CHAPTER 25
A Machine Learning Based Approach Using Complete Blood Count Parameters for Covid-19

Mr. M. MOHAMMED KASIM

Nehru Institute of Engineering and Technology, India

Ms. M. JEBA PAULIN

Nehru Institute of Engineering and Technology, India

Mr. P. PARTHIBAN

Nehru Institute of Engineering and Technology, India

Dr. P. K. MANOJ KUMAR

Nehru arts and science college, India

ABSTRACT

The corona-virus disease is an infectious disease caused by the SARS-CoV-2 virus. Most people who fall sick with COVID-19 will experience low to moderate symptoms and recover without special treatment. However, some will become seriously ill and require medical attention. An ensemble of machine learning algorithms on the basis of complete blood count parameters was developed. The severity of covid-19 was predicted by analysing the patients age and gender using Complete Blood Count parameters. A CBC is a regular blood test taken from the covid affected patients. The analysis the severity is categorized into three levels: low, moderate and high.

Keywords— Covid19, Machine Learning, Ensemble, Decision Tree, Random Forest Classifier, Logistic regression.

INTRODUCTION

Corona-virus disease 2019 (COVID-19) is a contagious disease caused by severe acute respiratory syndrome corona-virus 2 (SARS-CoV-2). The disease has since spread worldwide, leading to an ongoing pandemic. The SARS-CoV-2 corona- virus has infected more than 100 million people and has resulted in almost three million deaths worldwide. To mitigate this pandemic spread, the use of AI techniques to develop tools that is supportive for clinicians have increased. The development of ML models to predict ICU admission or death to stratify patients by risk level, has so far lagged behind and reported as limitations in the existing solutions. To address these limitations, in this work, we report a study aimed at developing Machine Learning (ML) models to predict the disease severity according to the patient's age,

ISBN: 978-93-5782-920-5, www.aetsjournal.com "Handbook of research in Big data Analytics, Artificial intelligence and Machine learning"

gender and CBC parameters. Complete Blood Count - CBC is a feature of routine blood test which is quick enough to get its wide application in a number of diagnostic and monitoring tasks. We present three models, which have been conceived as complementary decision support tools. One model, which is based on the ensembling of 3 models, has been selected for its high accuracy, despite its low clinical interpret-ability. The three models, i.e. a decision tree, a logistic regression and random forest algorithm have been selected because of their explain-ability, despite their lower accuracy with respect to the ensemble model mentioned above. Indeed, these models can provide clinicians with more interpret-able indications that can help them in their decision-making during the management and treatment of COVID- 19 patients.

LITERATURE SURVEY

Logistic regression was carried out to quantify the effects of the covariates on in-hospital mortality [1]. They included in the multivariate models the following covariates: age, gender and laboratory parameters derived from complete blood count. This study was adequately powered to compensate for variations in the mortality rate: for example, with a Neutrophil-to-Lymphocyte (NL) ratio above 7 in 25% of discharged patients and 50% of deceased patients. The estimated power is 0.9999 for a mortality rate of 0.33 (33%) and remains > 0.80 for all populations with mortality rate above 0.06 (6%).Thus, the data strongly support the idea that complete blood count, a routine test for most patients admitted to hospital, might be highly informative with regard to COVID-19 patients admitted to hospital.

Three different training data set of hematochemical values from 1,624 patients (52% COVID-19 positive), admitted at San Raphael Hospital (OSR) from February to May 2020, were used for developing machine learning (ML) models [2]. The complete OSR dataset (72 features: complete blood count

(CBC), biochemical, coagulation, hemogas analysis and CO-Oxymetry values, age, sex and specific symptoms at triage) and two sub-datasets (COVID- specific and CBC dataset, 32 and 21 features respectively) were used. 58 cases (50% COVID-19 positive) from another hospital, and 54 negative patient details, were used for internal-external and external validation.

They developed ML models for the complete OSR dataset, the area under the receiver operating characteristic curve (AUC) for the algorithms ranged from 0.83 to 0.90; For the COVID- specific dataset, it ranged from 0.83 to 0.87 and for the CBC dataset, it ranged from 0.74 to 0.86. The validations also achieved good results: AUC from 0.75 to 0.78 and specificity from 0.92 to 0.96.

Based on the data, a model was developed that predicts COVID-19 test results using eight binary features [3]. They were sex, age, known contact with an infected individual, and five initial clinical symptoms. The training- validation set was divided to training and validation sets at a ratio of 4:1.

Predictions were generated using a gradient-boosting machine model built with decision-tree based learners.Gradient boosting is widely considered as state of the art technique in predicting tabular data and is used by many successful algorithms in the field of machine learning. The missing values were inherently handled by the gradient- boosting predictor.

Polynomial Regression can be expressed as a special case of Linear Regression [4]. Linear Regression works on known continuous data and the two variables (target variable and independent variable) are correlated. If the variables are correlated but the relationship does not look linear, polynomial regression can be used to fit a polynomial equation to our dataset. Polynomial Regression is a supervised machine learning algorithm. They trained based on prior data and then tested the model on another dataset to validate its accuracy. The train and test data have been transformed for polynomial regression. The polynomial Regression algorithm shows an accuracy of 93%.

As Random Forest are a class of probability scoring classifiers (that is, for each instance the model assigns a probability score for every possible class) [5]. The abstention is performed on the basis of two

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

thresholds α, β[0,1]: if we denote with 1 the positive class and 0 the negative class, then each instance is classified as positive if score(1) > α and score(1) > score(0), negative if score(0) > β and score(0) > score(1) and, otherwise, the model abstains. In these models, the performance was evaluated only on the non-abstained instances, and the coverage is a further performance element to be considered.

SYSTEM DESIGN

A dataset of complete blood count parameters has been collected from COVID-19 patients admitted in the hospital. Complete Blood Count - CBC is a routine blood test which is quick enough to get for its wide application in a number of diagnostic and monitoring tasks.

The collected datasets are preprocessed, trained, tested and validated. Using the ensemble of ML algorithms the accuracy is analysed. The severity level is predicted according to age, gender of the patient. Then the admission of intense care unit is predicted for the severe risk.

Dataset

A dataset in collaboration with medical doctors from different COVID-19 patients were collected and were split for training and testing. Figure 1 shows the architecture of the database and analysis. Figure 2 shows the system design for predicting the severity of covid-19 using dataset.

Database

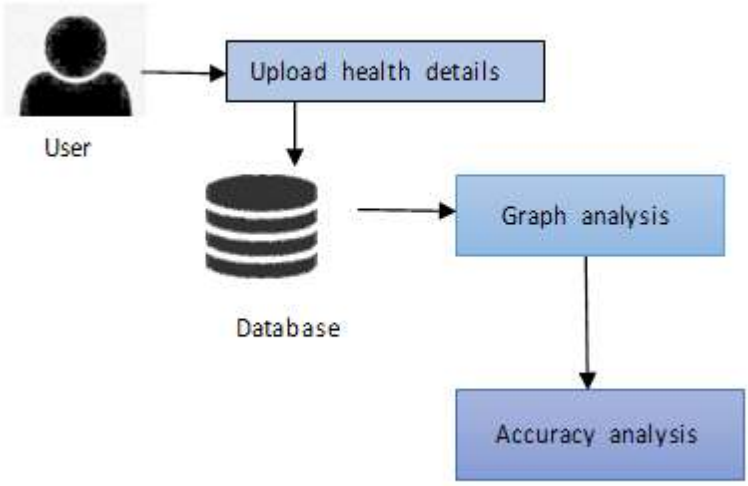

Figure 1: Architecture diagram

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

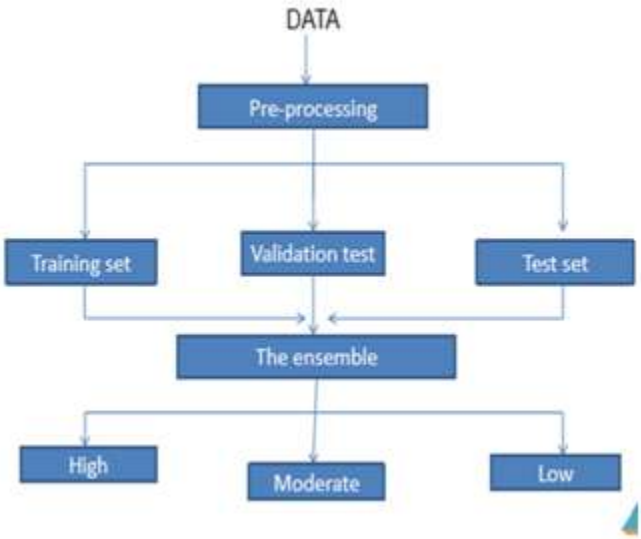

Figure 2: System Design

Preprocessing of Dataset

The pre-processing includes removing the unwanted data from the dataset. In this process, the null values such as missing values are replaced by zero.

Classification

Logistic regression

Logistic regression is one of the most popular machine learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic Regression can be used to classify the different types of observations and can easily determine the most effective variables that are needed for the classification.

Decision tree

Decision tree is a supervised learning technique that can be used for both classification and regression problems; But, mostly it is preferred for solving Classification problems. It is a tree- structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

Random forest algorithm

Random forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both classification and regression problems in ML. It is based on the concept of ensemble technique, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

The Ensemble

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produce more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

Our ensemble model includes

- Logistic regression
- Random forest classifier
- Decision tree

RESULTS AND DISCUSSION

This system is able to predict the severity of the covid affected patients. Using the ensemble of ML models the accuracy is compared. This system improves accuracy, prevent the rate of mortality at earlier stages and thus death rate can be controlled to a certain level. The accuracy of this system is 96%.

Figure 5 shows the accuracy, sensitivity, specificity of the best selected algorithm which is decision tree. Figure 6 shows the pie chart of ICU & NO ICU analysis according to the severity. Figure 7 shows the pie chart of age analysis according to the severity. Figure 8 shows the pie chart of gender analysis according to the severity.

Figure 5: Accuracy, Sensitivity, Specificity of the best selected algorithm which is decision tree.

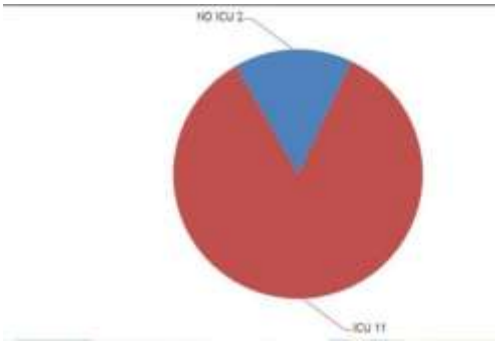

Figure 6: ICU & NO ICU analysis

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

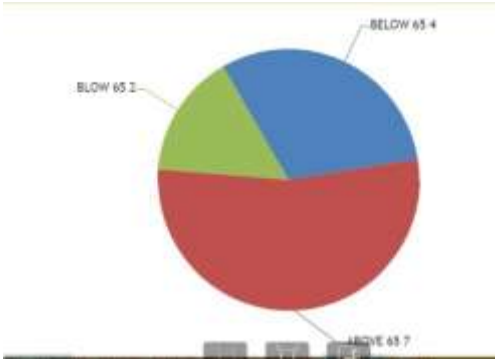

Figure 7: Severity according to age analysis

Figure 8: Severity according to gender analysis

CONCLUSION

In summary, we reported a study to address the challenging task of predicting the severity of a COVID-19, so that the patient can be transferred to ICU during their hospital stay on time. The proposed approach, based on ensemble models, reported good results. Also, our methods are cheap, as they ground mainly on age, gender and CBC test results only. This is the main strength of our approach in light of acceptable accuracy. For this reason, our models can be useful in healthcare facilities which have to manage a surge of ill patients who cannot afford the execution of more COVID-specific exams on a daily basis.

FUTURE ENHANCEMENT

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”

For future work, we aim to externally validate our models with data comingfrom other hospitals and other time periods. This would allow testing the model in light of possible virus mutations and different patient management and therapeutic policies. Since these latter ones depend on the number of cases to deal with and on the continuous advancement of what we know about COVID-19 and its effective treatment (changing its prognosis), phenomena related to concept drift cannot be ruled out in any existing predictive model, including ours.

REFERENCES

Mattia Bellan, Danila Azzolina, Eyal Hayden, Gianluca Gaidano, Mario Pirisi, Antonio Acquaviva, "Simple Parameters from Complete Blood Count Predict In- Hospital Mortality in COVID-19", Disease Markers, vol. 2021, Article ID 8863053, 7 pages, 2021.

F. Cabitza, A. Campagner, D. Ferrari,C. Di Resta, D. Ceriotti, E. Sabetta, A. Colombini, E. De Vecchi, G. Banfifi, M. Locatelli et al., "Development, evaluation, and validation of machine learning models for covid-19 detection based on routine blood tests," Clinical Chemistry and Laboratory Medicine (CCLM), vol. 59, no. 2, 2021.

Shira Deri-Rozov & Noam Shomron, Yazeed Zoabi, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms", npj digital medicine, vol 21, No.3, pp 1-5,2021.

Alankrit Gupta, Ekta Gambhir, Ritika Jain, Uma Tomer, "Regression Analysis of COVID-19 using Machine Learning Algorithms", IEEE, vol-20, no-5, pp 65-71,2020.

Davide Brinati , Andrea Campagner , Davide Ferrari , Massimo Locatelli, Giuseppe Banfi, Federico Cabitza, "Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning", J Med Syst, vol-44, no-8, pp 135, 2020 July 1.

H. S. Yang, Y. Hou, L. V. Vasovic, P. A. Steel, A. Chadburn, S. E. Racine- Brzostek, P. Velu, M. M. Cushing, M. Loda, R. Kaushal et al., "Routine laboratory blood tests predict sars-cov-2 infection using machine learning", Clinical chemistry, vol. 66, no. 11, pp. 1396–1404, 2020.

L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray et al., "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal", bmj, vol. 369, 2020.

E. J. Favaloro and G. Lippi, "Recommendations for minimal laboratory testing panels in patients with covid-19: Potential for prognostic monitoring." in Seminars in Thrombosis and Hemostasis, vol. 46, 2020, pp. 379– 382.

J. Linssen, A. Ermens, M. Berrevoets,M. Seghezzi, G. Previtali, H. Russcher,

A. Verbon, J. Gillis, J. Riedl, E. de Jongh et al., "A novel haemocytometriccovid-19 prognostic score developed and validated in an observational multi centre european hospital-based study," Elife, vol. 9, p. e63195, 2020.

T. Hernandez-Boussard, S. Bozkurt,J. P. Ioannidis, and N. H. Shah, "Minimar (minimum information for medical ai reporting): Developing reporting standards for artificial intelligence in health care," Journal of the American Medical Informatics Association, vol. 27, no. 12, pp. 2011–

2015, 2020.

A. K. Waljee, A. Mukherjee, A. G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, and P. D. Higgins, "Comparison of imputation methods for missing laboratory data in medicine," BMJ open, vol. 3, no. 8, 2013.

ISBN: 978-93-5782-920-5, www.aetsjournal.com "Handbook of research in Big data Analytics, Artificial intelligence and Machine learning"

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” Journal of artificial intelligence research, vol. 16, pp. 321– 357, 2002.

ISBN: 978-93-5782-920-5, www.aetsjournal.com “Handbook of research in Big data Analytics, Artificial intelligence and Machine learning”