*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 11*
***Article Received:*** *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

# Prediction of Covid-19 Using Fuzzy-Rough Nearest Neighbor Classification

**K.S.Padmashree[1*], Dr.P.Velmani[2], Dr.S.Loghambal[3]**

**[1]Research Scholar**
Register Number: 20121072292008
PG & Research Department of Mathematics
The M.D.T.Hindu College, Tirunelveli – 627 010
(Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627 012)

**[2]Research Supervisor**
Assistant Professor,Department of Computer Science
The M.D.T.Hindu College, Tirunelveli – 627 010
(Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627 012)

**[3]Research Joint Supervisor**
Assistant Professor,
Department of Mathematics,The M.D.T.Hindu College,Tirunelveli – 627 010
(Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627 012)
***Corresponding Author E-mail:** padmashree.abi@gmail.com

**Abstract**

Prediction refers to the process of using data and statistical or machine learning techniques to estimate or forecast future events or outcomes based on patterns and trends observed in historical data. The goal of prediction is to make accurate forecasts about what is likely to happen in the future, given what is known about past events and trends. The corona virus has created a global pandemic that significantly disrupted our daily schedule and behaviour patterns. Individuals who contract COVID-19 experience a range of symptoms, which can vary in severity. It is crucial to promptly assess the health condition of individuals affected by COVID-19 by evaluating their symptoms and obtaining essential information. . . To assist in this task, physicians rely on rapid and precise Artificial Intelligence (AI) techniques that aid in predicting patients' mortality risk and the severity of their conditions. Early identification of a patient's severity can help conserve hospital resources and prevent patient fatalities by facilitating immediate medical interventions. This research paper introduces an innovative approach that employs the FRNN technique to train a classifier capable of achieving remarkable accuracy in predicting the survival outcomes of COVID-19-affected people. The model is trained on 11 attributes, out of which five are the primary clinical symptoms of this fatal virus: Nasal-Congestion, cough, tiredness, runny nose, fever, sore throat, Diarrhea, and breath shortness, and the other three features are test indication, age, and gender. Our proposed approach, which employs the ENN-SMOTE algorithm to tackle the issue of imbalanced data, demonstrates remarkable effectiveness as evidenced by the experimental results.

**Keywords**--Fuzzy-Rough Nearest Neighbor(FRNN),Edited Nearest Neighbor(ENN),COVID-19

## I. INTRODUCTION

Prediction in machine learning involves utilizing a trained model to make projections about the outcome of future events. These events could be anything from the likelihood of a customer buying a product, to the probability of a disease outbreak. The goal of prediction in ML is to make accurate and reliable predictions based on historical data and patterns. To make accurate predictions, it is important to have access to high-quality data that is relevant to the event or outcome being predicted. This data is often used to construct predictive models, which are mathematical or statistical representations of the patterns and relationships observed in the data. Forecasting the severity risk of diseases in their early stages is crucial due to several beneficial impacts it can have, including reducing mortality rates, optimizing the utilization of hospital resources, and providing valuable decision-making support to healthcare professionals. The global health crisis caused by the novel corona virus has experienced a significant increase in both cases and fatalities globally. According to Johns Hopkins

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 11*
***Article Received:*** *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

University [1], there have been nearly 230 million confirmed cases and 4.7 million deaths reported worldwide. The United States, Brazil, India, France, Russia, and Italy are among the countries most heavily impacted by the crisis.

Currently, computational, mathematical, and surveillance-based methods play a significant role in studying infectious diseases [2]. Machine learning techniques are increasingly used for diagnosing diseases, creating prediction models, and recognising risk factors [3]. Machine learning in healthcare brings numerous benefits, including the ability to significantly improve the accuracy of diagnoses, automate certain responsibilities traditionally handled by radiologists and clinical/anatomic pathologists, and augment the capabilities of healthcare professionals in establishing diagnoses or prognoses [4]. Clinicians seeking to enhance their understanding of personalized patient treatment rely on machine learning as an essential resource. Our research employs a machine learning approach to detect individuals with COVID-19 patients who are at a higher risk of developing illness severely, allowing them to assign priority to their hospitalization accordingly. Adopting this method could potentially decrease the mortality rates of patients and alleviate the strain on healthcare resources.

The medical field heavily utilizes machine learning, as evidenced by various studies such as Abbasi et al. [5], who proposed a new method utilizing machine learning to optimize blood unit transportation in hospital networks resulting in a 29% reduction in average daily cost. Amiri et al. [6] employed machine learning to accurately measure urea, glucose concentrations, and potassium chloride in human blood. Arslan Tuncer and Ayyıldız [7] utilized k-nearest neighbor and support vector machine techniques to distinguish between IDA and Î - thalassemia. Banerjee et al. [8] used ML and statistical testing to improve the initial screening of corona virus-positive cases. This study introduces a novel algorithm that utilizes the Fuzzy Rough Nearest Neighbour classifier for the classification of COVID-19 data. The algorithm in our study predicts the class of a new test instance by calculating the sum of its memberships to the fuzzy-rough lower and upper approximation of each class. The instance is then assigned to the class with the highest sum. The findings of our study yield the following notable contributions:

This paper gives an innovative algorithmic approach for classifying COVID-19 data using the Fuzzy Rough Nearest Neighbour (FRNN) classifier. The FRNN algorithm classifies the new test instance by computing the combined fuzzy-rough lower and upper approximations of every class's memberships. The lower approximation membership represents the absence of the same elements from the opposite class and the upper approximation membership indicates the presence of similar elements from the same class. The test instance is assigned to the class having the higher sum. Therefore, this study presents the following noteworthy contributions:

- Application of the FRNN method for the first time in diagnosing the COVID-19 severity.
- Utilizing symptom data for diagnosing COVID-19 severity.
- FRNN model shows superior performance compared to other evolutionary-based methods.

By using our predictive outcomes, we can control the transmission, reduce the infection rate, and potentially eradicate the existing COVID-19 outbreak.

## II. Related Work

A thorough review of the existing literature on multivariate models and scoring systems employed to predict COVID-19 outcomes has identified 107 research articles that present a total of 145 prognostic models. Furthermore, among these models, 60 utilize radiological techniques to diagnose COVID-19 in individuals exhibiting symptoms indicative of infection. Additionally, 9 models focus on determining the severity of the disease, while 50 models propose prognostic models aimed at forecasting the progression of severe disease, mortality risk, Hospitalization in the intensive care unit, intubation, duration of hospitalization, and ventilation.Notably, this review was conducted up until May 5th, 2020, and offers detailed insights into the current state-of-the-art prediction models based on patient data. Moreover, reference [19] presents a noteworthy methodology that sheds light on the biases inherent in numerous risk prediction models, which can potentially impact the precision of published approaches. However, the study in [19] does not provide a detailed description of the various statistical approaches or machine learning used for prediction.

This study aimed to provide an updated survey on COVID-19 methodologies by analyzing research articles published until 7$^{th}$ October 2020.The focus of our analysis centered on prognostic models for COVID-19 patients, to identify the key processing steps employed in these models. Our research findings indicate that a notable number of the proposed methodologies did not include essential pre-processing steps, such as data standardization/normalization, missing value imputation, or feature selection. This lack of incorporation of these crucial steps could potentially compromise the robustness and performance of the models. While some studies solely presented descriptive statistics derived from univariate [20] or multivariate [21] analyses, the majority of approaches employed logistic regression classifiers [22]–[38]. Additional methods included RF classifiers [25], [26], [30], [31], [36]–[40], XGBoost [41], [42], SVMs [26], [30], [36], [39], [40], K-Nearest Neighbour classifiers [26], [30], Cox regression models [44], [45], and artificial neural networks [45]. It is important to highlight that, apart from a single study that employed a private dataset encompassing 929 COVID-19 patients [46], all other published methodologies were developed and evaluated using datasets of relatively limited sample sizes. This limitation poses a challenge to the

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 11*
***Article Received:*** *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

application of advanced learning models such as neural networks, as they excel in uncovering complex nonlinear relationships, but require larger datasets for optimal performance. Furthermore, the utilization of private datasets in the studies makes it impossible to objectively compare different methods.

## III. Proposed Prediction Model

In this research, the Fuzzy-Rough Nearest Neighbour (FRNN) Classification method was employed to identify COVID-19 diagnosis. To deal with the complexity of the COVID-19 symptoms data, digitalization and discretization were carried out. After processing the data, normalization was performed to ensure that all the values were within the same range. Additionally, feature extraction was carried out to identify the key data characteristics. To address the issue of imbalanced data, an oversampling algorithm was applied. The preprocessed data was subsequently divided into a training set and a test set. The model was trained using the training set and the optimal parameters were determined during this process. Subsequently, the trained model was loaded with the test data for classification purposes, and the performance of the algorithm was assessed by examining the classification results. In Figure 1, a visual representation of the proposed prediction model architecture is presented. The accuracy of predictions is influenced by various factors, such as the quality and relevance of the data, the appropriateness of the analytical techniques used, and the assumptions underlying the predictive models. Overall, prediction is a key tool for decision-making in many domains, and it is an important area of research and development in statistics and machine learning.

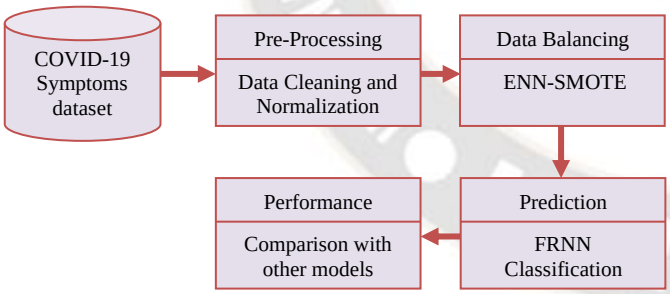


**Figure 1: Framework of the Proposed Prediction Model**

### A) Data Collection

The openly accessible dataset titled "Symptoms and COVID Presence" from Kaggle [10] was utilized in this study. The dataset, which was last updated on 18-08-2020, consists of data collected between 17-04-2020 and 29-08-2020. It comprises 20 features that indicate the existence of different symptoms, along with a class feature indicating whether the individual has COVID-19 or not. In total, the dataset comprises 5434 instances, with 4383 (80.7%) representing COVID-19 patients and 1051 (19.3%) representing healthy individuals. The presence of COVID-19 is indicated as either "Yes" or "No." COVID-19 affects individuals in diverse ways, with infected patients displaying varying symptom severity. In addition to the typical COVID-19 symptoms like fever, dry cough, and shortness of breath, some people have also experienced muscle aches, anosmia (loss of smell or taste), and fatigue [11].

### B) Data pre-processing

Pre-processing the dataset involves several steps, including data cleaning and attributes selection. During attribute selection, insignificant attributes are removed resulting in 11 selected attributes out of the original 14. The input attributes include cough, Tiredness, Nasal-Congestion, Runny-Nose, sore throat, fever, breath shortness, Diarrhea, gender, test indication, and age, while the target attribute is the severity with values of severe, moderate, mild, and none.

To pre-process the data, the min-max method is employed, along with feature normalization, to ensure that the data samples are maintained within the same magnitude range. The data pre-processing step utilizes Formula (1), which involves linear transformation to map the data onto a range of 0 to 1.

$$X^* = (X - \text{Min})/(\text{Max} - \text{Min}). \quad (1)$$

Formula (6) uses X* to represent the pre-processed sample data, while X represents the original sample data. Additionally, Max and Min respectively refer to the maximum and minimum values of the sample data.

### C) Maintenance of data imbalance

The class distributions in many real-world datasets are unbalanced, which can result in incorrect classification performance because most machine learning models perform best when there are nearly equal numbers of examples of each class [15]. The data imbalance, characterized by the dominance of the majority class over the minority class, often leads to classifiers that exhibit a bias towards the majority class. This causes the problem of unbalanced data [16].

To address this issue, various techniques have been developed, including the SMOTE (Synthetic Minority Oversampling Technique) [12]-[14].SMOTE is an oversampling technique commonly used to tackle the issue of imbalanced class distribution in machine learning. It creates synthetic samples for the minority class by interpolating between existing minority class samples. On the other hand, under sampling techniques like Edited Nearest Neighbor (ENN) can be employed to selectively remove some samples from the majority class, resulting in a more balanced dataset. Nevertheless, under sampling methods can potentially remove useful examples that are crucial in the learning process. In addition, in scenarios where the majority class substantially outnumbers the minority class, as evidenced in the heart dataset employed in this study, the efficacy of these techniques might diminish. The creation of duplicates through oversampling can potentially result in over fitting.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 11*
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

| Algorithm 1 SMOTE-ENN Technique |
|---|
| Input: Input data |
| ***Step 1: Oversampling:*** |
| 1: Randomly select a sample xi from the minority class. |
| 2: Find the K nearest neighbors of xi. |
| 3: Create a synthetic sample p by connecting xi to a randomly selected neighbor q, forming a line segment in the feature space. |
| 4: Assign the minority class label to the newly created synthetic sample. |
| 5: Produce additional synthetic samples by taking a convex combination of the two selected samples. |
| **Step 2: Under sampling:** |
| 6: Select a sample xi∈ S, where S denotes the total number of samples xi from the minority class |
| 7: Search for the K nearest neighbors of xi |
| 8: If xi have more neighbors from the other class, then discard xi |
| 9: Repeat 6—8 for all the examples in the dataset. Output: Balanced Covid-19 dataset |
| **Output:** Balanced COVID-19 dataset |

### D) FUZZY-ROUGH NEAREST NEIGHBOUR (FRNN) Classification

The FRNN classification algorithm, as presented in [32], is specifically designed for handling two-class imbalanced data. Its implementation revolves around determining the classification of a new test instance, denoted as x, by evaluating the combined membership value of x to the fuzzy-rough upper and lower approximations of both classes. The instance is then assigned to the class with the highest sum. Let I be an implicator[1],I represent a t-norm which is a binary operator on fuzzy sets. Additionally, let R be a fuzzy relation that signifies an approximation of Indiscernibility between instances. The lower approximation of P and N defines the membership degrees, denoted as P(x) and N(x) respectively, for the element x.

$$\underline{P}(x) = \min_{y \in U} \; I\big(R(x,y), P(y)\big) \quad (2)$$

$$\underline{N}(x) = \min_{y \in U} \; I\big(R(x,y), N(y)\big) \quad (3)$$

Interpretations can be assigned to values $\underline{P}(x)$ and $\underline{N}(x)$, which signifies the absence of objects outside P (belonging to N) that closely resemble x. Similarly, the upper approximation of P and N, under the fuzzy relation R, defines the membership degrees of x as $\overline{P}(x)$ and $\overline{N}(x)$ respectively.

$$\overline{P}(x) = \max_{y \in U} \; I\big(R(x,y), P(y)\big) \quad (4)$$

$$\overline{N}(x) = \max_{y \in U} \; T\big(R(x,y), N(y)\big) \quad (5)$$

[1]An implicatorI I is a $[0,1]^2 \longrightarrow [0,1]$mapping that is decreasing in its first argument and increasing in its second argument, satisfies I (0, 0) =I (0, 1) =I (1, 1) =1 and I (1, 0) =0.

The interpretation $\overline{P}(x)$is that it represents the degree to which there exists another element in P that is close to x and similarly for $\overline{N}(x)$.

In this work, we consider I and T defined by I (a, b)=max(1-a,b) and T (a, b)=min(a, b), for a, b in [0,1]. It can be verified that in this case Equations. (2)– (5) can be simplified to

$$\underline{P}(x) = \min_{y \in N} \; 1 - R(x,y) \quad (6)$$

$$\underline{N}(x) = \min_{y \in P} \; 1 - R(x,y) \quad (7)$$

$$\overline{P}(x) = \max_{y \in P} \; R(x,y) \quad (8)$$

$$\overline{N}(x) = \max_{y \in N} \; R(x,y) \quad (9)$$

To clarify, the value of $\underline{P}(x)$is established based on its resemblance to the nearest negative (majority) sample, while the value of $\underline{N}(x)$is found by its similarity to the closest positive (minority) sample. Meanwhile, for $\overline{P}(x)$and $\overline{N}(x)$, the most comparable element to x within the positive or negative class is identified. It's worth noting that the lower and upper approximations are interconnected, resulting in $\overline{P}(x) = 1 - \underline{N}(x)$ and $\overline{N}(x) = 1 - \underline{P}(x)$. The classification of the test instance x is decided by the FRNN algorithm, which involves the computation of...

$$\mu_P(x) = \frac{\underline{P}(x) + \overline{P}(x)}{2} - \frac{\underline{P}(x) + 1 - \underline{N}(x)}{2} \quad (10)$$

$$\mu_N(x) = \frac{\underline{N}(x) + \overline{N}(x)}{2} - \frac{\underline{N}(x) + 1 - \underline{P}(x)}{2} \quad (11)$$

If $\mu_P(x) \geq \mu_N(x)$,x is classified to the positive class otherwise, it is classified as the negative class.

## IV. EMPIRICAL RESULTS

This part outlines the experimental methodology utilized to validate our proposed approach, which involved using the COVID-19 symptoms dataset. We considered various configurations for the FRNN algorithm, as well as baseline and state-of-the-art methods. Additionally, statistical tests were conducted to facilitate effectiveness assessment. All experiments were conducted in Python using the sklearn package, along with associated libraries for generating reports such as confusion matrix, classification reports, AUC, and ROC curves.

To assess the effectiveness of the models, multiple metrics were utilized including accuracy, sensitivity/recall, precision, and f1 scores. The variable C corresponds to the number of classes, where i∈ C designates a particular class. TP as True Positive, TN as True Negative, FP as False Positive, and FN as False Negative respectively in the analysis.

- ❖ According to (12), the precision score is calculated as

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 11*
***Article Received:*** *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

the proportion of accurately predicted positive instances to the overall predicted positive instances.

$$\text{the Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (12)$$

- The formula (13) is used to determine the Recall/Sensitivity, which represents the proportion of accurate positive predictions to the total no. of observations in a genuine class.

$$\text{Recall}_i = \frac{TP_i}{TP_i + TN_i} \quad (13)$$

- The F1-score, obtained by (14) through the weighted average of Precision and Recall calculations, proves to be more beneficial than accuracy. This is because it considers both false negative and false positive measurements.

$$\text{the Recall}_i = \frac{TP_i}{TP_i + TN_i} \quad (14)$$

- In equation (15), Accuracy refers to the ratio of accurately predicted observations to the overall number of observations.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (15)$$

- It should be noted that these measurements were calculated for every classifier utilized in each experiment.

**B) Performance Analysis**

In this section, our objective is to evaluate the accuracy of a predictive model by employing four algorithms on a COVID-19 dataset and assessing their predictive performance. To determine the most efficient predictive model for optimal outcomes, we employed a ten-fold cross-validation technique with stratification as the testing methodology. We measured the performance of the predictive model using diverse metrics, including classification accuracy; recall (sensitivity), precision, and f-measure, in order to ensure reliable and precise outcomes. The performance measures of various classifiers on the COVID-19 dataset are shown in Table 1.

**Table 1: Performance Comparison of Predictive Models**

| Metric | KNN | NB | SVM | FRNN |
|---|---|---|---|---|
| Accuracy | 0.877 | 0.884 | 0.907 | 0.945 |
| Precision | 0.889 | 0.902 | 0.908 | 0.946 |
| Recall | 0.825 | 0.834 | 0.868 | 0.915 |
| F-Measure | 0.842 | 0.846 | 0.875 | 0.927 |

Table 1 show the proposed FRNN outperforms NB and SVM in terms of prediction performance, with a precision score of 0.946 compared to 0.902 and 0.908, respectively. On the other hand, NB performed the worst with a precision score of 0.889. However, it should be noted that our dataset had highly imbalanced classes, which often led to misclassification of the minority class during training. To address this issue, further experiments were conducted to reduce the class ratio, as described in the next subsection.

**Table 2: Performance Comparison of Predictive Models with ENN-SMOTE**

| Predictive Model | Data Balancing | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|---|
| KNN | ENN-SMOTE | 0.884 | 0.892 | 0.848 | 0.869 |
| NB | ENN-SMOTE | 0.892 | 0.911 | 0.889 | 0.891 |
| SVM | ENN-SMOTE | 0.917 | 0.912 | 0.901 | 0.905 |
| FRNN | ENN-SMOTE | 0.955 | 0.951 | 0.928 | 0.935 |

When the classifiers were applied in combination with ENN-SMOTE, it was observed that the performance of all predictive models consistently improved. Among these models, our proposed FRNN achieved the highest f-measure of 93.5%, next SVM at 90.5%, NB at 89.1%, and KNN at 86.9%. Additionally, it was noted that the ENN-SMOTE approach effectively raised the number of instances in the minority class through iterations and the selection of appropriate k values, thus achieving a balanced dataset in conjunction with the other classes. Figures 2 and 3 depict the actual scores and predictions before and after the ENN-SMOTE application, respectively. With the exception of the minority class, each predictive

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 11 Issue: 11*
***Article Received:*** *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

model's performance demonstrated a noteworthy enhancement in the majority classes.

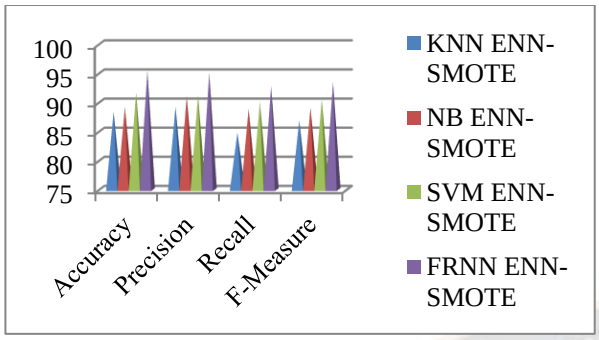


**Figure 2: Comparison of Correctly Classified by class with applied SMOTE**

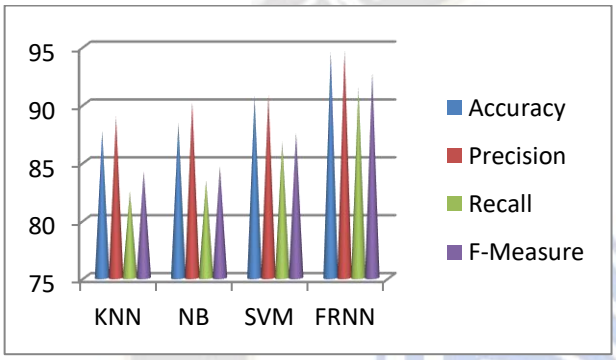


**Figure 3: Comparison of Accurately Classified, without the application of SMOTE**

## V. Conclusion

This study successfully developed an effective FRNN classification model to distinguish the severity of COVID-19 in accordance with the symptoms exhibited by patients. To address the imbalanced training dataset, ENN-SMOTE was employed, and FRNN was utilized for learning and generating the prediction model. The results obtained revealed a significant enhancement in the proposed model, with the highest f-measure of 93.5%. Moreover, a comparative analysis of ENN-SMOTE was conducted to evaluate its performance accuracy in COVID-19 prediction. The results demonstrated that the oversampling approach explored, ENN-SMOTE, consistently improved all predictive models. Overall, our findings provide a practical solution to address the imbalanced classification of COVID-19 prediction by employing data-level strategies.

## References

[1]. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). [Online]. Available: https://coronavirus.jhu.edu

[2]. E. Y. Li, C.-Y. Tung, and S.-H. Chang, ‘‘The wisdom of crowds in action: Forecasting epidemic diseases with a Web-based prediction market system,’’ Int. J. Med. Informat., vol. 92, pp. 35–43, Aug. 2016.

[3]. N. Kimura, Y. Aso, K. Yabuuchi, M. Ishibashi, D. Hori, Y. Sasaki, A. Nakamichi, S. Uesugi, H. Fujioka, S. Iwao, M. Jikumaru, T. Katayama, K. Sumi, A. Eguchi, S. Nonaka, M. Kakumu, and E. Matsubara, ‘‘Modifiable lifestyle factors and cognitive function in older people: A cross-sectional observational study,’’ Frontiers Neurol., vol. 10, p. 401, Apr. 2019.

[4]. Z. Obermeyer and E. J. Emanuel, ‘‘Predicting the future—Big data, machine learning, and clinical medicine,’’ New England J. Med., vol. 375, p. 1216, Sep. 29, 2016.

[5]. B. Abbasi, T. Babaei, Z. Hosseinifard, K. Smith-Miles, and M. Dehghani, ‘‘Predicting solutions of large-scale optimization problems via machine learning: A case study in blood supply chain management,’’ Comput. Oper. Res., vol. 119, Jul. 2020, Art. no. 104941.

[6]. I. S. Amiri, P. Yupapin, B. Mahapatra, S. K. Tripathy, and G. Palai, ‘‘Computation of PUG concentration in human blood using the combination of photonics and machine learning,’’ Optik, vol. 192, Sep. 2019, Art. no. 162968.

[7]. H. Ayyáz and S. Arslan Tuncer, ‘‘Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta-thalassemia via neighborhood component analysis feature selectionbased machine learning,’’ Chemometric Intell. Lab. Syst., vol. 196, Jan. 2020, Art. no. 103886.

[8]. A. Banerjee, S. Ray, B. Vorselaars, J. Kitson, M. Mamalakis, S. Weeks, M. Baker, and L. S. Mackenzie, ‘‘Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population,’’ Int. Immunopharmacol., vol. 86, Sep. 2020, Art. no. 106705.

[9]. A. A. Mousavi, C. Zhang, S. F. Masri, and G. Gholipour, ‘‘Structural damage localization and quantification based on a CEEMDAN Hilbert transform neural network approach: A model steel truss bridge case study,’’ Sensors, vol. 20, no. 5, p. 1271, Feb. 2020.

[10]. Kaggle, “Symptoms and COVID Presence”, 8 January 2022. Available online: https://www.kaggle.com/hemanthhari/symptoms-and-covidpresence

[11]. A. Tsatsakis et al., “SARS-CoV-2 pathophysiology and its clinical implications: An integrative overview of the pharmacotherapeutic management of COVID-19”, Food Chem. Toxicol., vol. 146,pp. 111769, 2020.

[12]. S. F. Abdoh, M. A. Rizka, and F. A. Maghraby, ‘‘Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques,’’ IEEE Access, vol. 6, pp. 59475–59485, 2018.

[13]. A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, ‘‘Improving the prediction of

heart failure Patients' survival using SMOTE and effective data mining techniques,'' IEEE Access, vol. 9, pp. 39707–39716, 2021.

[14]. Asniar, N. U. Maulidevi, and K. Surendro, ''SMOTE-LOF for noise identification in imbalanced data classification,'' J. King Saud Univ. Comput. Inf. Sci., Feb. 2021.

[15]. R. Longadge and S. Dongre, ''Class imbalance problem in data mining review,'' 2013, arXiv:1305.1707. [Online]. Available: http://arxiv.org/ abs/1305.1707

[16]. S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, ''Handling imbalanced datasets: A review,'' GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25–36, 2006

[17]. M. S. K. Inan, R. E. Ulfath, F. I. Alam, F. K. Bappee, and R. Hasan, ''Improved sampling and feature selection to support extreme gradient boosting for PCOS diagnosis,'' in Proc. IEEE 11th Annu. Comput. Commun. Workshop Conf. (CCWC), Jan. 2021, pp. 1046–1050.

[18]. T. Le, M. T. Vo, B. Vo, M. Y. Lee, and S. W. Baik, ''A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction,'' Complexity, vol. 2019, pp. 1–12, Aug. 2019

[19]. L. Wynants et al., ''Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal,'' BMJ, vol. 369, no. 369, 2020. [Online]. Available: https://www.bmj.com/content/ 369/bmj.m1328

[20]. A. Pervaiz, U. Pasha, S. Bashir, R. Arshad, M. Waseem, and O. Qasim, ''Neutrophil to lymphocyte ratio (NLR) can be a predictor of the outcome and the need for mechanical ventilation in patients with COVID-19 in Pakistan,'' Pakistan J. Pathol., vol. 31, no. 2, pp. 38–41, 2020.

[21]. H. Yildiz, J. C. Yombi, and D. Castanares-Zapatero, ''Validation of a risk score to predict patients at risk of critical illness with COVID-19,'' Infectious Diseases, pp. 1–3, Oct. 2020.

[22]. S. Schalekamp, M. Huisman, R. A. van Dijk, M. F. Boomsma, P. J. Freire Jorge, W. S. de Boer, G. J. M. Herder, M. Bonarius, O. A. Groot, E. Jong, A. Schreuder, and C. M. Schaefer-Prokop, ''Model-based prediction of critical illness in hospitalized patients with COVID-19,'' Radiology, to be published, doi: 10.1148/radiol.2020202723.

[23]. X. Bai, C. Fang, Y. Zhou, S. Bai, Z. Liu, Q. Chen, Y. Xu, T. Xia, S. Gong, X. Xie, D. Song, R. Du, C. Zhou, C. Chen, D. Nie, D. Tu, C. Zhang, X. Liu, L. Qin, and W. Chen, ''Predicting COVID-19 malignant progression with Ai techniques,'' medRxiv, 2020.[Online].Available:https://www.medrxiv.org/content /10.1101/2020.03.20.20037325v3

[24]. F. Caramelo, N. Ferreira, and B. Oliveiros, ''Estimation of risk factors for COVID-19 mortality-prelim-i-nary results,'' medRxiv, 2020,doi: 10.1101/2020.02.24.20027268

[25]. J. Xie, D. Hungerford, H. Chen, S. Abrams, S. Li, G. Wang, Y. Wang, H. Kang, L. Bonnett, R. Zheng, X. Li, Z. Tong, B. Du, H. Qiu, and C.-H. Toh, ''Development and external validation of a prognostic multivariable model on admission for hospitalized patients with COVID-19,'' medRxiv, 2020, doi: 10.1101/2020.03.28.20045997.

[26]. X. Qi, Z. Jiang, Q. Yu, C. Shao, H. Zhang, H. Yue, B. Ma, Y. Wang, C. Liu, X. Meng, and S. Huang, ''Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study,''medRxiv,2020.[Online]. Available: https://www.medrxiv.org/content/ 10.1101/2020.02.29.20029603v1

[27]. H. Huang, S. Cai, Y. Li, Y. Li, Y. Fan, L. Li, C. Lei, X. Tang, F. Hu, F. Li, and X. Deng, ''Prognostic factors for COVID-19 pneumonia progression to severe symptoms based on earlier clinical features: A retrospective analysis,'' Frontiers Med., vol. 7, p. 643,Oct.2020.[Online].Available:https://www.frontiersin. org/article/10.3389/fmed.2020.557453

[28]. M. Pourhomayoun and M. Shakibi, ''Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decisionmaking,'' medRxiv, 2020. [Online]. Available: https://europepmc.org/ article/ppr/ppr137985\

[29]. O. Y. Bello-Chavolla, J. P. Bahena-López, N. E. Antonio-Villa, A. Vargas-Vázquez, A. González-Díaz, A. Márquez-Salinas, C. A. Fermín-Martínez, J. J. Naveja, and C. A. Aguilar-Salinas, ''Predicting mortality due to SARS-CoV-2: A mechanistic score relating obesity and diabetes to COVID-19 outcomes in Mexico,'' J. Clin. Endocrinol. Metabolism, vol. 105, no. 8, pp. 2752–2761, Aug. 2020.

[30]. E. Carr et al., ''Supplementing the national early warning score (news2) for anticipating early deterioration among patients with COVID-19 infection,'' medRxiv, 2020.[Online].Available:https://www.medrxiv.org/content /10.1101/2020.04.24.20078006v4

[31]. D. Colombi, F. C. Bodini, M. Petrini, G. Maffi, N. Morelli, G. Milanese, M. Silva, N. Sverzellati, and E. Michieletti, ''Well-aerated lung on admitting chest CT to predict adverse outcome in COVID-19 pneumonia,'' Radiology, vol. 296, no. 2, pp. E86–E96, Aug. 2020.

[32]. A. K. Das, S. Mishra, and S. S. Gopalan, ''Predicting COVID-19 community mortality risk using machine learning and development of an online prognostic tool,'' PeerJ, vol. 8, Sep. 2020, Art. no. e10083. [92] X. Chen and Z. Liu, ''Early prediction of mortality risk among severe COVID-19 patients using machine learning,'' medRxiv, 2020, doi: 10.1101/2020.04.13.20064329.

[33]. Q. Liu, X. Fang, S. Tokuno, U. Chung, X. Chen, X. Dai, X. Liu, F. Xu, B. Wang, and P. Peng, ''Prediction of the clinical outcome of COVID-19 patients using T lymphocyte subsets with 340 cases from Wuhan, China: A retrospective cohort study and a Web visualization tool,'' 2020, doi: 10.2139/ssrn.3557995.

[34]. M. P. McRae, G. W. Simmons, N. J. Christodoulides, Z. Lu, S. K. Kang, D. Fenyo, T. Alcorn, I. P. Dapkins, I. Sharif, D. Vurmaz, S. S. Modak, K. Srinivasan, S. Warhadpande, R. Shrivastav, and J. T. McDevitt, ''Clinical decision support tool and rapid point-of-care platform for determining disease severity in patients with COVID-19,'' Lab Chip, vol. 20, no. 12, pp. 2075–2085, 2020.

[35]. C. V. Guillamet, R. V. Guillamet, A. A. Kramer, P. M. Maurer, G. A. Menke, C. L. Hill, and W. A. Knaus, ''Toward a COVID-19 score-risk assessments and registry,'' medRxiv,2020.[Online].Available:https://www.medrxiv.org/content/10.1101/2020.04.15.20066860v1

[36]. H. Zhang, T. Shi, X. Wu, X. Zhang, K. Wang, D. Bean, R. Dobson, J. T. Teo, J. Sun, P. Zhao, C. Li, K. Dhaliwal, H. Wu, Q. Li, and B. Guthrie, ''Risk prediction for poor outcome and death in hospital in-patients with COVID-19: Derivation in Wuhan, China and external validation in London, UK,'' medRxiv, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/05/03/2020.04.28.20082222

[37]. J. Gong, J. Ou, X. Qiu, Y. Jie, Y. Chen, L. Yuan, J. Cao, M. Tan, W. Xu, F. Zheng, and Y. Shi, ''A Tool to early predict severe coronavirus disease 2019 (COVID-19): A multicenter study using the risk nomogram in Wuhan and Guangdong, China,'' Clin. Infectious Diseases, vol. 71, pp. 833–840, Apr. 2020.

[38]. C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo, ''COVID-19 patient health prediction using boosted random forest algorithm,'' Frontiers Public Health, vol. 8, p. 357, Jul. 2020.

[39]. J. Sarkar and P. Chakrabarti, ''A machine learning model reveals older age and delayed hospitalization as predictors of mortality in patients with COVID-19,'' medRxiv, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.25.20043331v1

[40]. G. Chassagnon et al., ''AI-driven CT-based quantification, staging and short-term outcome prediction of COVID-19 pneumonia,'' 2020, arXiv:2004.12852. [Online]. Available: http://arxiv.org/abs/2004. 12852

[41]. X. Jiang, M. Coffee, A. Bari, J. Wang, X. Jiang, J. Huang, J. Shi, J. Dai, J. Cai, T. Zhang, and Z. Wu, ''Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity,'' Comput., Mater. Continua, vol. 63, pp. 537–551, May 2020.

[42]. L. Yan, H.-T. Zhang, Y. Xiao, M. Wang, C. Sun, J. Liang, S. Li, M. Zhang, Y. Guo, Y. Xiao, and X. Tang, ''Prediction of criticality in patients with severe COVID-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan,'' medRxiv, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v2

[43]. A. Vaid et al., ''Machine learning to predict mortality and critical events in COVID-19 positive new york city patients: A cohort study (preprint),'' J. Med. Internet Res., to be published, doi: 10.2196/24018.

[44]. J. Lu et al., ''ACP risk grade: A simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of the outbreak in Wuhan, China,'' 2020, doi: 10.2139/ssrn.3543603.

[45]. D. Ji, D. Zhang, J. Xu, Z. Chen, T. Yang, P. Zhao, G. Chen, G. Cheng, Y. Wang, J. Bi, L. Tan, G. Lau, and E. Qin, ''Prediction for progression risk in patients with COVID-19 pneumonia: The CALL score,'' Clin. Infectious Diseases, vol. 71, no. 6, pp. 1393–1399, Sep. 2020.

[46]. H. Al-Najjar and N. Al-Rousan, ''A classifier prediction model to predict the status of coronavirus COVID-19 patients in South Korea,'' Eur. Rev. Med. Pharmacol. Sci., vol. 24, no. 6, pp. 3400–3403, 2020.