		1	0	0	662	0.039	1719	0.166	0	0	0	0
5	**Neutrophil**	0	4425	1	0	0	1345	0.130	0	0	0	0
		0.1-0.5	0	0	16742	1	2587	0.251	4895	0.465	17590	1
		0.6-0.9	0	0	0	0	5966	0.579	4293	0.408	0	0
		1	0	0	0	0	398	0.038	1331	0.126	0	0
6	**Chest X-ray**	0	0	0	0	0	1873	0.181	0	0	0	0
		0.1-0.5	4425	1	0	0	3959	0.384	1040	0.098	17590	1
		0.6-0.9	0	0	13503	0.806	4464	0.433	9479	0.901	0	0
		1	0	0	3239	0.193	0	0	0	0	0	0
7	**Treatment**	0	0	0	3978	0.237	0	0	0	0	0	0
		0.1-0.5	0	0	12764	0.762	3932	0.381	10519	1	17590	1
		0.6-0.9	0	0	0	0	6364	0.618	0	0	0	0
		1	4425	1	0	0	0	0	0	0	0	0
8	**Respiratory rate**	0	4425	1	0	0	0	0	0	0	0	0
		0.1-0.5	0	0	13503	0.806	3375	0.327	10519	1	17590	1
		0.6-0.9	0	0	0	0	6921	0.672	0	0	0	0
		1	0	0	3239	0.193	0	0	0	0	0	0

9	**O2 saturation**	0	0	0	0	0	4464	0.433	0	0	0	0
		0.6-0.9	4425	1	16742	1	3561	0.345	10519	1	17590	1
		1	0	0	0	0	2271	0.220	0	0	0	0
10	**O2 supplement**	0	0	0	0	0	7839	0.761	0	0	0	0
		0.3-0.7	0	0	0	0	2457	0.238	10519	1	17590	1
		1	4425	1	16742	1	0	0	0	0	0	0
11	**Temperature**	0	4425	1	0	0	1873	0.181	0	0	0	0
		0.1-0.5	0	0	12216	0.729	1502	0.145	10519	1	17590	1
		0.6-0.9	0	0	4526	0.270	5178	0.502	0	0	0	0
		1	0	0	0	0	1743	0.169	0	0	0	0
12	**Systolic BP**	0	0	0	4526	0.270	5966	0.579	0	0	0	0
		0.1-0.5	4425	1	7217	0.431	1112	0.108	10519	1	17590	1
		0.6-0.9	0	0	0	0	3218	0.312	0	0	0	0
		1	0	0	4999	0.298	0	0	0	0	0	0
13	**Heart rate**	0	0	0	0	0	398	0.038	0	0	0	0
		0.1-0.5	4425	1	12764	0.762	3932	0.381	10519	1	17590	1
		0.6-0.7	0	0	3978	0.237	1502	0.145	0	0	0	0
		1	0	0	0	0	4464	0.433	0	0	0	0
14	**Consciousness**	0	4425	1	16742	1	1345	0.130	0	0	0	0
		0.5	0	0	0	0	7449	0.723	10519	1	17590	1
		1	0	0	0	0	1502	0.145	0	0	0	0
15	**NEWS score**	0	0	0	8965	0.535	12	0.001	0	0	0	0
		0.1-0.5	4413	0.997	7765	0.463	3218	0.312	10519	1	17590	1
		0.6-0.7	12	0.002	12	0.001	2590	0.251	0	0	0	0
		1	0	0	0	0	4476	0.434	0	0	0	0
16	**Severity**	0	0	0	16742	1	0	0	0	0	0	0
		0.5	4425	1	0	0	0	0	10519	1	17590	1
		1	0	0	0	0	10296	1	0	0	0	0
17	**Cell type**	0	490	0.110	1333	0.079	915	0.088	322	0.030	1285	0.073
		0.1-0.5	2781	0.628	10429	0.622	5528	0.536	7669	0.729	10546	0.599
		0.6-0.9	1139	0.2574	4906	0.293	3839	0.372	2520	0.239	5688	0.323
		1	15	0.003	74	0.004	14	0.001	8	0.001	71	0.004

Table 8: Frequency and CMI computation for Comorbidity feature and Neutrophil feature from Class 3

Comorbidity feature value	**Neutrophil feature Value**	**Freq. of Comorbidity feature**	**Freq. of Neutrophil**	**Freq. of Comorbidity & Neutrophil**	CMIP
0	0	0	1345	0	0
0.1-0.5	0	5832	1345	1345	1.07060116
0.6-0.9	0	4464	1345	1345	0.81947249
1	0	0	1345	0	0
0	0.1-0.5	0	2587	0	0
0.1-0.5	0.1-0.5	5832	2587	2587	0.55661328
0.6-0.9	0.1-0.5	4464	2587	2587	0.42604967
1	0.1-0.5	0	2587	0	0
0	0.6-0.9	0	5966	0	0
0.1-0.5	0.6-0.9	5832	5966	5832	0.2413608
0.6-0.9	0.6-0.9	4464	5966	4464	0.18474531
1	0.6-0.9	0	5966	0	0
0	1	0	398	0	0
0.1-0.5	1	5832	398	398	3.61798633
0.6-0.9	1	4464	398	398	2.76932287
1	1	0	398	0	0
				CMI=$\sum CMIP =$	**9.68615193**

Table 9: Weight Values between Feature Comorbidity and the other Features of dataset

	Pair of Features	**Weight Value**		**Pair of Features**	**Weight Value**
1	W (Comorbidity & Disease group)	0.0080	*9*	W (Comorbidity & O2 supplement)	0.0004
2	W (Comorbidity &Hospital)	0.0080	*10*	W (Comorbidity & Temperature)	0.0017
3	W (Comorbidity &WBC)	0.0009	*11*	W (Comorbidity &Systolic)	0.0012
4	W (Comorbidity, Neutrophil)	0.0034	*12*	W (Comorbidity &Heart Rate)	0.0033
5	W (Comorbidity & Chest X-ray)	0.0009	*13*	W (Comorbidity & Consciousness)	0.0013
6	W (Comorbidity &treatment)	0.0003	*14*	W (Comorbidity &news)	0.0763
7	W (Comorbidity & Respiratory rate)	0.0003	*15*	W (Comorbidity & Severity)	8.8068
8	W (Comorbidity & O2 saturation)	0.0008	*16*	W (Comorbidity &Cell-Type)	0.9081

In testing phase, the hidden value is used to classify the tested records. Table 10 shows our testing record implementation using values from COVID-19 test set.

Table 10: Sample of Testing Record from COVID-19 Test Set

	Features description		Classes				
No.	**Feature Name**	**Interval**	**Asymptomatic**	**COVID-19(Mild)**	**COVID-19(Sever)**	**Flu**	**Normal**
1	Disease group	0.5	1	0	0	0	0
2	Comorbidity	0.3	1	0.73	0.57	0.79	1.00
3	Hospital-day	0.3	1	1	1	1	1
4	WBC/micro-L	0.2	1	0.51	0.36	0.40	0
5	Neutrophil	0.5	0	1	0.25	0.79	1
6	Chest X-ray	0.9	0	0.81	0.43	1	0
7	Treatment	0.4	0	0.76	0.38	1	1
8	Respiratory rate	0.1	0	0.81	0.33	1	1
9	O2 saturation	0.8	1	1	0.35	1	1
10	O2 supplement	0.2	0	0	0.25	1	1
11	Temperature	0.5	0	0.73	0.15	1	1
12	Systolic BP	0	0	0.27	0.58	0	0
13	Heart rate (BPM)	0.7	0	0.24	0.15	0	0
14	Consciousness	1	0	0	0.15	0	0
15	NEWS score	0.4	0.99729	0.46	0.31	1	1
16	Severity	0.5	1	0	0	1	1
17	Cell type	0	0.11073	0.08	0.09	0.03	0.13

8. Experimental Results and Discussion

In this section, a proposed multi-class model using NCA as a feature selection method and Hidden Naïve Bayes HNB as a statistical type classifier on the COVID-19 dataset will be evaluated and discussed. The model is implemented in multiple stages, first considering the COVID-19 dataset for evaluation. A total of about 59,570 patient records are submitted to pre-processing steps (cleaning, normalization, and discretization). Following that, the dataset is split into 2/3.

Training set, which is about 41,700 records, and 1/3 testing set, which is about 17,870 records. Using the training set, the feature selection method is utilized by adapting the NCA algorithm to select the most influent features extracted from patients's records. After applying the feature selection method, we calculated and computed the interval of each feature. A distribution of each selected feature is depicted in Figure 3. Then, the data is forwarded to the HNB classifier for classification. Figure 4 shows our distribution of each class in the COVID-19 training dataset.

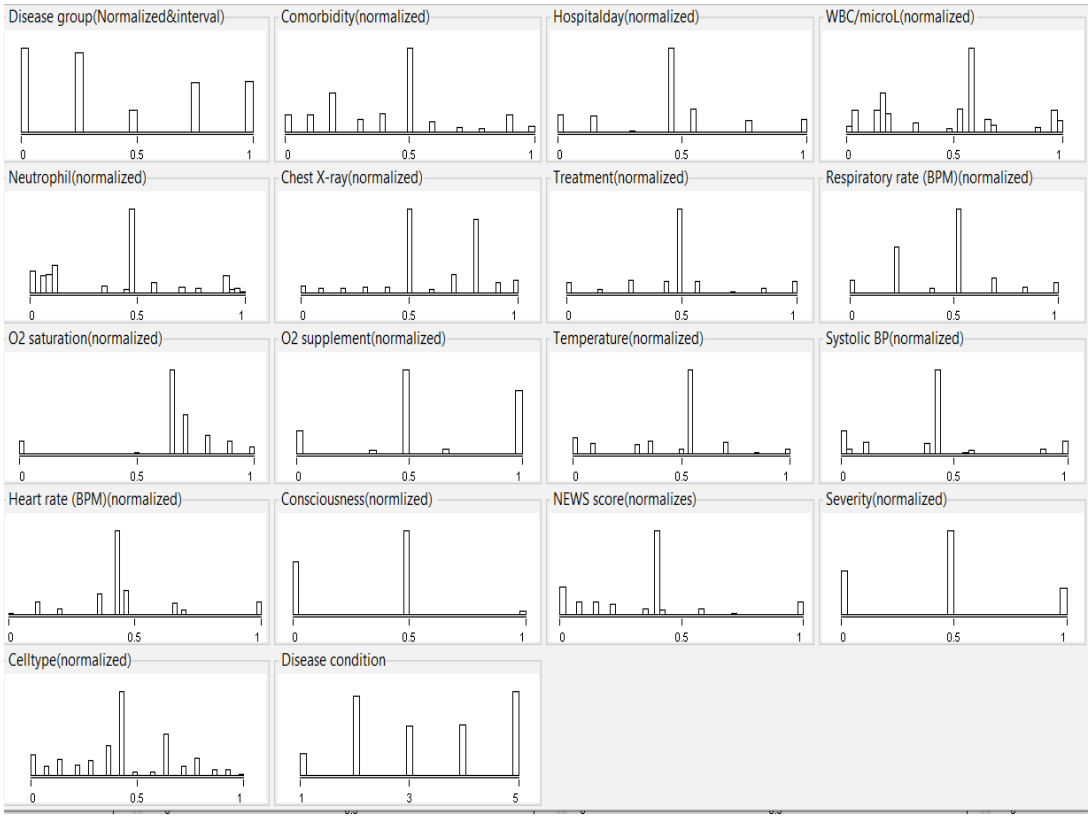

Figure 3: Dataset Features Distribution

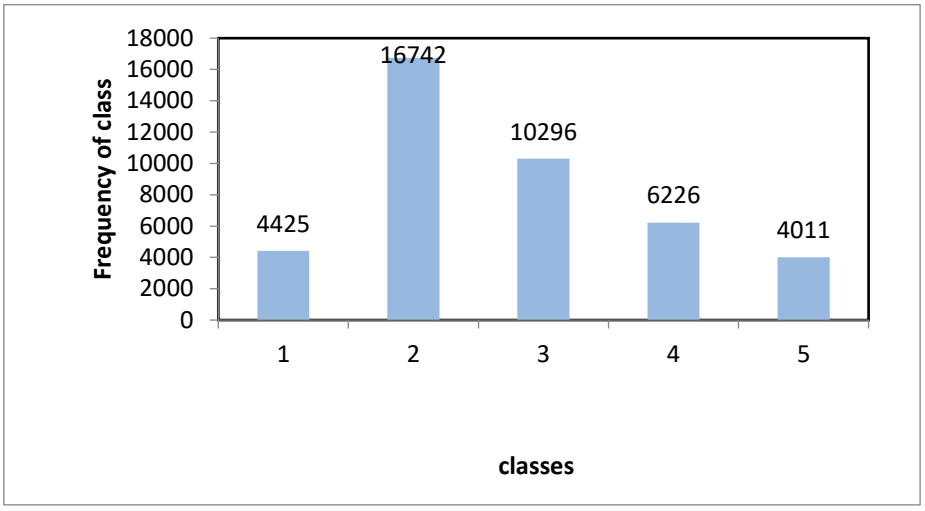

Figure 4: Distribution of each Class

To be noted, during the feature weighting module, for each feature, a hidden parent value is computed considering all other features. Therefore, many calculations are done during the HNB process. These resulted in a considerable number of values that can't be shown as a table. For this reason, in Figure 5, which depicts 16 charts, we show an example of conducting feature neutrophil intervals with all remaining features, including their intervals for class 3 (Severe COVID-19).

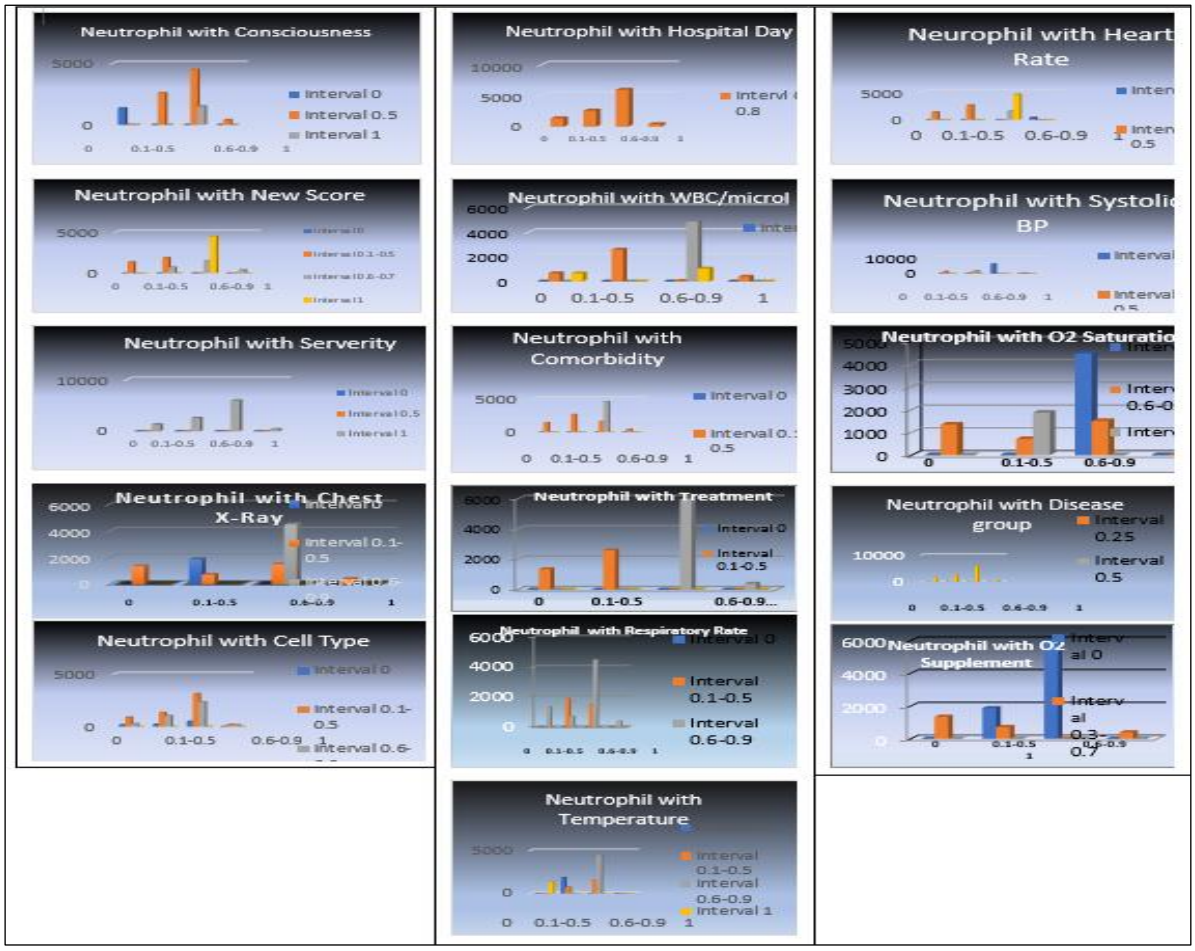

Figure 5: Neutrophil Feature with the other Remaining Features in Class 3

Throughout the conducted experiments, four assessment measures (accuracy is given by Eq. (8), recall is given by Eq. (9), precision is given by Eq. (10), and the F measure is given by Eq. (11)) were used to assess the proposed model [19]:

Accuracy is the rate of classification. Its formula is:

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{8}$$

The recall measures how many times it predicts yes. Its formula is:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{9}$$

Precision measures the number of times correctly predicted yes. Its formula is:

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{10}$$

The F1 measure represents the recall and precision in the following formula:

$$F1Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{11}$$

In Table 11, we show the confusion matrix for the COVID-19 test set based on five classes.

Table 11: Multi Class Confusion Matrix

	Asymptomatic	COVID-19(Mild)	COVID-19(Sever)	Flu	Normal
Asymptomatic	1858	500	500	200	800
COVID-19(Mild)	300	2711	500	1800	200
COVID-19(Sever)	500	500	1344	350	250
Flu	300	1000	500	1600	10
Normal	800	1000	150	20	1222

Table 12 shows our evaluation of the proposed model using the COVID-19 testing dataset. The testing phase used 17800 patients' records to evaluate the model's performance.

Table 12: Performance Measure for each Class in COVID-19 Dataset

Class type	Accuracy	Recall	Precision	F measure
Asymptomatic	0.95	0.92	0.97	0.94
COVID-19(Mild)	0.95	0.96	0.90	0.92
COVID-19(Sever)	0.86	0.84	0.81	0.82
Flu	0.81	0.88	0.67	0.76
Normal	0.88	0.62	0.96	0.75
The Average	0.89	0.84	0.86	0.83

According to the value of the results shown in Table 12, the accuracy of detecting flu was low, whereas the accuracy of detecting asymptomatic and COVID-19 (mild) was 95%. The precision measure for asymptomatic and normal classes reached 97% and 96%, respectively. Additionally, the recall for COVID-19 (mild) reached the highest, which is approximately 96% compared with other classes. A comparison is made on the accuracy between the proposed model and ref. [11] using the same COVID-19 dataset. It was shown that despite the vast calculations using HNB, the average accuracy was raised to 0.89 in comparison with the results yielded from ref. [11]. This is because each attribute (symptom) has a hidden parent value that reflects the influence of other attribute values. Table 13 shows our calculated accuracy value after implementing our proposed model and comparing it with the accuracy value found in [11].

Table 13: Accuracy Comparison with Reference [11]

	Method	Dataset	Accuracy
Ref [11]	Back propagation	COVID-19 dataset	0.79
Our Proposed Model	HNB	COVID-19 dataset	0.89

9. Conclusions

This paper suggests a proposed model to provide medical assistance in diagnosing COVID-19 according to specified symptoms. The model used a statistical classification method called Hidden Naïve Bayes to alleviate the feature's conditional independence assumption. Compared to other statistical classifiers, HNB creates a hidden parent value for each feature that synthesizes all of the other qualified features' influences. As a consequence, this assumption led to an improvement in the quality of the COVID-19 diagnosis because every symptom had a weight computed from all other symptoms, and a class probability was calculated taking into account the entire condition of each symptom. Hence, all symptoms contribute to diagnosing a patient's illness, which is a very important issue in obtaining accurate performance. Furthermore, HNB showed better efficiency than the work in [11], which used a backpropagation classifier.

References

[1] B. A. Al-Hameli, A. A. Alsewari and M. Alsarem, "Prediction of Diabetes Using Hidden Naïve Bayes: Comparative Study," in *Proceedings of Advances on Smart and Soft Computing*: Springer, pp. 223-233, Jan. 2021. Doi: 10.1007/978-981-15-6048-4_20.

[2] S. Y. Ilu, P. Rajesh and H. Mohammed, "Prediction of COVID-19 using Long Short-Term Memory by Integrating Principal Component Analysis and Clustering Techniques," *Informatics in Medicine Unlocked*, vol. 31, pp.1-7, 2022. Doi: https://doi.org/10.1016/j.imu.2022.100990.

[3] D. Jain and V. Singh, "Feature Selection and classification Systems for Chronic Disease Prediction: A Review," *Egyptian Informatics Journal*, vol. 19, no. 3, pp. 179-189, 2018. Doi: https://doi.org/10.1016/j.eij.2018.03.002.

[4] M. Jabbar and S. Samreen, "Heart Disease Prediction System Based On Hidden Naïve Bayes Classifier," in *Proceedings of the 2nd International Conference on Circuits, Controls, Communications and Computing (I4C)* IEEE, pp. 1-5. , Bangalore, India, 2016. Doi: 10.1109/CIMCA.2016.8053261.

[5] E. Elbasi, A. Zreikat, S. Mathew and A. E. Topcu, "Classification of Influenza H1N1 and COVID-19 Patient Data using Machine Learning," in *Proceedings of the 44th International Conference on Telecommunications and Signal Processing (TSP)*, IEEE, pp. 278-282, Jul. 2021. Doi: 10.1109/TSP52935.2021.9522591.

[6] 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. Doi: https://doi.org/10.1007/s42979-020-00216-w.

[7] T. Alsmadi, N. Alqudah and H. Najadat, "Prediction of COVID-19 patients States using Data Mining Techniques," in *Proceedings of the International Conference on Information Technology (ICIT)*: IEEE, pp. 251-256 , 2021. Doi: 10.1109/ICIT52682.2021.9491716.

[8] S. Kotsiantis and V. Tampakas, "Increasing the Accuracy of Hidden Naive Bayes Model," in *Proceedings of the 6th International Conference on Advanced Information Management and Service (IMS)*: IEEE, pp. 247-252, 2010. Electronic ISBN:978-89-88678-32-9.

[9] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, N. Rouf and M. M. U. Din, "Machine Learning based Approaches for Detecting COVID-19 using Clinical Text Data," *International Journal of Information Technology*, vol. 12, no. 3, pp. 731-739, 2020. Doi: https://doi.org/10.1007/s41870-020-00495-9.

[10] S. Bhatia and J. Malhotra, "Naïve Bayes Classifier for Predicting the Novel Coronavirus," in *Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*: IEEE, pp. 880-883, 2021, Doi:10.1109/ICICV50876.2021.9388410.

[11] A. F. Al-Zubidi, N. F. AL-Bakri, R. K. Hasoun, S. H. Hashim and H. T. Alrikabi, "Mobile Application to Detect COVID-19 Pandemic by Using Classification Techniques: Proposed System," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 16, pp. 34-51 2021. Doi: https://doi.org/10.3991/ijim.v15i16.24195.

[12] Y. Luo, "Research on Classification Methods Based on Improved TF-IDF and Double Hidden Naive Bayes," in *Proceedings of the* 7th *International Conference on Computing, Control and Industrial Engineering*, Springer, pp. 417-426, 2023. Doi: https://doi.org/10.1007/978-981-99-2730-2_41.

[13] N. F. AL-Bakri, A. F. Al-Zubidi, A. B. Alnajjar and E. Qahtan, "Multi Label Restaurant Classification using Support Vector Machine," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 9, no. 2, pp. 774-783, 2021. Doi: HTTP://dx.doi.org/10.21533/pen.v9i2.1876.

[14] H. A. Mahmood and S. H. Hashem, "Network Intrusion Detection System (NIDS) in Cloud Environment Based on Hidden Naïve Bayes Multiclass Classifier," *Al-Mustansiriyah Journal of Science*, vol. 28, no. 2, pp. 134-142, 2018. Doi: https://doi.org/10.23851/mjs.v28i2.508.

[15] N. F. AL-Bakri and S. H. Hashim, "Collaborative Filtering Recommendation Model Based on k-means Clustering," *Al-Nahrain Journal of Science*, vol. 22, no. 1, pp. 74-79, 2019. doi:10.22401/ANJS.22.1.10.

[16] N. F. AL-Bakri and S. Hassan, "A Proposed Model to Solve Cold Start Problem using Fuzzy User-Based Clustering," in *Proceedings of the 2nd Scientific Conference of Computer Sciences (SCCS)*: IEEE, pp. 121-125, 2019. Doi:10.1109/SCCS.2019.8852624.

[17] H. Zhang, L. Jiang and J. Su, "Hidden Naïve Bayes," in *Proceedings of the AAAI Conference on Artificial Intelligence 20*, pp. 919-924, 2005. Corpus ID: 1779699.

[18] A. R. Muhsen, G. G. Jumaa, N. F. AL Bakri and A. T. Sadiq, "Feature Selection Strategy for Network Intrusion Detection System (NIDS) Using Meerkat Clan Algorithm," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 16, pp. 158-171, 2021. Doi: https://doi.org/10.3991/ijim.v15i16.24173.

[19] N. F. AL-Bakri, J. F. Yonan, A. T. Sadiq and A. S. Abid, "Tourism Companies Assessment via Social Media Using Sentiment Analysis," *Baghdad Science Journal,* vol. 19, no. 2, pp. 422-429, 2022. Doi: 10.21123/bsj.2022.19.2.0422.