Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a KNN Variant Algorithm

Ahmed Hamed (ahmed_hamed@ci.suez.edu.eg)
Suez Canal University https://orcid.org/0000-0003-0928-548X
Ahmed Sobhy
Suez Canal University
Hamed Nassar
Suez Canal University

Research Article

Keywords: COVID-19, Classification, KNN, Incomplete data, Heterogeneous data

Posted Date: May 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-27186/v1

Version of Record: A version of this preprint was published at Arabian Journal for Science and Engineering on March 4th, 2021. See the published version at https://doi.org/10.1007/s13369-020-05212-z.

Noname manuscript No.
(will be inserted by the editor)

Accurate Classification of COVID-19 Based on Incomplete Heterogeneous Data using a *K*NN Variant Algorithm

Ahmed Hamed · **Ahmed Sobhy** · **Hamed Nassar**

Received: date / Accepted: date

Abstract The coronavirus 2019 disease (COVID-19) is wreaking havoc around the world, and great efforts are underway to control it. Millions of people are now being tested and their data keeps accumulating in large volumes. This data can be used to classify newly tested persons as whether they have the disease or not. However, normal classification techniques are hampered by the fact that the data is typically both incomplete and heterogeneous. To address this two-fold obstacle, we propose a *K*NN variant (*K*NNV) algorithm which accurately and efficiently classifies COVID-19. The main two ideas behind the proposed algorithm are that for each instance to be classified it chooses the parameter K adaptively and calculates the distances to other instances in a novel way. The *K*NNV was implemented and tested on a COVID-19 dataset from the Italian society of medical and intervention radiology society. It was also compared to three algorithms of its category. The test results show that the *K*NNV can efficiently and accurately classify COVID-19 patients. The comparison results show that the algorithm greatly outperforms all its competitors in terms of four metrics: precision, recall, accuracy, and F-Score.

Keywords COVID-19 · Classification · *K*NN · Incomplete data · Heterogeneous data.

1 Introduction

The coronavirus disease 2019 (COVID-19) [1] is now sweeping the world. It is causing a major threat to human life, with severe economic consequences. Its symptoms include cough, fever and respiratory complications. The hazardous side of COVID-19 is its rapid spreading [2] because of it is transmitted by contact and by small droplets produced when

A. Hamed
Suez Canal University - Ismailia - Egypt
Tel.: +2-012-2725375
E-mail: ahmed_hamed@ci.suez.edu.eg

A. Sobhy
E-mail: ahmed_sobhy@ci.suez.edu.eg

H. Nassar
E-mail: nassar@ci.suez.edu.eg

people cough, sneeze or talk. To make matters worse, COVID-19 can survive on surfaces up to 72 hours [3], causing people to catch it by touching apparently normal objects.

The best way to improve a COVID-19 patient survival rate is through the early detection of the disease [4], and here is where AI techniques, such as what is employed in the present work, can help. It is widely believed that AI has the potential to tackle the pressing issues related by COVID-19 if there is information collected about the patient [5]. However, this information can be heterogeneous, in the sense that their features are of two types, categorical and numerical [6]. Categorical features are qualitative, e.g. cough, whereas numerical features are quantitative, e.g. body temperature. The presence of both types complicates processing. What is more, this information can also be incomplete, in the sense that some feature have missing values [7]. Values can be missing due to negligence, difficulty or cost.

These two issues, heterogeneity and incompleteness, present tremendous challenges for the classification of COVID-19 cases in the data collected about numerous persons. Inherently, common classification algorithms do not handle heterogeneity and missing values explicitly. Therefore, one needs to adapt an algorithm to handle incomplete heterogeneous COVID-19 (IHC) datasets efficiently accurately, and this is the goal of the present work.

Formally, an IHC dataset can be defined by a triple $(\mathscr{U},\mathscr{A},\mathscr{V})$. The set $\mathscr{U}=\{u_1,u_2,\ldots,u_m\}$ is a non-empty finite set of patients called the universe. The set $\mathscr{A}=\{a_1,a_2,\ldots,a_n,d\}$ is a non-empty finite set of features describing the patients. It contains $n+1$ features, n of which, namely a_1,a_2,...,a_n, are conditional features and the $n+1^{\text{st}}$ feature is a decision label. The values of some of the a_i may be incomplete, with an incomplete value represented by "$*$". If $\mathbb{V}=\{\zeta_1,\zeta_2,\ldots,\zeta_r\}$ is the value set of d, then d partitions the universe $\mathscr{U}$ into r decision subsets, $\mathfrak{U}_{\zeta_1},\mathfrak{U}_{\zeta_2},\ldots,\mathfrak{U}_{\zeta_r}$, where $\mathfrak{U}_{\zeta_k}$ is the subset of all patients with decision label value ζ_k. Finally, the set $\mathscr{V}$ is the union of the value sets of all features. That is,

$$\mathscr{V}=\left\{v_{u_i,a_j}\,\middle|\,u_i\in\mathscr{U},a_j\in\mathscr{A}^-\right\},$$

where, v_{u_i,a_j} is the value of feature a_j of patient u_i and

$$\mathscr{A}^-=\mathscr{A}-\{d\}.$$

For example, $v_{u_1,a_2}=3$ means that feature a_2 of patient u_1 has the value 3. Accordingly, $\mathscr{A}^-=\mathscr{C}\cup\mathscr{N}$, where $\mathscr{C}$ and $\mathscr{N}$ are two disjoint sets of categorical and numerical features, respectively. Table 1 shows a toy IHC dataset, used repeatedly in the sequel to illustrate the proposed algorithm.

In general, classification is a two-step process [8]. In the first, a classification algorithm learns from a set of patients $u_i\in\mathscr{U}$ whose decision values are known. In the second, the algorithm uses what it has learnt to classify an patient u_0 whose decision value is not known. A classification algorithm, used in the present work, that can be applied to IHC data with heterogeneous and missing values, is the K nearest neighbor (KNN) algorithm [9]. Its main idea is to search for K neighbor patients nearest the unknown patient u_0 and then predict the decision value of the latter by a majority vote of the neighbors. It is useful in the present work because it does not require the decision subsets $\mathfrak{U}_{\zeta_j}$ to be linearly separable.

As good as it is, KNN has three problems. First, the selection of a proper K value, which influences the performance of KNN, is challenging. A small K increases the influence of noise on prediction, while a large K increases computational complexity. Second, KNN uses the same K for all unknown patients to be classified, whereas K is to a great extent patient dependent. This causes to a high percentage of misclassification. Third, KNN computes distances inaccurately, as discussed in Section 3, also causing a high percentage of misclassification. These three problems are mitigated by the present work.

Table 1 A toy IHC, where $\mathscr{U} = \{u_1, u_2, \ldots, u_{13}\}$ is the set of patients, $\mathscr{A} = \{a_1, a_2, \ldots, a_7, d\}$ with $\mathscr{C} = \{a_1, a_2, \ldots, a_5\}$ being categorical features, $\mathscr{N} = \{a_6, a_7\}$ being numerical features and d is the decision label. Note that an $*$ denotes a missing value and in this example it represents about 20% of the data.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	d
u_1	Male	$*$	Yes	Yes	Yes	0.97	$*$	COVID-19
u_2	$*$	Yes	$*$	No	Yes	$*$	0.39	COVID-19
u_3	Female	No	$*$	No	Yes	0.77	0.79	Flu
u_4	Male	Yes	No	Yes	No	0.08	0.8	Flu
u_5	Female	Yes	Yes	$*$	No	0.8	0.1	COVID-19
u_6	Female	$*$	$*$	No	Yes	0.42	0.55	Flu
u_7	Male	Yes	Yes	No	Yes	0.98	$*$	COVID-19
u_8	Male	Yes	No	Yes	Yes	0.43	0.42	Flu
u_9	$*$	Yes	No	No	Yes	0.96	$*$	Flu
u_{10}	Female	$*$	No	Yes	Yes	$*$	0.34	COVID-19
u_{11}	Male	Yes	Yes	Yes	$*$	0.38	0.39	Flu
u_{12}	Male	Yes	Yes	Yes	No	0.85	$*$	COVID-19
u_{13}	Female	No	Yes	$*$	$*$	0.31	0.59	Flu

We introduce an *K*NN variant (*K*NNV) classification algorithm able to accurately identify COVID-19 cases even when the test data is incomplete and heterogeneous. Specifically, *K*NNV chooses K adaptively for each unknown patient and computes accurately the distances between patients. For validation, the *K*NNV is tested on publicly available IHC datasets from the Italian Society of Medical and Intervention Radiology (SIRM) [10]. The test results show excellent performance compared with algorithms of the same category.

The rest of this article is organized as follows. Section 2 covers related work. Section 3 describes the proposed algorithm. In Section 4, experimental work is carried out to validate the algorithm and discussion about the findings is given. Finally, concluding remarks are presented in Section 5.

2 Related work

Researchers in different fields are now actively fighting COVID-19, and information science researchers are no exception. Roosa et al. [11] propose a forecasting algorithm for COVID-19 in China. They use phenomenological models to develop and assess short-term forecasts of the cumulative number of confirmed COVID-19 patients in the Chinese Hubei province. McCall [12] reports that AI is doing a paradigm shift in health care and is, therefore, considered a promising tool to trap COVID-19. He recommends using AI algorithms to predict the location of the next outbreak. Hu et al. [13] propose an AI system for real-time forecasting of COVID-19 with the aim of estimating the life time of the virus. They propose a modified stacked auto-encoder system for modeling the transmission dynamics of the pandemic throughout China. Pirouz et al. [14] propose a binary classification system using artificial neural networks (ANN) to predict the number of confirmed COVID-19 cases in Hubei province. They use a variety of features such as maximum, minimum, and average daily temperature, the density of city, humidity and wind speed as inputs to the ANN, with the aim of predicting the confirmed number of COVID-19 patients in the next 30 days. Du et al. [15] propose a hybrid AI model that combines the strengths of both a natural language processing module and a long short-term memory network. Their objective is to analyze the change in the infectious capacity of COVID-19 patients within a few days after they catch the virus. Santosh [16] proposes an AI-driven system able to predict the time of the next COVID-19 outbreak, while forecasting the COVID-19 possibility to spread across the

globe. Boldog et al. [17] propose a computational tool able to assess the risks of COVID-19 outbreaks outside China. They compute the probability of a major outbreak in a country through testing three features: cumulative number of patients in the non-locked down Chinese provinces, connectivity of that country with China and efficacy of control measures in that country. All the above proposals focus on predicting outbreaks, overlooking classification of existing cases. For this latter objective, the endeavors next have been made.

Gozes et al. [18] propose an AI based computational tomography (CT) image classification algorithm for the detection, quantification and tracking of COVID-19. This algorithm has the ability to distinguish COVID-19 patients from other patients. They use a deep learning model to classify COVID-19 from CT images. Ai et al. [19] propose a CT image algorithm for COVID-19 identification that uses a reverse-transcription polymerase chain reaction test. Barstugan et al. [20] propose a feature extraction process of CT images and a discrete wavelet transform algorithm to improve COVID-19 classification. Specifically, they use a grey level co-occurrence matrix, local directional pattern, grey level run length matrix and grey-level size zone matrix. Afterwards, they use support vector machines to classify based on extracted features. Xu et al. [21] propose a CT image classification algorithm for the early classification of COVID-19 using deep learning techniques. The algorithm is able to distinguish COVID-19 patients from Flu patients. They first segment the CT images using a 3D deep learning model, then the segmented images are binary classified. Wang et al. [22] use CT images to extract COVID-19 conditional graphical features that can then be used to distinguish COVID-19 patients from other patients. Li et al. [23] propose a CT image classification algorithm for the early classification of COVID-19. They exploit both deep learning and ANN to extract visual features from chest CT images. All the above endeavors depend on medical imaging, which may not be available or accessible. An alternative classification direction, in which the present work falls, depends on available personal and test data. Numerous attempts have been made in this direction as described next.

Peng et al. [24] use AI techniques to improve the classification accuracy of COVID-19. They use sparse rescaled linear square regression, evolutionary non-dominated radial slots based algorithm, attribute reduction with multi-objective decomposition-ensemble optimizer, gradient boosted feature selection, and recursive feature elimination. Rao and Vazquez [25] propose to classify COVID-19 from collected data about travel history along with common manifestations using a phone-based online survey. This data is then used to divide patients into four decision subsets: no-risk, minimal-risk, moderate-risk and high-risk. Maghdid et al. [26] propose a framework for the early classification of COVID-19 using on-board smart-phone sensors. They make use of temperature, inertial, proximity, color, humidity, and wireless chipset sensors embedded in smart-phones. An interesting observation about this attempt is its ability to provide low-cost classification compared to other attempts. All the above attempts have one thing in common—they impute missing values in collected data. This imputation is harmful because it affects the data distribution, on the one hand, and breaks down some relations between conditional features and the decision label, on the other [27]. Besides, the above attempts do not handle heterogeneous data directly. They convert categorical values to 0's and 1's as a turn around, negatively impacting classification accuracy. The present work, though falling in the same direction, aims at avoiding all these drawbacks.

3 Proposed approach

In this section, we introduce the *K*NNV classification algorithm as a tool to identify COVID-19 cases in incomplete heterogeneous test data. The contribution of *K*NNV is twofold. First, for each unknown patient u_0, it computes a special K value suitable for that patient. Second, accurate distance calculations are employed.

3.1 Preliminaries

Given a set $\mathscr{U}$ of patients described by a set of features $\mathscr{A}^-$, a normalization function is applied for the set $\mathscr{N} \subset \mathscr{A}^-$ to scale the numerical features to the interval $[0,1]$ to prevent features with large ranges from outweighing those with smaller ranges. A normalized feature value $\widehat{v}_{u_i,a_j}$ is obtained from its raw counterpart v_{u_i,a_j} by

$$\widehat{v}_{u_i,a_j} = \begin{cases} \frac{v_{u_i,a_j} - \min\limits_{u_n}\left(v_{u_n,a_j}\right)}{\max\limits_{u_n}\left(v_{u_n,a_j}\right) - \min\limits_{u_n}\left(v_{u_n,a_j}\right)} & \text{if } v_{u_i,a_j} \neq * \\ * & \text{if } v_{u_i,a_j} = * \end{cases},$$

where $\min\limits_{u_n}\left(v_{u_n,a_j}\right)$ and $\max\limits_{u_n}\left(v_{u_n,a_j}\right)$ are the minimum and maximum values, respectively, of feature a_j across all patients u_n, .

For any unknown patient u_0, *K*NN searches the IHC data for a subset $\mathbb{U}_{K,u_0} \subset \mathscr{U}$ of patients, of arbitrary size $K = |\mathbb{U}_{K,u_0}|$, whose members $u_i \in \mathbb{U}$ are closest to u_0. Closeness between the unknown patient u_0 and another patient $u_i \in \mathbb{U}_{K,u_0}$ is measured with respect to some set $\mathscr{M} \subseteq \mathscr{A}^-$ of features using the Euclidean distance given by

$$l_{\mathscr{M}}(u_0,u_i) = \sqrt{\sum_{n=1}^{|\mathscr{M}|}\left(v_{u_0,a_n} - v_{u_i,a_n}\right)^2}. \tag{1}$$

Let us comment on how to calculate the difference under the square root. First, for categorical features, the difference between two feature values is calculated via the Hamming distance as follows.

- If both values are existing and identical (e.g., v_{u_i,a_n} is male and v_{u_0,a_n} is male) or both are missing, the difference is considered 0.
- Else (i.e. if both values are existing and different (e.g., v_{u_i,a_n} is male and v_{u_0,a_n} is female) or one of them is missing), then the difference is considered 1.

For numerical features, the difference between two feature values is calculated via the Euclidean distance as follows.

- If both values are existing, the difference is calculated normally by subtraction.
- Else (i.e. if one value is missing), the difference is considered the existing value.

Finally, given the set $\mathbb{U}_{K,u_0} \subseteq \mathscr{U}$ of the K nearest neighbors of the unknown patient u_0, then the decision label $v_{u_0,d}$ of u_0 is assigned a value $\zeta_j \in \mathbb{V}$, $j = 1,2,\ldots,r$, given by

$$v_{u_0,d} = \underset{\zeta_j}{\operatorname{argmax}} \sum_{u_i \in \mathbb{U}_{K,u_0}} \delta_{v_{u_i,d},\zeta_j}, \tag{2}$$

where $\delta_{x,y}$ is the Kronecker delta function given by

$$\delta_{x,y} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}.$$

In (2), we essentially carry out a majority vote as follows. The sum is evaluated r times, for $\zeta_1, \zeta_2, \ldots, \zeta_r$, and the ζ_j that results in the highest value of the sum will be the voted decision label. The sum itself is evaluated in K steps, for the K members $u_i \in \mathbb{U}_{K,u_0}$. Starting at 0, the sum is incremented by 1 if the label $v_{u_i,d}$ of u_i is identical to the label ζ_j currently being considered by argmax_{ζ_j} and is not incremented otherwise.

3.2 *K*NNV algorithm

In reference to the items mentioned above in the comment on how to calculate the differences for (1), we can note the following. Since numerical values are normalized to $[0,1]$, the difference between any two existing values is also in $[0,1]$. In the means time, since categorical features are in $\{0,1\}$,they lead to differences also in $\{0,1\}$. Therefore, in calculating Euclidean distances between patients, categorical and missing values have the greatest impact. This means that special attention should be paid to these types of values.

In *K*NNV, for each unknown patient u_0, three nearest neighbor classes are defined with respect to the sets $\mathscr{C}$, $\mathscr{N}$ and $\mathscr{A}^-$, with the aim to tackle the issue of data incompleteness. The first is the neighborhood class $\theta_{\mathscr{C},u_0}$ which contains the nearest patients to u_0 with respect to categorical features $\mathscr{C}$. Before defining $\theta_{\mathscr{C},u_0}$ formally, we introduce two definitions.

Definition 1. (Feature similarity relation, $\mapsto$): Let u_0 and u_i, $u_0 \neq u_i$, be two different patients, and let v_{u_0,a_n} and v_{u_i,a_n} be the values of their n^{th} feature. We say that the feature value v_{u_0,a_n} is similar to the feature value v_{u_i,a_n}, and denote that by $v_{u_0,a_n} \mapsto v_{u_i,a_n}$, if and only if $v_{u_0,a_n} = *$ and $v_{u_i,a_n} \neq *$.

Now, we employ Definition 1 to define the neighborhood relation N, which relates an unknown patient u_0 to its nearest patient u_i with respect to the categorical features $\mathscr{C}$.

Definition 2. (Neighborhood relation, N): Let u_0 and u_i be two patients, $u_0 \neq u_i$, and let v_{u_0,a_n} and v_{u_i,a_n} be the values of their n^{th} feature. We define the neighborhood relation N as the set of pairs (u_0,u_i), such that for every $a_n \in \mathscr{C}$, the value v_{u_i,a_n} is either similar or equal to the value v_{u_0,a_n}, or the value v_{u_0,a_n} is either similar or equal to the value v_{u_i,a_n}. Using quantifiers, this is expressed as

$$N = \left\{ (u_0,u_i) \mid \forall u_i \in \mathscr{U}, \forall a_n \in \mathscr{C} : v_{u_i,a_n} \mapsto v_{u_0,a_n} \vee v_{u_0,a_n} \mapsto v_{u_i,a_n} \vee v_{u_i,a_n} = v_{u_0,a_n} \right\}.$$

The neighborhood relation N of Definition 2 is used to define, for each unknown patient u_0, the neighborhood class $\theta_{\mathscr{C},u_0}$, given by

$$\theta_{\mathscr{C},u_0} = \{ u_i \,|\, u_i \in \mathscr{U} \wedge (u_0,u_i) \in N \}. \tag{3}$$

An interesting observation about the class $\theta_{\mathscr{C},u_0}$ is that it gathers the exact number of nearest patients without the need for a parameter K. Actually, the class $\theta_{\mathscr{C},u_0}$ is used to determine the proper K value for the unknown patient u_0. Specifically, for each unknown patient u_0, the value K is set to the cardinality of its neighborhood class $\theta_{\mathscr{C},u_0}$, i.e. $K = |\theta_{\mathscr{C},u_0}|$.

Now, *K*NNV proceeds to compute the neighborhood class $\theta_{\mathscr{N},u_0}$, which relates an unknown patient u_0 to its K nearest neighbors with respect to numerical features $\mathscr{N}$, using $K = |\theta_{\mathscr{C},u_0}|$ to classify the patient as follows.

Let $L = (l_1, l_2, \ldots, l_{|\mathscr{U}|})$ be the list of distances, with respect to $\mathscr{N}$, between the unknown patient u_0 and the patients $u_1, u_2, \ldots, u_{|\mathscr{U}|}$, respectively. Further, let $\mathscr{Z} = \{z_1, z_2, \ldots, z_K\}$ be the set of indices of the K smallest values in L. Clearly, then $\mathscr{Z}$ has the indices of the nearest K patients to u_0 with respect to $\mathscr{N}$. The set $\mathscr{Z}$ is now used to define the neighborhood class $\theta_{\mathscr{N},u_0}$ with respect to $\mathscr{N}$ as follows.

$$\theta_{\mathscr{N},u_0} = \{u_i \mid u_i \in \mathscr{U} \wedge i \in L\}. \tag{4}$$

With the above in mind, the neighborhood class $\theta_{\mathscr{A}^-,u_0}$ of an unknown patient u_0 with respect to the entire set of features $\mathscr{A}^-$ is given by

$$\theta_{\mathscr{A}^-,u_0} = \begin{cases} \theta_{\mathscr{C},u_0} \bigcap \theta_{\mathscr{N},u_0} & \text{if } \theta_{\mathscr{C},u_0} \bigcap \theta_{\mathscr{N},u_0} \neq \emptyset \\ \theta_{\mathscr{C},u_0} \bigcup \theta_{\mathscr{N},u_0} & \text{otherwise} \end{cases}. \tag{5}$$

The class $\theta_{\mathscr{A}^-,u_0}$ is basically the set of K nearest neighbors of the unknown patient u_0, that is sufficient to discover the decision label $v_{u_0,d}$, which is the ultimate goal of the KNNV algorithm. Simply, the algorithm now just employs (2), using $\mathbb{U}_{K,u_0} = \theta_{\mathscr{A}^-,u_0}$ and $\mathscr{M} = \mathscr{A}^-$, to find $v_{u_0,d}$. The KNNV pseudocode that formalizes all the above steps is shown in Algorithm 1.

Before leaving this section, we should emphasize the advantage of using the above three-step procedure for determining K dynamically over using a fixed K for all test cases. The problem is that it is hard to find the exact number of nearest patients with respect to the numerical feature set $\mathscr{N}$. This is because a neighborhood threshold should be applied which, in turn, would result in a different set of nearest patients whenever the threshold changes.

Algorithm 1 K Nearest Neighbor Variant (KNNV) algorithm

Input: $(\mathscr{U}, \mathscr{A}, \mathscr{V})$ //IHC
u_0 //Unknown patient with no assigned decision label d
Output: $v_{u_0,d}$ //Decision label of u_0
//Calculations with categorical features $\mathscr{C}$:
Calculate the neighborhood class $\theta_{\mathscr{C},u^-}$ as per (3)
$K := |\theta_{\mathscr{C},u_0}|$
//Calculations with numerical features $\mathscr{N}$:
$L := \emptyset$ //Set of Euclidean distances
$m := |\mathscr{U}|$ //Number of patients in dataset
for i = 1 **to** m **do**
$L := L \bigcup \{l_{\mathscr{N}}(u_0, u_i)\}$ as per (1) //According to the numerical features $\mathscr{N}$ only
end for
Construct the set $\mathscr{Z}$ with the indices of K nearest neighbor patients to u_0
Calculate the neighborhood class $\theta_{\mathscr{N},u_0}$ with K patients as per (4)
//Calculations with the whole set of features $\mathscr{A}^-$:
Calculate the neighborhood class $\theta_{\mathscr{A}^-,u_0}$ as per (5)
Find the decision label $v_{u_0,d}$ as per (2)

3.3 Illustrative example

Below, we explain how the KNNV predicts the decision label of an unknown patient $u_0 = <\text{Female}, *, \text{Yes}, \text{No}, *, 0.43, 0.79>$ based on the toy IHC dataset shown in Table 1.

Neighborhood class $\theta_{\mathscr{C},u_0}$ with respect to $\mathscr{C} = \{a_1, a_2, \ldots, a_5\}$:

As per Definition 2, we search for the neighboring patients $u_i \in \mathscr{U}$ in which the values of the categorical feature $a_j \in \mathscr{C}$ of both u_i and u_0 are either similar or equal to each other. In this context, patient u_3, with categorical feature vector <Female, No, $*$, No Yes>, is a neighbor of u_0 with categorical feature vector <Female, $*$, Yes, No $*$>. This is because the categorical feature values of u_3 and u_0 are either equal (features a_1, a_4) or similar (features a_2, a_3, a_5). Likewise u_5, u_6 and u_{13} are neighbors of u_0 with respect to $\mathscr{C}$. Therefore, as per (3), the neighborhood class $\theta_{\mathscr{C},u_0}$ of u_0 with respect to the categorical features $\mathscr{C}$ is $\theta_{\mathscr{C},u_0} = \{u_3, u_5, u_6, u_{13}\}$. Consequently, for this unknown patient u_0, $K = |\theta_{\mathscr{C},u_0}| = 4$.

Neighborhood class $\theta_{\mathscr{N},u_0}$ with respect to $\mathscr{N} = \{a_6, a_7\}$:

We compute the Euclidean distance, as per (1), between the unknown patient u_0 and every patient $u_i \in \mathscr{U}$ with respect to the numerical features $a_j \in \mathscr{N}$. Let us find the distance between u_1 with numerical feature vector $< 0.97, * >$ and u_0 with numerical feature vector $< 0.43, 0.79 >$. For feature a_6 with both $v_{u_1,a_6} \neq *$ and $v_{u_0,a_6} \neq *$, the difference is taken 0.54, while for a_7, the difference is taken 0 because $v_{u_1,a_7} = *$. Therefore, $l_{\mathscr{N}}(u_0, u_1) = \sqrt{0.54^2 + 0^2} = 0.54$.

Repeating the previous exercise with all $u_i \in \mathscr{U}$ ends up with the list

$L = (0.54, 0.4, 0.34, 0.35, 0.78, 0.24, 0.55, 0.37, 0.53, 0.45, 0.4, 0.42, 0.23)$. The next step is to find the set $\mathscr{Z}$ of indices of the four (since $K = 4$) closest patients to the unknown patient. By inspection, $\mathscr{Z} = \{13, 6, 3, 8\}$, corresponding to the distances $\{0.23.0.24, 0.34, 0.37\}$. It follows from (4) that the neighborhood class $\theta_{\mathscr{N},u_0} = \{u_3, u_6, u_8, u_{13}\}$.

Neighborhood class $\theta_{\mathscr{A}^-,u_0}$ with respect to $\mathscr{A}^- = \{a_1, a_2, \ldots, a_7\}$:

As per (5), the final K nearest neighbor ($\theta_{\mathscr{A}^-,u_0}$) patients to u_0 with respect to the entire set of feature $\mathscr{A}^-$ is the intersection between both $\theta_{\mathscr{C},u_0}$ and $\theta_{\mathscr{N},u_0}$. Since this intersection is nonempty, it follows that $\theta_{\mathscr{A}^-,u_0} = \{u_3, u_6, u_{13}\}$. According to the majority vote (2), noting that $v_{u_3,d} = v_{u_6,d} =$ Flu and $v_{u_{13},d} =$ COVID-19, the decision label of u_0 is $v_{u_0,d} =$ Flu.

4 Experimental work

The *K*NNV algorithm was coded in Matlab R16a, and run on a PC with Centos 7, Intel(R) Core(TM) i7 CPU2.4 GHz with 16 GB of main memory.

There are two objectives for the experiments carried out in this section. The first is to validate the *K*NNV algorithm by showing its superiority to related algorithms that can handle both heterogeneity and incompleteness in the data, namely Modified *K*NN (M*K*NN) [28], *K*NN for imperfect data (KNN$_{\text{imp}}$) [29] and cost-sensitive *K*NN (cs*K*NN) [30]. For each algorithm, the precision, recall, accuracy, and F-Score [31] metrics were evaluated. The second objective is to test the classification significance of each conditional feature with respect to the final classification decision.

The classification performance was analyzed using a 10-fold cross-validation method. The whole IHC dataset was split into ten equal sub-datasets; nine served for training, and the last one for testing. This means that each patient appeared in a test set once, and appeared in a training set nine times. *K*NNV and its counterparts were independently run ten times and, then, the averages of the results attained are presented. As the K value affects the performance of M*K*NN, KNN$_{\text{imp}}$ and cs*K*NN, we tested their performance for $K = 2, 3, \ldots, 10$, and for each algorithm we chose the value giving the best results. Accordingly, $K = 5$ was chosen for M*K*NN, $K = 7$ for KNN$_{\text{imp}}$ and $K = 4$ for cs*K*NN.

Table 2 IHC dataset used in the experiments, with 68 COVID patients and 62 Flu patients described by 16 conditional features a_i and one decision feature d.

	Feature	Type	Value set
a_1	Age	Numerical	[4-90]
a_2	Gender	Categorical	{Male, Female}
a_3	Fever	Categorical	{Yes, No}
a_4	Dyspnea	Categorical	{Yes, No}
a_5	Nasal	Categorical	{Yes, No}
a_6	Cough	Categorical	{Yes, No}
a_7	Partial pressure of oxygen (PO2)	Numerical	[32-292]
a_8	C-reactive protein (CRP)	Numerical	[0.75-23]
a_9	Asthenia	Categorical	{Yes,No}
a_{10}	Leukopenia	Categorical	{Yes,No}
a_{11}	Exposure to COVID-19 patients	Categorical	{Yes, No}
a_{12}	Coming from high risk zone	Categorical	{Yes, No}
a_{13}	Temperature	Numerical	[35.7-40]
a_{14}	Blood test	Categorical	{Yes, No}
a_{15}	Polymerase chain reaction (RT-PCR)	Categorical	{Positive, Negative}
a_{16}	Medical History	Categorical	{Cancer, Croonic, Astham, COPD, Chronic, DM}
d	Decision label	Categorical	{COVID-19, Flu}

4.1 Dataset

We need a dataset containing COVID-19 cases that we can test the classification algorithms on. However, we could not find such a dataset with a mix of COVID-19 and non-COVID-19 cases. So, the only solution to have the desired mixed dataset was to compose it manually. First, we composed a dataset of 68 COVID-19 cases from the SIRM database [10]. Second, for Non-COVID-19 cases, we composed a dataset of 62 Flu cases from the influenza research database (IRD) [32]. Then we merged the two datasets and shuffled them randomly to obtain an IHC dataset with two decision subsets; {COVID-19, Flu}.

It should be noted that we faced a problem with data of the 68 COVID-19 cases obtained from SIRM. Specifically, that data was unstructured, in the sense that the attributes of each case were described verbally in a separate paragraph. Therefore, we had to first structure the data in a matrix format consistent with that of the Flu cases. The resulting mixed dataset contains features that are categorical, such as cough, and features that are numerical, such as partial pressure of oxygen (PO2). Additionally, since not all patients undertook the same medical tests (for example, some patients had a blood test and others did not), there were missing feature values. As such, the dataset used in the present work is heterogeneous with about 52% of the feature values missing. This dataset is described in Table 2.

4.2 Experiment 1: Performance evaluation of *K*NNV

We have assessed the performance of *K*NNV using the following four metrics

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

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

Table 3 Average values of Precision, Recall, Accuracy and F-Score achieved by *K*NNV and three related algorithms over 10 runs.

	M*K*NN	cs*K*NN	*K*NN$_{\text{imp}}$	*K*NNV
Precision	0.42	0.52	0.77	**0.95**
Recall	0.59	0.62	0.70	**0.93**
Accuracy	0.41	0.51	0.66	**0.93**
F-Score	0.44	0.54	0.79	**0.92**

Table 4 Maximum values of Precision, Recall, Accuracy and F-Score achieved by *K*NNV and three related algorithms over 10 runs.

	M*K*NN	cs*K*NN	*K*NN$_{\text{imp}}$	*K*NNV
Precision	0.75	0.81	0.87	**1**
Recall	0.83	1	0.90	**1**
Accuracy	0.68	0.85	0.81	**1**
F-Score	0.63	0.90	0.94	**1**

Table 5 Minimum values of Precision, Recall, Accuracy and F-Score achieved by *K*NNV and three related algorithms over 10 runs.

	M*K*NN	cs*K*NN	*K*NN$_{\text{imp}}$	*K*NNV
Precision	0.27	0.38	0.67	**0.75**
Recall	0.11	0.22	0.37	**0.79**
Accuracy	0.28	0.33	0.56	**0.85**
F-Score	0.17	0.31	0.62	**0.82**

$$\text{F-Score} = 2\left(\frac{\text{Precision * Recall}}{\text{Precision + Recall}}\right), \tag{9}$$

where,

1. TP (True positive): Number of COVID-19 patients that are properly classified.
2. FP (False positive): Number of COVID-19 patients that are wrongly classified as Flu.
3. TN (True negative): Number of Flu patients that are properly classified.
4. FN (False negative): Number of Flu patients that are wrongly classified as COVID-19.

The values of these metrics for *K*NNV and some related algorithm are reported in Table 3. The Table shows vividly that *K*NNV achieves better results than the related algorithms. In view of the equations of the four evaluation metrics, this indicates that it has high values for both TP and TN, and low values for both FP and FN. This is due to the careful calculation of the distances in the *K*NNV in the presence of a percentage of missing values. These results are reaffirmed in Tables 4 and 5 which outline the best and worst values for the four evaluation metrics for *K*NNV and the related algorithms. Over the 10 runs made, *K*NNV achieved the best results, reason enough to conclude that *K*NNV can accurately identify COVID-19 cases even when the data is both heterogeneous and incomplete.

4.3 Experiment 2: Feature significance

To better understand COVID-19, this experiment is dedicated to investigate the impact of each individual feature on the its classification. In particular, the classification accuracy is computed for each feature independently in the aim to rank its impact on the classification decision. While testing numerical features, $\theta_{\mathscr{C},u_0} = \emptyset$, and therefore we set $K = 1$.

Table 6 Mean classification accuracy of *K*NNV for each feature separately.

	Feature	Accuracy	Rank
a_1	Age	0.50	9
a_2	Gender	0.52	5
a_3	Fever	0.59	2
a_4	Dyspnea	0.46	13
a_5	Nasal	0.65	1
a_6	Cough	0.47	12
a_7	Partial pressure of oxygen (PO2)	0.48	11
a_8	C-reactive protein (CRP)	0.52	6
a_9	Asthenia	0.54	3
a_{10}	Leukopenia	0.53	4
a_{11}	Exposure to COVID-19 patients	0.52	7
a_{12}	Coming from high risk zone	0.52	8
a_{13}	Temperature	0.44	14
a_{14}	Blood test	0.56	15
a_{15}	Polymerase chain reaction (RT-PCR)	0.49	10
a_{16}	Medical History	0.53	16

Table 6 shows the mean classification accuracy of *K*NNV with each feature over 10 runs. A look at the Table shows that Nasal has the highest classification impact. This is reasonable as COVID-19 spreads mainly by droplets produced when people cough, sneeze or talk. Fever, Asthenia and Leukopenia come in second, third and fourth, respectively. This agrees with what the world health organization (WHO) states in its report [33]: at the beginning, the symptoms of COVID-19 are similar to those of a Flu. Also, as per this report, if we are not sure about the Nasal, Fever, Asthenia and Leukopenia features, we should look at the patient age. If the patient is old, having a weak immunity system, he/she is likely to be COVID-19 positive and should undertake other tests like RT-PCR and CRP. Indeed, Table 6 shows that the Age, RT-PCR and CRP features come after Nasal, Fever, Asthenia and Leukopenia features. It also shows that blood analysis and medical history features have the least impact on COVID-19 classification, agreeing with the mentioned WHO report which does not even mention blood analysis and medical history as relevant in diagnosing COVID-19.

5 Conclusions

The *K*NNV algorithm proposed in this article is designed principally to classify COVID-19 using IHC data. The key edge of the algorithm is that it inherits the merits of *K*NN while computing different K values for each unknown patient independently. Additionally, efficient calculations for the distances between patients in employed. This edge has greatly improved the classification accuracy. To assess the algorithm performance, the proposed *K*NNV algorithm is used and applied for the classification of COVID-19 and its performance is compared with three algorithms over a publicly available IHC dataset. The results show that the proposed algorithm outperforms the related algorithms in terms of four metrics: precision, recall, accuracy, and F-Score. All four metrics show that *K*NNV is far better than the related algorithms.

Declarations

Funding Not applicable.

Conflicts of interest/Competing interests Not applicable.
Availability of data and material Available upon request.
Code availability Available upon request.

References

1. World Health Organization (2020) Coronavirus disease 2019 (COVID-19): situation report, 72.
2. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Du B (2020) Clinical characteristics of coronavirus disease 2019 in China. New England Journal of Medicine. https://www.nejm.org/doi/10.1056/NEJMoa2002032
3. Cao J, Tu WJ, Cheng W, Yu L, Liu YK, Hu X, Liu Q (2020) Clinical Features and Short-term Outcomes of 102 Patients with Corona Virus Disease 2019 in Wuhan, China. Clinical Infectious Diseases. https://doi.org/10.1093/cid/ciaa243
4. Li K, Fang Y, Li W, Pan C, Qin P, Zhong Y, Li S (2020) CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19). European Radiology 1–10. https://doi.org/10.1007/s00330-020-06817-6
5. Chowdhury ME, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, Reaz MBI (2020) Can AI help in screening Viral and COVID-19 pneumonia?. arXiv preprint arXiv:2003.13145.
6. Wang Q, Qian Y, Liang X, Guo Q, Liang J (2018) Local neighborhood rough set. Knowledge-Based Systems 153:53–64. https://doi.org/10.1016/j.knosys.2018.04.023
7. Gautret P, Lagier JC, Parola P, Meddeb L, Mailhe M, Doudier B, Honoré S (2020) Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. International Journal of Antimicrobial Agents, 105949. https://doi.org/10.1016/j.ijantimicag.2020.105949
8. Zhang Y, Wang Y, Liu XY, Mi S, Zhang ML (2020) Large-scale multi-label classification using unknown streaming images. Pattern Recognition, 99, 107100. https://doi.org/10.1016/j.patcog.2019.107100
9. Deng Z, Zhu X, Cheng D, Zong M, Zhang S (2016) Efficient kNN classification algorithm for big data. Neurocomputing 195:143–148. https://doi.org/10.1016/j.neucom.2015.08.112
10. Italian Society of Medical and Intervention Radiology (SIRM). https://www.sirm.org/en/category/articles/covid-19-database/
11. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, Chowell G (2020) Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infectious Disease Modelling, 5:256–263. https://doi.org/10.1016/j.idm.2020.02.002
12. McCall B (2020) COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. The Lancet Digital Health 2(4). https://doi.org/10.1016/S2589-7500(20)30054-6
13. Hu Z, Ge Q, Jin L, Xiong M (2020) Artificial intelligence forecasting of covid-19 in china. arXiv preprint arXiv:2002.07112.
14. Pirouz B, Haghshenas SS, Haghshenas, SS., Piro P (2020) Investigating a Serious Challenge in the Sustainable Development Process: Analysis of Confirmed cases of COVID-19 (New Type of Coronavirus) Through a Binary Classification Using Artificial Intelligence and Regression Analysis. Sustainability 12(6), 2427. https://doi.org/10.3390/su12062427
15. Du S, Wang J, Zhang H, Cui W, Kang Z, Yang T, Yuan Q (2020) Predicting COVID-19 Using Hybrid AI Model. http://dx.doi.org/10.2139/ssrn.3555202
16. Santosh KC (2020) AI-Driven Tools for Coronavirus Outbreak: Need of Active Learning and Cross-Population Train/Test Models on Multitudinal/Multimodal Data. Journal of Medical Systems 44(5):1–5. https://doi.org/10.1007/s10916-020-01562-1
17. Boldog P, Tekeli T, Vizi Z, Dénes A, Bartha FA, Röst G (2020) Risk assessment of novel coronavirus COVID-19 outbreaks outside China. Journal of clinical medicine 9(2):571. https://doi.org/10.3390/jcm9020571
18. Gozes O, Frid-Adar M, Greenspan H, Browning PD, Zhang H, Ji W, Siegel E (2020) Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. arXiv preprint arXiv:2003.05037
19. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Xia L (2020) Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology, 200642. https://doi.org/10.1148/radiol.2020200642
20. Barstugan M, Ozkaya U, Ozturk S (2020) Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods. arXiv preprint arXiv:2003.09424
21. Xu X, Jiang X, Ma C, Du P, Li X, Lv S, Li Y (2020) Deep learning system to screen coronavirus disease 2019 pneumonia. arXiv preprint arXiv:2002.09334.

22. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J, Xu B (2020) A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). medRxiv. https://doi.org/10.1101/2020.02.14.20023028
23. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Cao K (2020) Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. Radiology, 200905. https://doi.org/10.1148/radiol.2020200905
24. Peng M, Yang J, Shi Q, Ying L, Zhu H, Zhu G, Yan H (2020) Artificial Intelligence Application in COVID-19 Diagnosis and Prediction. http://dx.doi.org/10.2139/ssrn.3541119
25. Rao ASS, Vazquez JA (2020) Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey in the populations when cities/towns are under quarantine. Infection Control & Hospital Epidemiology 1–18. https://doi.org/10.1017/ice.2020.61
26. Maghdid HS, Ghafoor KZ, Sadiq AS, Curran K, Rabie K (2020) A novel ai-enabled framework to diagnose coronavirus covid 19 using smartphone embedded sensors: Design study. arXiv preprint arXiv:2003.07434.
27. Cao T, Yamada K, Unehara M, Suzuki I, Do VN (2017) Rough Set Model in Incomplete Decision Systems. Journal of Advanced Computational Intelligence and Intelligent Informatics 21(7):1221–1231. https://doi.org/10.20965/jaciii.2017.p1221
28. Ayyad SM, Saleh AI, Labib LM (2019) Gene expression cancer classification using modified K-Nearest Neighbors technique. BioSystems 176:41–51. https://doi.org/10.1016/j.biosystems.2018.12.009
29. Cadenas JM, Garrido MC, Martínez R, Muñoz E, Bonissone PP (2018) A fuzzy K-nearest neighbor classifier to deal with imperfect data. Soft Computing, 22(10):3313–3330. https://doi.org/10.1007/s00500-017-2567-x
30. Zhang S (2019) Cost-sensitive KNN classification. Neurocomputing. https://doi.org/10.1016/j.neucom.2018.11.101
31. Goutte C, Gaussier E (2005) A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: European Conference on Information Retrieval (pp. 345-359). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25
32. Influenza Research Database. https://www.fludb.org/brc/home.spg?decorator=influenza.
33. World Health Organization (2020) Laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases: interim guidance, 2 March 2020 (No. WHO/COVID-19/laboratory/2020.4). World Health Organization