Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
A Decision-Level Fusion Method for COVID-19 Patient Health Prediction

Abdu Gumaei\textsuperscript{a,b,*}, Walaa N. Ismail\textsuperscript{c,d}, Md. Rafiul Hassan\textsuperscript{e}, Mohammad Mehedi Hassan\textsuperscript{a,*}, Ebtsam Mohamed\textsuperscript{d}, Abdullah Alelaiwi\textsuperscript{f}, Giancarlo Fortino\textsuperscript{g}

\textsuperscript{a} STC’s Artificial Intelligence Chair, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
\textsuperscript{b} Computer Science Department, Faculty of Applied Sciences, Taiz University, Taiz 8803, Yemen
\textsuperscript{c} Department of Management Information Systems, College of Business Administration, Al Yamamah University, Riyadh 11512, Saudi Arabia
\textsuperscript{d} Faculty of Computers and Information, Minia University, Minia 61519, Egypt
\textsuperscript{e} College of Arts and Sciences, University of Maine at Presque Isle, ME 04769, United States
\textsuperscript{f} College of Medicine, King Saud University, Riyadh 11543, Saudi Arabia
\textsuperscript{g} Department of Informatics, Modeling, Electronics, and Systems, University of Calabria, 87036 Rende, Italy

\section*{Article Info}

Article history:
Received 26 October 2020
Received in revised form 11 August 2021
Accepted 28 October 2021
Available online 8 November 2021

Keywords:
COVID-19
Patient health prediction
Decision fusion
Random forest
Gradient boosting
Extreme gradient boosting

\section*{Abstract}

With the continuous attempts to develop effective machine learning methods, information fusion approaches play an important role in integrating data from multiple sources and improving these methods’ performance. Among the different fusion techniques, decision-level fusion has unique advantages to fuse the decisions of various classifiers and getting an effective outcome. In this paper, we propose a decision-level fusion method that combines three well-calibrated ensemble classifiers, namely, a random forest (RF), gradient boosting (GB), and extreme gradient boosting (XGB) methods. It is used to predict the COVID-19 patient health for early monitoring and efficient treatment. A soft voting technique is used to generate the final decision result from the predictions of these calibrated classifiers. The method uses the COVID-19 patient’s health information, travel demographic, and geographical data to predict the possible outcome of the COVID-19 case, recovered, or death. A different set of experiments is conducted on a public novel Corona Virus 2019 dataset using a different ratio of test sets. The experimental results show that the proposed fusion method achieved an accuracy of 97.24\% and an F1-score of 0.97, which is higher than the current related work that has an accuracy of 94\% and an F1-score 0.86, on 20\% test set taken from the dataset.

© 2021 Elsevier Inc. All rights reserved.

1. Introduction

The novel coronavirus disease (SARS-CoV-2) has led the world to a pandemic situation, and it is still spreading aggressively [1–3]. We are experiencing the second wave of this virus because of its profound infectious nature and continuous spread. It results in a significant disturbance to general well-being and an immense loss of human life [4]. Moreover, because of the administrative guideline for isolation and social distancing, it has a severe adverse effect on the world economy. Thus, it is vital to resist the widespread of the virus and monitors the spread of the new COVID-19 patient cases continually. It is of central significance to distinguish which people are contaminated at as early as possible to restrict the spread of the disease through people contact tracking and isolation.

COVID-19 has set off a monetary downturn in countless nations [5]. Reverse Transcription-Polymerase Chain Reaction (RT-PCR) is considered the best quality test for SARS-CoV-2 recognition [6], which is assumed to have similar characteristics to COVID-19 virus. This test depends on viral nucleic discovery in sputum or nasopharyngeal swab. Even though it has high specificity, it has a few disadvantages. The RT-PCR test is obtrusive and insufficient.

Nonetheless, it experiences a generally low recognition rate in the beginning phases of the disease. SARS-CoV-2 contamination can be tested with an antibody test [6]. However, the titers are distinguishable from the second week of sickness and require venipuncture, which, in blend with very long handling time, makes it less ideal for quick screening of patient health. Hence, there is
a need for an alternative detection method that is efficient and effective for repeated large-scale testing of SARS-CoV-2/COVID-19.

Artificial Intelligence (AI) has been applied to detect, prevent, and predict many health anomalies, including the COVID-19 disease [7]. Numerous approaches have produced excellent results on image classification problems in the presence of large-size categorized data [8]. ImageNet [9], a technique that prevents overfitting on image classification problems has revealed exceptional results when dealing with large-scale categorized data for training. Additionally, convolutional neural networks (e.g., GoogleNet [10], SqueezeNet [11], and inceptionv3 [12]) are most advantageous because their architecture is thoroughly trained from the start to the end, which removes the need for handcrafted feature extractors. Wang et al. [13] applied a CNN-based deep learning approach to computed tomography (CT) images for screening Corona Virus disease. Song et al. [14] identified that an accurate diagnosis could be achieved with the application of deep learning on CT images. They enhanced the well-known ResNet-50 network, which is already a pre-trained network using numerous images to reduce the training time for the application of efficient feature extraction for any image processing related problem. At the same time, Pal et al. [15] utilized the Long Short Term Memory (LSTM) model to characterize the explicit danger of COVID-19 for a specific country, utilizing the patterns and climate information for this country.

To combat the spread of the current pandemic, researchers have started to investigate machine learning (ML) methods to process a wide range of activities, including health organization reports, social tweets, formal news, and media activity [16,17]. Bayes and Valdivieso [18] utilized the Bayesian approach to predict the number of deaths in Peru for the future of 70 days. In [19], the authors implemented the random forest algorithm for detecting the COVID-19 disease using the CT scans. Despite the remarkable ability of the proposed works, they train only one classifier with the collected data that may decrease the classification accuracy, especially with the imbalanced data set of COVID-19.

Despite the remarkable ability of the various CNN architectures and the traditional forecasting models, such as ARIMA [20], regression models [21-23], recurrent neural network (RNN) models [24], and Bayesian approaches [25] they also have some limitations, in particular, (i) they are failed to support long term prediction, which is very helpful in decision making activities for the epidemics; (ii) parameter learning from limited training examples will overfit such models, regardless of how prevention methods are used, (iii) extensive historical data on such individual base model performance is not available.

For pandemic prediction, a more robust option for decision-making and governmental decision planning is mandatory. Existing studies show that decision fusion approaches that incorporate multiple models’ predictions into a combined prediction/classification result have outperformed the single based models [26-28]. Therefore, assembling the extraordinary ability of ML approaches using decision fusion will overcome the mentioned limitations to predict/classify the COVID-19 disease accurately.

In the proposed research, we suggest a deep decision fusion model comprising of random forest, gradient boosting, and extreme gradient boosting classifiers models. Compared to current fine-tuned pretrained models, this approach provides excellent results for COVID-19 health prediction. Such design can be used to build an automated intelligent software that will analyze the dynamics of the COVID-19 outbreak considering the history and implicit correlation-ship among different countries that suffer from the epidemic. Additionally, the fusion of such ML models can help in the understanding of the disease spread, which will improve the predictions of the infection distribution among different countries and allow the simulation of the full epidemic dynamics. We can summarize the list of following contributions in this work:

- We improve the accuracy of COVID-19 patient health prediction by reducing the individual classifier’s overfitting problem by using effective ensemble learning classifiers such as gradient boosting and extreme gradient boosting.
- We propose a robust decision-level fusion method that fuses the decision of random forest, gradient boosting, and extreme gradient boosting classifiers. By exploiting fused classifiers’ diversity, patient health prediction accuracy will be improved, and the method will be more reliable and more robust.
- We select a soft voting technique for obtaining the final fused prediction results because it is recommended for the classifiers, which are substantial, diverse, and well-calibrated as the adopted classifiers.
- We evaluate the proposed method using an extensive set of experiments for 10-fold cross-validation and a different ratio of test examples.
- We analyze and validate the reliability and stability of the proposed method using a statistical t-test measurement.
- We compare the accuracy of the proposed decision-level fusion method with the accuracy of the current related work.

2. Literature review

The healthcare expert can use radiology images like X-ray and CT scans as routine tools to diagnose and monitor the patient who tested positive for COVID19 disease. Due to the high outburst of COVID-19 Pandemic, the performance of such devices is very limited. Consequently, there is increased attention in investigating the potential of applying deep learning ML and AI models for discovering new ways for robust forecasting of COVID-19 activity [16,17]. The researcher in [29] suggested an expert system for the COVID-19 disease prediction. The model utilizes deep convolutional networks (CNNs) to classify 108 COVID-19 infected patients using their CT images scan. The resulting accuracy and specificity are 86.27%, 83.33%, respectively. In [30], a new forecasting model is proposed to characterize the severity of COVID-19 disease on the Canadian people and worldwide. The model employs the long short-term memory (LSTM) network to determine the key factors intended for predicting the disease. Karar et al. [31] use the Chest X-Ray scans for pneumonia detection via a Generative Adversarial Network-based fine-tuned model. Wang and Wong [13] found the transfer learning technique very efficient to extract features from X-ray images of pneumonia patients. They identified COVID-19 and normal patients using the extracted features through transfer learning and CNN. Farooq and Hafeez [32] used the same dataset as Wang and Wong [13] to classify COVID-19 from pneumonia patients. Their framework was designed based on CNNs.

A hybrid method is devised in [33] to encourage and facilitate a bit of knowledge for Alzheimer patients to recall their memories. A promising approach is presented in [34] using pre-trained deep learning-based drug-target interaction model to recognize the accessible medications that could be utilized to treat COVID-19 patients. Sujatha et al. [35] authors used linear regression and the Multilayer perceptron and vector auto-regression model method for predicting the widespread of COVID-19 cases. The authors worked over the epidemiological trends collected from KAGGLE resources. The Sujatha et al. [36] researchers decompose the expected COVID-19 client perspectives in two parts; the first is eHealth applications in China. The second one is the eHealth Framework in Ukraine, which provided a new perceptive and improvement of an eHealth application. Sethy et al. [37] extracted features using pre-trained models from ImageNet and then used these features as input to SVM for identifying COVID-19 positive and negative cases.

The machine learning techniques have been used in many applications [38-41] and recently gained attention for the COVID-
19 outbreak spreading prediction [42,43] due to its generalization ability and speed. A two-stage estimation telediagnosis and telemonitoring systems for Parkinson’s disease (PD) is proposed in [44]. The model starts by selecting the significant feature sets by applying a wrapper-based feature selection algorithm, called Boruta. Then the significant features are feed into a decision tree-based boosting algorithm. In [19], a random forest model is used to process symptoms, travel history, and patient history to predict new Covidien cases. The researcher in [45] uses data-driven ensemble forecast modeling to guide caregivers in selecting and identifying suitable sites for COVID-19 vaccine efficacy trials. Ensemble approaches are developed by fusing shared targets models and data formats, then the final classification/prediction decision is constructed and evaluated [46,47]. The ensemble model demonstrates outstanding robustness for complex classification tasks like influenza [48] and dengue [49], and more significantly in the area of the disease diagnosis/prediction, and drugs/vaccine discovery [50]. In [50], the authors characterize the COVID symptoms from textual clinical reports using classical and ensemble machine learning algorithms. Ensemble learning techniques have been used to classify the collected data into four classes COVID, SARS, ARDS, and Both (COVID, ARDS). Logistic regression and Multinomial Naive Bayes showed better results than other ML algorithms by having 96.2% testing accuracy. In the future, recurrent neural networks can be used for better accuracy. The work in [38] initiated a novel model for forecasting and predicting the pandemic on a range of 1-3 to 6 days ahead. The model uses the stack-ensemble method with support vector regression algorithm for predicting 10 COVID-19 Brazilian patients. In [51], a shared forecasting ensemble model for COVID-19 is proposed for predicting the U.S. deaths attributable to COVID-19, called the COVID-19 Forecast Hub [https://covid19forecasthub.org/]. Recently, the researchers in [39] initialize a novel classifier using an extreme gradient boosting (XGBoost) algorithm. A multilayered decision rules are built recursively from clinical and mammographic factor datasets. The model classifies the infected patients into four classes including that have a high risk or that need an intensive care to reduce the death rate.

For COVID-19 classification, the majority of approaches rely on deep learning, fine-tuning, and ensemble learning. Although the COVID-19 classification methods have shown encouraging results on modest datasets, there is no guarantee that these designs will yield similar results on a more extensive unbalanced set of data. Moreover, in a comparison of the performance of various CNN architectures, there is substantial improvement in classification accuracy with decision fusion as opposed to a single deep architecture. In the majority of approaches, decision-fusion is not used to predict/classify the pandemic (see Table 1). Therefore, in the current study, we propose a robust decision-level fusion method that combines the prediction results of random forest, gradient boosting, and extreme gradient boosting classifiers. The proposed method exploits the diversity of these three calibrated classifiers to improve further the accuracy results and make the prediction of patient health status is more reliable than relying on a decision of individual methods.

### 3. Materials and methods

#### 3.1. Dataset

The World Health Organization and John Hopkins University [58] have compiled this dataset, and it has been pre-processed further to fit well with the current study needs. Table 2 presents the features of the data. The list of these features includes symptoms such as fever, cough, cold, fatigue, body pain, and malaise along with a set of demographic features such as location, country, vis_wuhan, from_wuhan, (hosp_vis—sym_on), age, gender, which can be used to classify the health risk of COVID-19 patients. In the dataset, there are two classes/health status for each patient: dead or recovered. The dataset contains numeric values in a different format (i.e., Date, String, and Numeric) and categorical variables (i.e., Yes/No, NA).

#### 3.2. Research approach

Our proposed approach consists of several steps that employ and fuse several methods together, as described in the following subsection. Fig. 1 illustrates the flowchart of the approach. It starts with the pre-processing step to train multiple ML classification methods, including a random forest (RF) method, a gradient boosting (GB) method, and an extreme gradient boosting (XGB) method.

#### 3.2.1. Data pre-processing

In this step, we process the dataset to convert categorical variables such as date and string to numeric values. The ML methods used in our approach require the input data to be in numerical format. The well-known label-encoding technique is used in the data pre-processing step to enumerate string and date values of non-numerical/categorical columns of the dataset. The patient id column is removed from the dataset at this stage of data pre-processing because it does not have any beneficial value in the prediction task. In addition, this step replaces all the missing values that have “NA” with zero to prevent from causing errors when they are directly passed as inputs. Since the data columns are not directly used, feature engineering has been applied. To provide the number of duration of days between the date of COVID-19 symptom being noticed and the date that the patient being visited the hospital, a new column has been added with the value of the difference between these two dates (hosp_vis—sym_on).

#### 3.2.2. Prediction using random forest method

Random forest (RF) is one of the popular ML techniques that add an additional layer of randomness through bagging [59]. It
works by constructing a tree using a different bootstrap sample of the data. Besides, random forests can change how the classification is constructed. The RF classifier [59] consists of multiple decision trees (ntree). In our proposed approach the RF algorithm identifies the relationship between the input features of COVID-19 dataset and corresponding health status. In this process, the algorithm first divides the dataset into a number of (n) smaller subsets. For each of the subsets, it will add branches to the tree simultaneously following the decision tree induction method [59], and finally, the final tree will be generated. The leaf node, which is the target value of the generated tree, contains the decision value as either dead or recovered. The decision trees generated from the subsets are represented as $h_1(x), h_2(x) \ldots h_k(x)$ where the training data is represented as vectors $<X, Y>$. 

The margin function $(mg(X, Y))$ for each decision tree $h_k(x)$ is expressed as follows:

$$mg(X, Y) = \text{avg}_k I(h_k(X) = Y) - \max_{j \neq Y} \text{avg}_k I(h_k(X) = j)$$  \hspace{2cm} (1)$$

The generalization error for each of the trees with probability over the $X, Y$ space, is given as follows:

$$P\text{E}^* = P_{X, Y}(mg(X, Y) < 0)$$  \hspace{2cm} (2)$$

In random forests, the number of classifiers increases with about $h_k(X) = h(X, \theta k)$ for all trees sequences, therefore the prediction accuracy of each classifier from the strong law of large numbers and tree structure, with probability PE* converges, defined as follows:

$$P_{X, Y} = P_\theta(h(X - \theta) = Y) - \max_{j \neq Y} P_\theta(h(X - \theta) = j)$$  \hspace{2cm} (3)$$

The number $n$ of decision trees generated above creates a forest. To produce a final outcome from the forest or number of decision trees, it sums up the various decisions to decide the final class of the object as described in the following pseudocode.

---

**Table 2**

A description of the dataset and its features.

| Column  | Description                      | Values (for categorical variables) | Type               |
|---------|----------------------------------|------------------------------------|--------------------|
| id      | Patient Id                       | NA                                 | Numeric            |
| location| The Patient's location           | NA                                 | Date               |
| country | Patient’s native country         | NA                                 | String, Categorical|
| gender  | Patient’s gender                 | NA                                 | String, Categorical|
| age     | Patient’s age                    | NA                                 | Numeric            |
| sym_on  | The symptoms start date of patient | NA                              | Date               |
| hosp_vis| The hospital visiting date       | NA                                 | Numeric, Categorical|
| from_wuhan| The patient track history to be in Wuhan, China or not | Yes (1), No (0) | Numeric, Categorical|
| from_wuhan| The patient is Wuhan, China citizen or not | Yes (1), No (0) | Numeric, Categorical|
| death   | The patient health severity (passed away or not) | Yes (1), No (0) | Numeric, Categorical|
| recovered| Whether the patient recovered     | Multiple symptoms noticed by the patients | Numeric, Categorical|
The random forests algorithm:
Input: the COVID-19 dataset after preprocessing.
Output: The classification decision
Begin
Step1: Draw bootstrap samples from vectors X, Y.
For each of the bootstrap samples
Build the decision node represented by the patient feature value (including patient demographics, patient symptoms, etc.),
Build the leaf node with the patient’s prospective condition value.
For Each node
Grow an unpruned classification,
Choose the best split from among those variables.
End
Create k forest from the selected data as h1(x), h2(x)…hk(x);
End
For Each tree i in k
Sums up the 1 votes;
Aggregate the predictions values of the k trees.
Return predicted values.
End

3.2.3. Prediction using gradient boosting method

Ensemble learning is a de facto model for sample classification, in which many classifiers are being used to efficiently classify the data and aggregate their results. Boosting [60] resolves the classification problems by combining several individual predictors (models) to produce combined predictors. The overall errors of the predictors will be minimized by assigning extra weight to samples incorrectly predicted by previous predictors. In the end, a weighted vote is taken for prediction. One example of these methods is the gradient boosting ensemble AdaBoost [61], which aggregates the prediction of multiple trees to improve the weak classification performance of single tree models. Moreover, gradient boosting has a significant classification performance on imbalanced datasets [62]. Therefore, we will use it to improve the classification accuracy of patient state prediction of the imbalanced COVID-19 dataset. This algorithm greedily builds decision trees from the training dataset samples. The gradient boosting works as follows:

Input: A training sample instance dataset with (x1, y1), …(xn, yn) where
x1 ∈ X, y1 ∈ Y, a trained random forest tree RFTree using distribution D1.
Output: The final decision tree.
For each iteration from t = 1 to T.
Initialize Dt,i = 1/T;
Calculate the hypothesis ht : X → {−1, +1} using
Calculate et = 1/T * [yi/h(t)(xi) ≠ yi];
Select a random subsample of the training data with αt = ln(1−εt)/εt.
Update the model using the selected subsample with
Dt,i+1 = exp(αt ht(xi)) / Zt for h(t)(xi) = yti et
Where Zt represents a normalization factor.
Greedyly grow a tree f(t)(x);
Add f(t)(x) to the model.
End.

After each boosting iteration, the sample that has been hard to classify in the previous steps will be adjusted. Therefore, the misclassified examples weights will be increased, and correctly classified examples weights decreased. Hence, each successive classifier focuses on examples. After a number of iterations, the final series of random modified trees are combined by a weighted majority vote into a final prediction.

3.2.4. Prediction using extreme gradient boosting method

Friedman [63] comes up with a gradient boosting decision tree (GBDT), an iterative algorithm for decision trees. Extreme gradient boosting (XGB) yields a model with hundreds or even thousands of trees. For each iteration, the XGB can provide deep insight into the collected samples, and their predictive powers surpass the prediction powers of most conventional methods. Moreover, the XGB method can implement distributed and parallel computing flexibly [64]. Therefore, in our current work, XGBoost is utilized to prevent model overfitting over imbalanced data and make a general prediction for COVID19 pandemic disease. XGB consists of two parts: the regression tree and the gradient boosting, with appropriate trimming. The approach utilizes a sequential tree growing mechanism, like in boosting with each tree gradient down by learning from the k-1 number of regression tree. Finally, after a set of iterations, the XGB’s loss function adds additional regularization term, which helps to smooth the final learned weights to avoid over-fitting of the aggregation process [64]. The estimated output ŷ of XGB tree model over Dt space can be expressed as the sum of the prediction score of all trees, as follows:

\[ ŷ = \sum_{k=1}^{k} f_k(x_i, y_i), f_k \in D_t \]

For the training sample instances (x1, y1), …(xm, ym) where x1 ∈ X, y1 ∈ Y, the leaf weight (prediction score) is f_k(x_i, y_i), for each leaf node j.

3.2.5. Proposed decision-level fusion method

The decision-level fusion method developed in our approach is a meta-method classifier that combines practically different ML classifiers for classification and prediction through a soft voting technique [65]. The soft voting technique predicts the class label based on the predicted probabilities of each classifier. This technique is recommended if adopted classifiers are diverse and well-calibrated [66].

\[ ŷ = \arg \max_l \sum_{j=1}^{m} w_j p_{l,j} \]

Where w_j is a weight that can be given to determine the contribution of each classifier and p_{l,j} represents the predicted probability of the class label l and the classifier j.

Suppose we formulate a COVID-19 patient health prediction task as a binary classification with a class label l ∈ {0, 1}, where class number zero means the patient will be recovered and the class number 1 means that the outcome of the patient is dead. In addition, we suppose that the probabilities of prediction obtained by three classifiers (C1, C2, C3) to predict an instance x are given as follows:

\[ C_1(x) = [p_{0,1}, p_{1,1}] \]
\[ C_2(x) = [p_{0,2}, p_{1,2}] \]
\[ C_3(x) = [p_{0,3}, p_{1,3}] \]

In our case, C1, C2 and C3 are the RF, GB, and XGB classifiers. By using uniform weights and the default values of classifiers’ parameters, the average probabilities of prediction can be calculated as follows:

\[ p(l = 0, x) = \frac{p_{0,1} + p_{0,2} + p_{0,3}}{3} \]
\[ p(l = 1, x) = \frac{p_{1,1} + p_{1,2} + p_{1,3}}{3} \]

\[ ŷ = \arg \max_l \{ p(l = 0, x), p(l = 1, x) \} \]

The value of ŷ is the final prediction result of the proposed decision-level fusion method.
4. Experiments and discussion

This section presents two experiments with statistical analysis to evaluate and validate the research approach on the dataset described in section 3. The implementation of the approach is coded with Python programming language and the statistical analysis is performed with the statistical SPSS software platform. After preprocessing the dataset, we use it to train the RF, GB, and XGB classifiers on the training set individually. Then, we test the trained models of these classifiers on the testing set to predict the COVID-19 patient’s health status. In the first experiment (referred to as Experiment 1 in the remaining part of the paper), we evaluate the proposed fusion method on 20% of test samples selected randomly from the original dataset. In the second experiment (referred to as Experiment 2 in the remaining part of the paper), we apply a 10-fold cross-validation technique to evaluate the three classification methods and the proposed method. To experimental results of all these methods are compared in terms of several performance metrics such as accuracy, F1-score, recall, and precision. Since the COVID-19 dataset is an imbalanced dataset (Dead: 5.81%, Recovered: 94.19%), F1-score will guide for such an imbalanced dataset to select the best performing method. In the following subsections, the evaluation metrics with the results and statistical analysis will be explained in more detail.

4.1. Evaluation metrics

To predict the COVID-19 of a particular outcome of the patient, many factors/attributes from patient history (ex. travel, demographic, etc.) must be investigated. Since COVID-19 health status prediction is very crucial, and the training data is usually imbalanced, accuracy is not enough to validate the results. Thus, in this study, we considered more than one evaluation metric for evaluating the proposed method’s performance and other existing methods used in this study. The following terms are used in the equations: true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN).

TP and TN: representing a situation whereby the classifier correctly classified positive and negative (“Death” and “Recovered” in our case) instances, respectively. FP: the number of cases where the model misclassified positive as negative. FN: represents patients negative incorrectly classified as positive.

4.1.1. Accuracy

Accuracy is a critical factor in assessing the performance of the classification and prediction model. Given a dataset with (TP+TN) cases, the accuracy can be calculated as the ratio of the total correct classifier prediction to the total number of cases as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

4.1.2. Precision and recall

Precision and recall are two significant metrics to identify the correctly classified patients in an imbalanced class dataset. The precision metric is calculated by the division of the number of true-positive (TP) instances divided by the total number of true-positive (TP) and false-positive (FP) instances, written as presented:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

The recall is an essential factor in identifying correctly classified patients out of all the correctly predicted cases. Recall the ratio of the true-positive (TP) samples to the sum of true-positive (TP) and false-negative (FN) samples. The following equation represents the equation for computing recall provided the values of TP and FN are known.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

4.1.3. F1-score

F1-score is the most significant measure that can be used for the correct evaluation of a method/classifier if the dataset is imbalanced/skewed. Therefore, we emphasize to F1-score to evaluate our proposed method in classifying the imbalanced COVID-19 patients. F1-score is a weighted average of precision and recall; it can be calculated as follows:

\[
\text{F1-score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}
\]

Micro average F1-Score:

For the imbalanced dataset, the Micro Average-sore has been defined as the global F1-score for the whole dataset, i.e., the total true positives and false positives of the model for different classes [67]. To calculate the micro average F1-score, the micro average precision and recall are computed first for each class label from n classes, then the unweighted arithmetic mean is calculated as follows:

\[
\text{Micro avg. precision (MIP)} = \frac{\sum_{i=1}^{n} TP}{\sum_{i=1}^{n} TP + \sum_{i=1}^{n} FP}
\]

\[
\text{Micro avg. recall (MAR)} = \frac{\sum_{i=1}^{n} TP}{\sum_{i=1}^{n} TP + \sum_{i=1}^{n} FN}
\]

\[
\text{Micro avg. F1-score} = \frac{2 \times (\text{MIP} \times \text{MAR})}{\text{MIP} + \text{MAR}}
\]

Macro average F1-Score:

The Macro average score can be used when you want to know how the overall model performance across the sets of data [67]. In this study, we employ the macro average as a performance measure of the different classes. To calculate the macro average F1-score, the macro average precision and recall are computed first for each class label from n classes, then the unweighted arithmetic mean is calculated as follows:

\[
\text{Macro avg. precision (MAP)} = \frac{\sum_{i=1}^{n} PC}{n}
\]

\[
\text{Macro avg. recall (MAR)} = \frac{\sum_{i=1}^{n} RC}{n}
\]

\[
\text{Macro avg. F1-score} = \frac{2 \times (\text{MAP} \times \text{MAR})}{\text{MAP} + \text{MAR}}
\]

Weighted average F1-score:

For evaluating the performance on imbalanced data samples, the weighted average F1-score is used. The number of samples weighs the F1-score from that class for each class [67].

\[
\text{Weighted avg. F1-score} = \frac{(1 + \beta) \times (P \times R)}{(R + (\beta^2 \times P))}
\]

Where 0 < \beta < 1 gives more weight to the Precision, and \beta > 1 gives more weight to the recall.

4.2. Results of experiment 1

In experiment 1, we randomly select 80% of the dataset to train the models, and the remaining 20% is used for testing. Table 3 shows the number of cases for both Recovered and Death classes and their distributions in the training and testing datasets.
Table 3
The number of Recovered and Death cases in the training set (80% of the dataset) and testing set (20% of the dataset).

| Class Name | Number of Training cases | Percentage of Training cases | Number of Testing cases | Percentage of Testing cases |
|------------|--------------------------|------------------------------|-------------------------|-----------------------------|
| Recovered  | 820                      | 94.47%                       | 202                     | 93.09%                      |
| Death      | 48                       | 5.53%                        | 15                      | 6.91%                       |
| Total      | 868                      | 100%                         | 217                     | 100%                        |

Fig. 2. Confusion matrix of Recovered and Death prediction on 20% of the dataset for testing.

Table 4
Results of performance evaluation measures for predicting Recovered and Death classes on 20% of the dataset for testing using the proposed decision-level fusion approach.

| Class Name | Precision | Recall | F1-score |
|------------|-----------|--------|----------|
| Recovered  | 0.976     | 0.995  | 0.985    |
| Death      | 0.909     | 0.667  | 0.769    |
| Micro avg. | 0.972     | 0.972  | 0.972    |
| Macro avg. | 0.942     | 0.831  | 0.877    |
| Weighted avg. | 0.971   | 0.972  | 0.970    |
| Accuracy   | 97.24%    |        |          |

As shown in Table 3, we can see that the distribution of cases in the training dataset is highly imbalanced as there is 94.47% of the total sample represents the Recovered class and 5.53% of the full sample represents the Death class.

At first, we test the efficacy of our proposed method using the abovementioned testing dataset. Fig. 2 shows the confusion matrix for prediction results of experiment 1 in which the TP and TN values are highlighted with a light gray color.

As the Figure shows, our proposed method can correctly predict 201 recovered cases out of 202 and predict 10 death cases out of 15. These corrected cases confirm the method's ability and robustness to mitigate the imbalanced data problem caused by the dominant Recovered class. Table 4 lists the results of precision, recall, and F1-score measures for predicting Recovered and Death classes, as well as the accuracy result of both classes. As we see in Table 4, the proposed method achieves impressive prediction performance with an accuracy of 97.24%. For an imbalanced dataset, the accuracy could be misleading, and hence we look for the F1-score in this regard. As shown in the table, we can see that F1-score has a value of 0.985 for the Recovered class, which is close to 0.99, and a value of for the Death class, which is also close approximately to 0.80.

4.3. Results of experiment 2

After having a very impressive performance by our proposed method for train-test (training: 80% and test: 20% of the whole dataset) validation, we conduct another experiment using a 10-fold cross-validation scheme to compare the adopted methods (RF, GB, and XGB) that have been fused in our proposed method. Tables 5 to 8 show the results of COVID-19 health status classification accuracies and F1-scores for the RF, GB, and XGB methods and the proposed method, respectively.

As shown in Tables 5 to 8, among the performances of individual methods, namely RF, GB, and XGB, respectively, XGB outperforms the other methods. Because the dataset is imbalanced data, the F1-score for the dominant (majority) class (i.e., Recovered) is very high for each of the methods. However, the F1-score for the minority class (i.e., Death) varies from 0.626 to 0.724, where a maximum F1-score of 0.724 is achieved using the XGB method. The micro average F1-scores for each of the methods range from 0.965 (for RF) to 0.972 (for SGB). However, the variation among the macro average F1-scores for each of the methods is very high.
compared to those of micro F1-scores. It should be noted here that, for an imbalanced dataset macro average F1-score indicates a true evaluation of a method/classifier [67]. We notice that RF score underperforms among all the methods in terms of macro F1-score, which is 0.804. XGB provides the best performance among all methods, considering the macro F1-score, which is 0.855. Note that a score close to 1 indicates very good classification performance. A similar performance trend is achieved in terms of weighted average F1-score and accuracy metrics for the methods RF (see Table 5), GB (see Table 6), and XGB (see Table 7), where XGB slightly outperforms the other methods.

Table 8 shows the performance of our proposed method for the COVID-19 health status prediction. As shown in the table, our method can discriminate recovered cases and death cases with an F1-score of 0.986 and 0.748, respectively. This result is 2.4% (0.724 vs. 0.748) higher than that of the best performing method among RF, GB, and XGB for identifying death cases. As mentioned above, the macro average F1-score is an accurate evaluating performance metric for an imbalanced dataset. Our proposed method achieves a macro average F1-score of 0.867, which is close to 1. We notice from Tables 5 to 8, among the individual methods, the XGB achieves a macro average F1-core of 0.855. Thereby, our proposed method is also better than the best individual method in terms of the macro average F1-score (0.867 vs. 0.855; i.e., 1.2% higher than XGB). It should be noted that a single misclassification/mis-prediction would result in death of a patient and hence having an accurate macro average F1-score for such an imbalanced dataset is very important. In that, 1.2% better performance in terms of macro average F1-score can be considered significant considering the dataset is an imbalanced one.

Fig. 4 provides a visual comparison of performances among the individual and the proposed method in terms of accuracy. In contrast, Fig. 5 compares the performance in terms of micro average F1-score, macro average F1-score, and weighted average F1-score. As Figs. 3 and 4 show and Tables 5 to 8 represent, our proposed method performs consistently better than the other individual methods. These better performances are due to considering the weighted importance of each individual method while fusing the outputs of each individual method. Furthermore, one of our major contributions could be the application of XGB in predicting COVID-19 health status. As discussed in Section 2, there is no study applied XGB in the analysis and prediction of COVID-19 patients so far. As shown in Fig. 5, the performance of XGB is significantly better than the other individual methods (e.g., RF and GB). Interestingly, the proposed fusion method is even better than that of the XGB in all performance metrics considered in this study. Such an impressive COVID-19 health status prediction will certainly help professionals and physicians in the prognosis of COVID-19 patients’ health condition if our proposed fusion method or XGB is applied in practice.

### 4.4. Comparison with recent related work

This subsection compares the performance results of this study with that of current related work. Table 9 summarized experimental results of the decision-level fusion method proposed in this work compared with the results of the fine-tuned RF model boosted by the AdaBoost proposed by Iwendi et al. [19], on the same dataset for train-test scheme where 80% of the dataset is used as training data and the rest 20% is used as test set.

As expected, our fusion model’s performance shows a considerable improvement (0.97 vs. 0.86 in terms of F1-score) compared to the recent study’s experimental result on the same dataset. Such an impressive performance in predicting COVID-19 patient health status suggests a possible application of our proposed fusion method to help professionals/physicians in the prognosis of COVID-19 patient. We further emphasize one such efficacy of the fusion method due to the diversity of RF, GB, and XGB classifiers for reducing the effects of imbalanced and overfitting problems during the training process.

For more comparison, a 10-fold cross-validation technique is used on the same dataset to replicate the prediction results of the fine-tuned RF boosted by the AdaBoost method [19]. Table 10 compares the 10-fold cross-validation results of our fusion method with the fine-tuned RF method in terms of accuracy, micro average F1-score, and macro average F1-score.

![Fig. 3. Average accuracy results of the 10-fold cross-validation technique for the proposed decision fusion method compared with its single methods. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)](image-url)
It is illustrated from the 10-fold cross-validation results highlighted in boldface font of Table 10 that the mean of micro, macro, and weighted average F1-score results of our method is higher than the mean of the results fine-tuned RF boosted by AdaBoost method. For deciding whether this improvement is statistically significant or not, independent samples t-test will be performed in the next subsection.

4.5. Statistical analysis of F1-score results

In this subsection, we conduct a statistical analysis of the F1-score results to show whether the improvement achieved by this work against the current related work is statistically significant or not. In other words, we need to answer the research question written as, “Are the micro, macro, and weighted F1-score results of our proposed method statistically significant from the results of the fine-tuned RF boosted by AdaBoost method in a recent related work?” By using a statistical analysis test, the hypotheses of the research question can be formulated as follows:

Null hypothesis: There is no statistically significant difference between the mean of the average F1-score results of our proposed method and the mean of the average F1-score results of the fine-tuned RF boosted by the AdaBoost method.

Alternative hypothesis: There is a statistically significant difference between the mean of the average F1-score results of our proposed method and the mean of the average F1-score results of the fine-tuned RF boosted by the AdaBoost method.

As shown in Table 11, we can see that the mean of the average F1-score results for our method reached a value of 0.937 with a standard deviation of 0.075, which outperforms the mean of the average F1-score results, 0.847, with a standard deviation of 0.162 for the fine-tuned RF with AdaBoost method. The standard errors of these mean results are illustrated in Fig. 5.

Additionally, based on the results in Table 11, we see that the Levene’s test (F) for the equality of variances has a value of 23.7306. The t-test for the equality of means has a value of 2.7660 with a probability value (p-value) of 0.00009, which is smaller than the significance level, 0.05. Therefore, we reject the
null hypothesis and accept the alternative hypothesis, which states a statistically significant difference between the mean of the average F1-score results of our proposed method and the mean of the average F1-score results fine-tuned RF boosted by the AdaBoost method.

5. Conclusions and future work

In the field of medical informatics, ML-based applications play an essential role to analyze patient data, predict the outcome, and give efficient treatment plans. Fusing data from different sources is one of the best methods to improve the accuracy of ML methods in many applications. In this paper, we propose a decision-level fusion method that combines RF, GB, XGB classification methods to predict the COVID-19 patient health for efficient monitoring and early treatment. The final decision result of prediction is computed by combining the predictions of the three methods using a soft voting technique. The COVID-19 patient’s health information, such as travel demographic and geographical data, are used by the adopted classification methods and proposed fusion method to predict the possible outcome of the COVID-19 cases. The potential outcome of the prediction could be recovered or death based on the information of patients. A different set of experiments is conducted on a public novel Corona Virus 2019 dataset using a different ratio of test sets. The experimental results show that the proposed fusion method achieved the highest accuracy than the single prediction models and the current related work. One limitation of using COVID-19 patient information to predict its health is the unavailability of large amounts of data that could be exploited by a deep learning approach for predicting the COVID-19 patient health outcome. Therefore, in the future, we plan to use augmentation techniques to extend the size of the available dataset and study the performance of deep learning methods on it. Besides, we will build a framework to combine X-ray-based vision methods with these types of patient healthcare data to improve the reliability of ML-prediction methods.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgement

The authors are grateful to the Deanship of Scientific Research, King Saud University for funding through Vice Deanship of Scientific Research Chairs.

### References

[1] K. Prem, Y. Liu, T.W. Russell, A.J. Kucharski, R.M. Eggo, N. Davies, S. Flasche, S. Clifford, C.A. Pearson, J.D. Munday, The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study, Lancet Public Health (2020).
[2] Z. Liu, P. Magal, O. Seydi, G. Webb, Understanding unreported cases in the COVID-19 epidemic outbreak in Wuhan, China, and the importance of major public health interventions, Biology 9 (2020) 50.
[3] J. Qazi, K. Naqeeb, A. Qazi, H. AlSalman, U. Naseem, S. Yang, G. Hardaker, A. Gumaei, Evolution to Online Education around the globe during a SARS-CoV-2 Coronavirus (COVID-19) Pandemic: do develop and underdeveloped cope alike?, Child. Youth Serv. Rev. (2020) 105582.
[4] V. Surveillances, The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020, China CDC Wkly. 2 (2020) 113–122.
[5] T.K. Tsang, P. Wu, Y. Lin, E.H. Lau, G.M. Leung, B.J. Cowling, Effect of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China: a modelling study, Lancet Public Health (2020).
[6] S.F. Ahmed, A.A. Quadeer, M.R. McKay, Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies, Viruses 12 (2020) 254.
[7] A. Amini, W. Chen, G. Fortino, Y. Li, Y. Pan, M.D. Wang, Editorial Special Issue on “AI-Driven Informatics, Sensing, Imaging and Big Data Analytics for Fighting the COVID-19 Pandemic”, IEEE J. Biomed. Health Inform. 24 (2020) 2731–2732.
[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.
[10] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size, arXiv preprint, arXiv:1602.07360, 2016.
[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[13] L. Wang, Z.Q. Lin, A. Wong, Covid-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images, Sci. Rep. 10 (2020) 1–12.
[14] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, H. Zhao, Y. Jie, R. Wang, Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images, medRxiv, 2020.
[15] R. Pal, A.A. Sekh, S. Kar, D.K. Prasad, Neural network based country wise risk prediction of COVID-19, arXiv preprint, arXiv:2004.00959, 2020.
[16] S. Lalmanawma, J. Hussain, L. Chakhchhuk, Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review, Chaos Solitons Fractals (2020) 110059.
[17] K.S. Sharov, Creating and applying SIR modified compartmental model for calculation of COVID-19 lockdown efficiency, Chaos Solitons Fractals (2020) 110295.
[18] J. Fu, X. Chen, S. Wu, C. Shi, J. Zhao, J. Xian, Ship detention situation prediction via optimized analytic hierarchy process and naive Bayes model, Math. Probl. Eng. 2020 (2020).
A. Gumaei, W.N. Ismail, Md. Rafiul Hassan et al. Big Data Research 27 (2022) 100287

959–972.

960.

961.

962.

963.

964.

965.

966.

967.

968.

969.

970.

971.

972.

973.

974.

975.

976.

977.

978.

979.

980.

981.

982.

983.

984.

985.

986.

987.

988.

989.

990.

991.

992.

993.

994.

995.

996.

997.

998.

999.

1000.