BOOST MODEL ACCURACY OF IMBALANCED COVID-19 MORTALITY PREDICTION USING GAN-BASED OVERSAMPLING TECHNIQUE

T Bindhu Bhargavi,
India 2021

ABSTRACT

Integration of artificial intelligence (AI) techniques in wireless infrastructure, real-time collection, and processing of end-user devices is now in high demand. It is now superlative to use AI to detect and predict pandemics of a colossal nature. The Coronavirus disease 2019 (COVID-19) pandemic, which originated in Wuhan China, has had disastrous effects on the global community and has overburdened advanced healthcare systems throughout the world. Globally; over 4,063,525 confirmed cases and 282,244 deaths have been recorded as of 11th May 2020, according to the European Centre for Disease Prevention and Control agency. This paper proposes a fine-tuned Random Forest model boosted by the AdaBoost algorithm. The data analysis reveals a positive correlation between patients' gender and deaths, and also indicates that the majority of patients are aged between 20 and 70 years. These datasets have limited samples concerned with the positive COVID-19 cases, which raise the challenge for unbiased learning. The model has an accuracy of 94% and a F1 Score of 0.86 on the dataset used. The data analysis reveals a positive correlation between patients' gender and deaths, and also indicates that the majority of patients are aged between 20 and 70 years.

1 INTRODUCTION

The article covers the use of Generative Adversarial Networks (GAN), an Oversampling technique on real word skewed Covid-19 data in predicting the risk of mortality. This story gives us a better understanding of how data preparation steps like handling imbalanced data will improve our model performance.

The data and the core model for this article are considered from the recent study (July 2020) on "COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm" by Celestine Iwendi, Ali Kashif Bashir, Atharva Peshkar. et al. This study used the Random Forest algorithm boosted by the AdaBoost model and predicted the mortality of individual patients with 94% accuracy. In this article, the same model and model parameters were considered to clearly analyse the improvement of existing model accuracies by using GAN- based Oversampling Technique.

One of the best ways to learn good practices for aspiring Data Scientist would be participating in hackathons on different forums like Analytics Vidhya, Kaggle, or other. In addition, taking the solved cases and data from these forums or published research publications; understand their methodology, and try to improve accuracy or reduce the error with additional steps. This will form a strong basis and enable us to think deeply for the application of additional techniques we learned across the value chain of data science.

2 GENERATIVE ADVERSARIAL NETWORKS (GAN)

Generative adversarial networks are based on a game-theoretic scenario in which the generator network must compete against an adversary. As GAN learns to mimic the distribution of data, it is applied in various fields such as music, video, and natural language, and more recently to imbalanced data problems. Generative adversarial networks are based on a game-theoretic scenario in which the generator network must compete against an adversary. Oversampling based on Generative Adversarial Networks (GAN) over comes the limitations of conventional method such as overfitting, and allows the development of a highly accurate prediction model of imbalanced data.

- Two neural networks compete against each other to learn the target distribution and generate artificial data
  - A generator network G: mimic training samples to fool the discriminator.
  - A discriminator network D: discriminate training samples and generated samples.

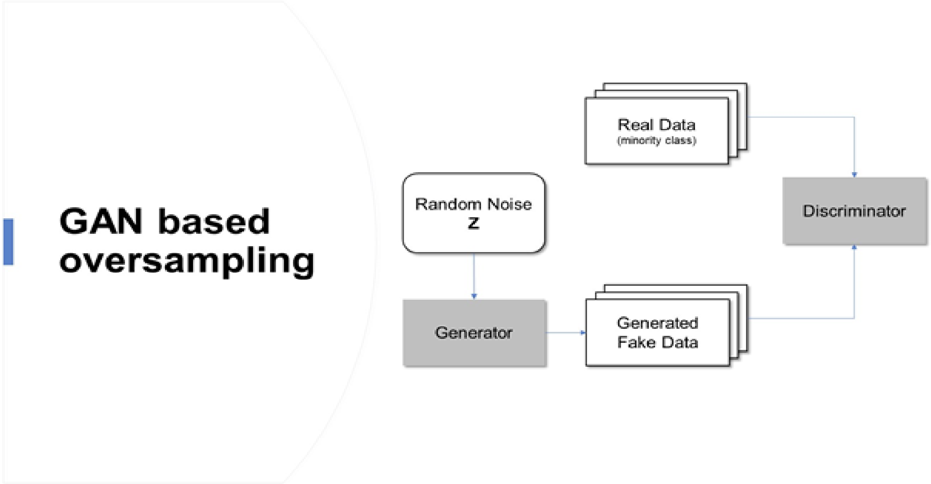
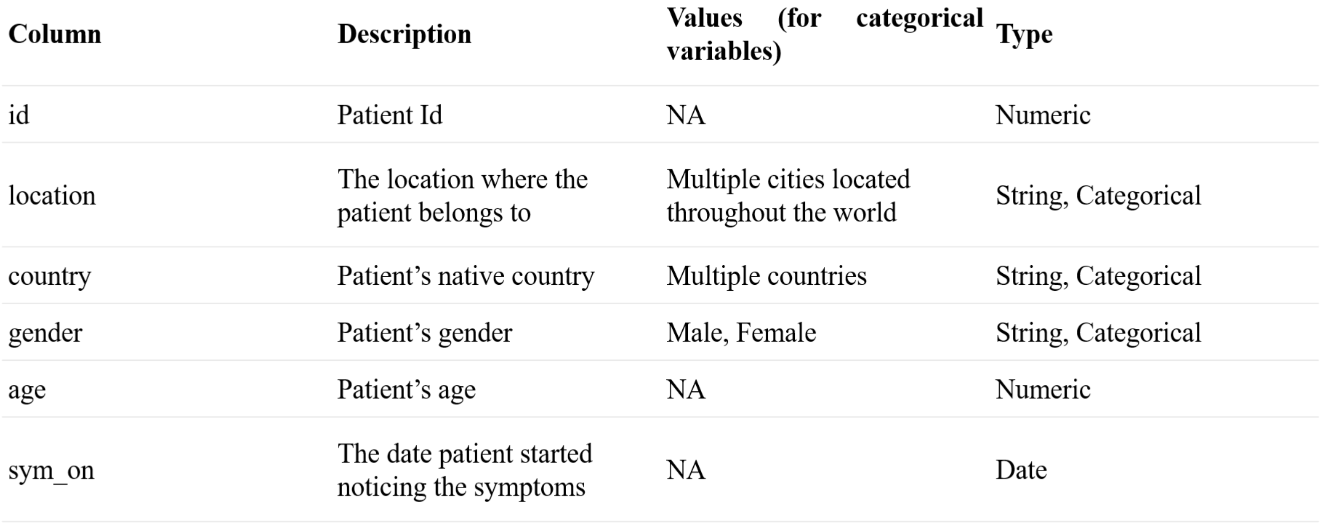


Fig 1. Generative Adversarial Networks

3 DATASETS



Column	Description	Values (for categorical variables)	Type
id	Patient Id	NA	Numeric
location	The location where the patient belongs to	Multiple cities located throughout the world	String, Categorical
country	Patient's native country	Multiple countries	String, Categorical
gender	Patient's gender	Male, Female	String, Categorical
age	Patient's age	NA	Numeric
sym_on	The date patient started noticing the symptoms	NA	Date

3.1 Data Pre-processing

- The dataset consists of columns with the data being the Date, String, and Numeric type. We also have categorical variables in the dataset.
- Since the ML model requires all the data that is passed as input to be in the numeric form, we performed label-encoding of the categorical variables.
- This assigns a number to every unique categorical value in the column.

3.2 Defining Generator

- The generator takes input from latent space and generates new synthetic samples. The leaky rectified linear activation unit (LeakyReLU) is a good practice to use in both the generator and the discriminator model for handling some negative values.
- It is used with the default recommended value of 0.2 and the appropriate weight initializer "It uniforms".
- In the output layer, the SoftMax activation function is used for categorical variables and sigmoid is used for continuous variables.

3.3 Defining Discriminator

- The discriminator model will take a sample from our data, such as a vector, and output a classification prediction as to whether the sample is real or fake.
- This is a binary classification problem, so sigmoid activation is used in the output layer and binary cross-entropy loss function is used in model compilation.
- The Adam optimization algorithm with the learning rate LR of 0.0002 and the recommended beta1 momentum value of 0.5 is used.

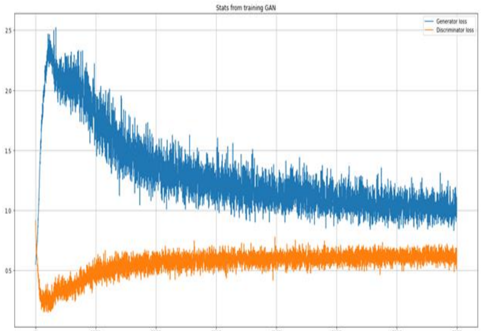


Fig 2: The Trained Model

3.4 Data Analysis

- Fever, cough, cold, fatigue, body pain, and malaise were the most common symptoms that were noticed in patients whose data is available in this dataset.

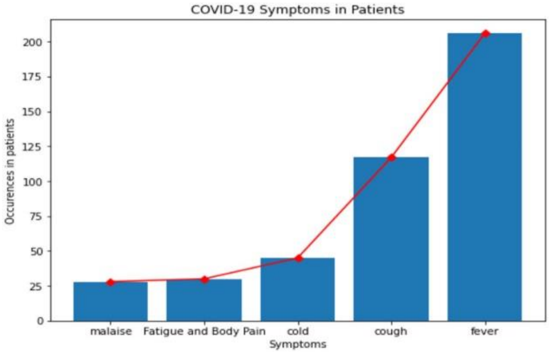


Fig 3: Data Analysis

- Correlation between features of the dataset provides crucial information about the features and the degree of influence they have over the target value.

4 EVALUATION METRICS

The purpose of evaluating the model we considered three evaluation metrics for this study.

4.1 Accuracy

- Given a dataset consisting of (*TP* + *TN*) data points, the accuracy is equal to the ratio of total correct predictions (*TP* + *TN* + *FP* + *FN*) by the classifier to the total data points.

- Accuracy is an important measure which is used to assess the performance of the classification model. Accuracy is calculated as shown in Equation (1) as follows:

- Accuracy = TP + TN / TP + TN + FP + FN (0.0<Accuracy<1.0)

4.2 Precision

- Precision is equal to the ratio of the True Positive (*TP*) samples to the sum of True Positive (*TP*) and False Positive (*FP*) samples.
- Precision is also a key metric to identify the number of correctly classified patients in an imbalanced class dataset.
- Precision = TP / TP + FP.

4.3 Recall

- Recall is equal to the ratio of the True Positive (*TP*) samples to the sum of True Positive (*TP*) and False Negative (*FN*) samples.
- Recall is a significant metric to identify the number of correctly classified patients in an imbalanced class dataset out of all the patients that could have been correctly predicted.
- Recall = TP/TP + FN.

4.4 F1 Score

- F1 Score is equal to the harmonic mean of Recall and Precision value.

- The F1 Score strikes the perfect balance between Precision and Recall thereby providing a correct evaluation of the model's performance in classifying COVID-19 patients.
- This is the most significant measure that we will be using to evaluate the model.
- F1 Score = 2 × Precision × Recall / Precision + Recall.

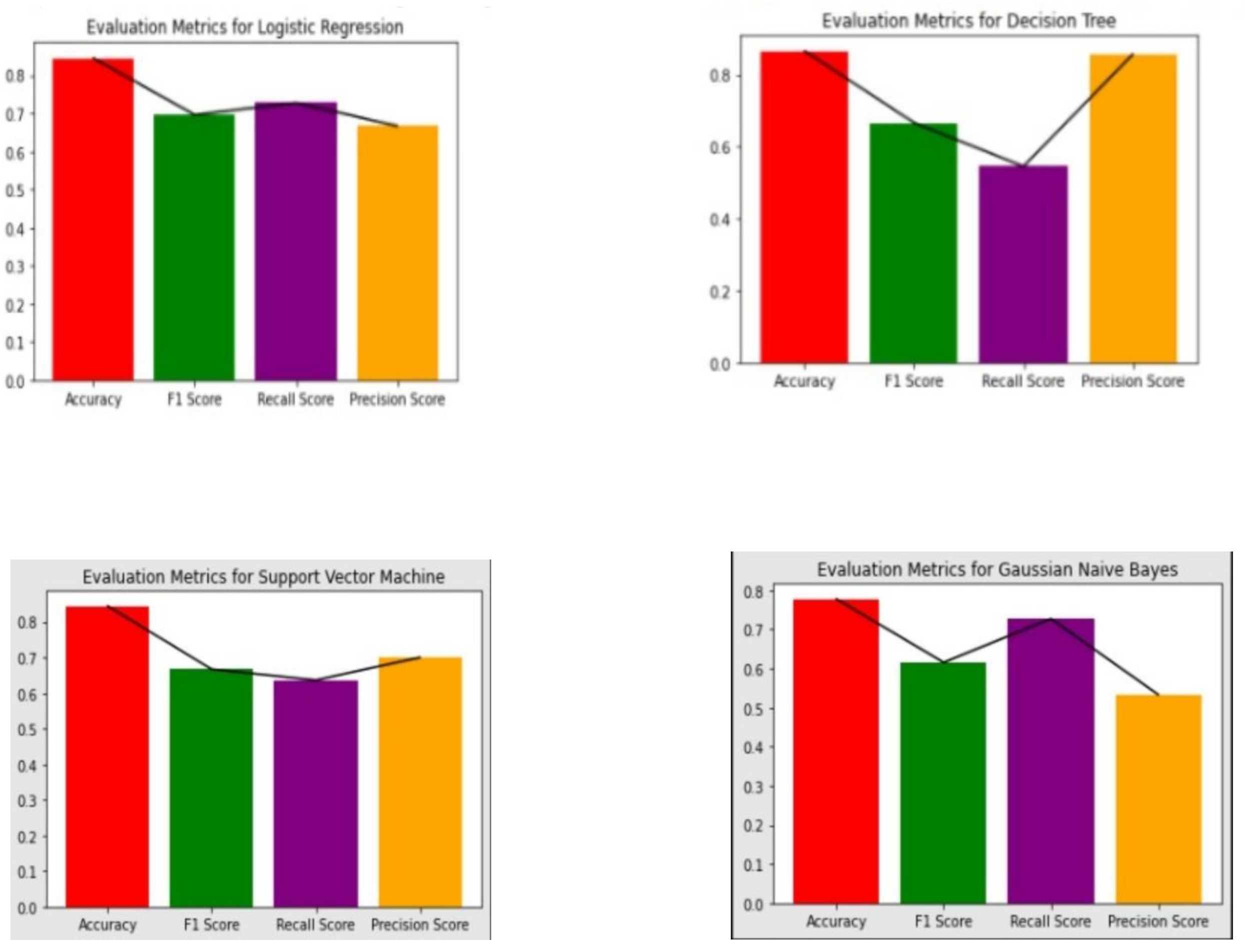


Fig 4: Evaluation Metrics

5 VISUALIZATION

Training the model and getting the results is not sufficient unless it is understood that what is triggering the concerned output. To handle this Blackbox, visualization techniques assist in illustrating the basis of prediction of the model. There are many visualization techniques for example class activation maps (CAM), saliency maps (SM), local interpretable model-agnostic explanations (LIME), and a lot more. In this article, activation maps and LIME techniques are utilized to present the model perception of identifying and classifying the COVID-19 samples from CXR images.

CAM aims at understanding the feature space of an input image that influences the prediction, whereas LIME is an innovative explanation technique to represent the model prediction with local fidelity, interpretability and model agnostic. For instance, fine-tuned NASNetLarge architecture is considered to generate LIME explanations for some samples taken from the test set, whereas class activation maps tend to present the patterns learned by the model for classification of samples as shown in figures presents the LIME technique applied to four samples belonging to four distinct classes as COVID-19, other types of pneumonia, tuberculosis and normal cases. The red and green areas in the LIME generated explanation correspond to the regions that contributed against the predicted class and towards the predicted class respectively.

6 OUTPUT

- The model performance is tested on the actual (original) split test data.
- After splitting the original data into train and test, generated data from GAN is added to the train data to compare the performance with the base model.

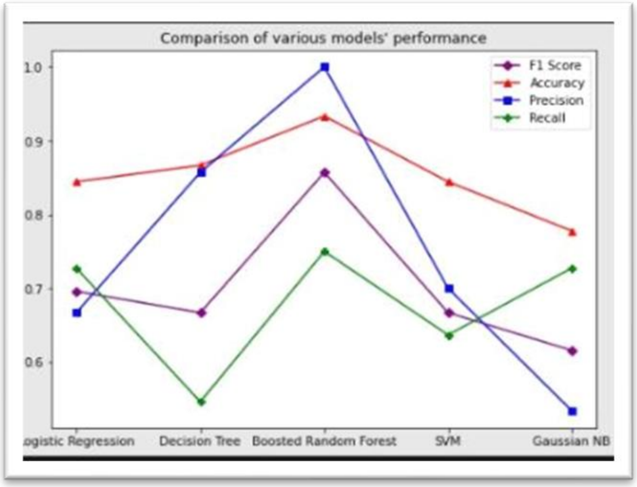
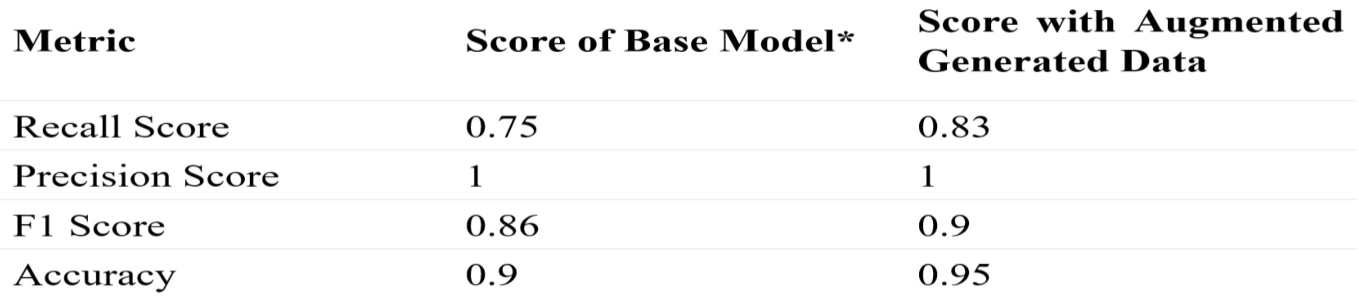


Fig 5: Model Comparison



Metric	Score of Base Model*	Score with Augmented Generated Data
Recall Score	0.75	0.83
Precision Score	1	1
F1 Score	0.86	0.9
Accuracy	0.9	0.95

Table 1. Model Comparison

7 CONCLUSION

The proposed model provides a more accurate and robust result compared to that of the based model, showing that GAN-based oversampling overcomes the limitations of the imbalanced data and it appropriately inflates the minority class. This article proposes to leverage the state-of-the-art deep learning models to aid in early identification and diagnosis of COVID-19 virus by using the limited posteroanterior chest X-ray images. Each trained model was evaluated using benchmark performance metrics e.g., accuracy, precision, recall, area under curve, specificity, and F1 score under four different scenarios concerned with imbalanced learning and classification strategy.

The extensive trials, it was observed that models achieve different scores in different scenarios, among which NASNetLarge displayed better performance specially in binary classification of COVID-19 samples. The visual representation based on local interpretable model agnostic explanations is utilized to understand the basis of prediction of the model. As an extension to this work deeper learning models and pre-processing techniques can be explored to achieve better results.

8 REFERENCES

[1] WHO Situation Report-94 Coronavirus disease 2019 (COVID-19) (2020).

[2] Sujatha R, Chatterjee JM, Hassanien AE. (2020).

[3] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017).

[4] Kathiresan S, Sait ARW, Gupta D, Lakshmanaprabu SK, Pandey HM (2020).] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition.

[5] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks.

[6] Novel Corona Virus 2019 Dataset (accessed April 23, 2020).

[7] Bayes C, Valdivieso L. Modelling death rates due to COVID-19 (May 5, 2020).