Validation and Optimization of supervised machine learning model for rapid COVID-19 Diagnoses using clinical symptoms

Vandana Dutt[1*], Paramjeet Singh[2] and Shaveta Rani[3]

[1,2,3] *Department of Computer Science and Engineering, GZSCCET, MRSPTU, Bathinda*
[*1]*Vandanadutt21@gmail.com*
[2]*param2009@yahoo.com*
[3] *garg_shavy@yahoo.com.*

KEYWORDS

Logistic regression, COVID-19 Diagnosis, Machine Learning, Precision-Recall Curve, ROC Curve

ABSTRACT

This research focuses on the development, validation and optimization of a supervised machine learning trained model for the rapid diagnosis of COVID-19 disease detection using patient's clinical symptoms. Using the logistic regression algorithm, it is a linear model which is known for its robustness and accurate result, we aimed to improve the diagnostic process by improving its capability to handle missing data values from the dataset, assess feature importance, and mitigate overfitting and underfitting problems. The model was instructed and evaluated using a medical dataset from Kaggle, with four performance metrics including accuracy, precision, recall and F1 scores calculated. The results found using LR model are high accuracy (98.5%), with a perfect precision (100%) and a recall of approximately (96.9%), diagnosing the patients effectively by using the model in accurately identifying positive cases. ROC curve analysis revealed a high area under the curve (AUC), confirming the excellent differentiation among positive and negative cases. Also, the Precision-Recall curve illustrated the high precision and robustness of this model. Future work will focus on expanding the dataset, comparing with other different types of models, refining the hyperparameters, and validating the model in real clinical settings to further improve its diagnostic capabilities. This model represents a significant advance in the rapid and accurate detection of COVID-19 infectious disease, with the potential to improve diagnostic efficiency and support public health efforts.

I. INTRODUCTION

In today's digital era, data analytics continuously focusing on creating automated models that learn from existed or collected data and past experience with minimal human involvement. Machine learning has gained remarkable insights over time, particularly through the analysis and evaluation of epidemiological datasets. There is a progressive trend in the healthcare sector to store patient's clinical data for future reference, enabling the development of efficient automated systems which are capable of diagnosing, categorizing and identifying anomalies. Models for the early detection of disease are planned by correlating the information by data scientists and the field health officers. There is a need to screen COVID-19 patients promptly and accurately in a bid to contain the illness due to new strains that have come up and the constantly increasing caseload. Although conventional approaches for identification of such infections are accurate, they can be time

consuming and cumbersome, in cases of public health crisis this can prove disadvantageous. Machine learning (ML) has developed as an influential technique to enhance the diagnostic procedures and outcomes based on data analysis methodologies for accurate prediction of complex patterns.

In this study, the following objectives are of concern: Applications of Logistic Regressions, a strong ensemble learning technique for quick and effective diagnosis of COVID-19 disease using health symptoms. For logistic regression, it is a binary prediction model which offers the results in two parts once it arrives at the result of every node in its structure. A more effective approach to developing a classifying model is proven by using the binary logit model as it can handle a large set, can handle missing values and give insight into feature extraction.

Robustness against overfitting: To extend the developed model to unseen clinical records, overfitting has to be avoided and this is done by use of logistic regression. This robustness makes it possible for the model to perform well with new patient data to clean.

Feature importance: In another type of analysis, feature importance can also be inferred through logistic regression, to deduce which symptoms are most likely to be associated with COVID-19. This assists doctors to reduce the number of important symptoms that indicates a certain disease.

Handling missing data: If there are data columns with missing entries it can be solved with the help of the logistic regression which is applicable for clinical data as in most cases there is more missing data. This feature is very useful in making sure that the model does not go out of its way just because a few data have been left out.

Proposed model: Logistic regression was selected due to its efficiency in classification problems, namely, it is appropriate to use in cases of medic data complexity. To verify the accuracy of the model and allow for its fine tuning, rigorous validation steps will be conducted in order to make sure that the model conforms to the highly stringent requisite of a diagnostic tool for diseases with many signs and symptoms.

Validation and Optimization

- **Validation:** In machine learning, validation refers to the steps taken for coming up with a model to test for correctness in a different data set apart from the training set not to be prone to overfitting – a situation. Holds where the model makes high accuracy in the training data as compared to the low accuracy in new data. Other methods like the cross validation are useful when training and validating the model on different portion of a given dataset divided into several sections. This helps in tuning the parameters of the models to choose the best model characteristics for each of them, and also ensure that the gained model is capable of the right forecast of outcome in other data.

- **Optimization** in machine learning means altering model parameters with an aim of minimizing or maximizing an objective function or cost function. That is how the model is trained in the given training data set so that it can improve its reliability for future applications.

Essential Ideas in Optimization.

• Goal Role:

Sometimes it is called the cost function or the loss function: it defines the measure of error or the difference between the target values and the model's predictions. That is why the minimization of this function constitutes the topic of optimization.

• Learning Pace:

An important hyperparameter in the gradient descent wherein determinate governs how much Mass of the model needs to be shifted based on the gradient value. Sometimes the learning rate could be too high and makes overshooting the minimum or conversely, the learning rate could be too low, making the learning very slow.

• The Random Gradient Descent (SGD) Method:

Sometimes called 'stochastic gradient descent,' this is different from the basic gradient descent in that the update is made for the model at every training example. This often results in faster convergence but at the same time brings high update noise level in the process.

Steps for validation and optimization:

To ensure the accuracy and reliability of the model, the study will include several key steps:To ensure the accuracy and reliability of the model, the study will include several key steps:

1. **Data preprocessing:** The first process is to clean and prepare the data set for analysis and other methods that covers absence in the data and norms the data values.

2. **Model Training:** The entire set of data undergoes a data preprocessing stage followed by feature extraction and feature selection After the feature extraction and feature selection stage the data set is split into training and testing where the conventional logistic regression model is trained and tested using cross validation methods for hyperparameter tuning.

3. **Model Validation:** Model evaluation can be computed with respect to the desired goals by using several measurements including accuracy, precision, recall and F1- score before Cross validation.

4. **Model optimization:** Optimizing hyperparameters in the model and analyzing a set of design choices to improve the model's performance.

5. **Model evaluation:** The performance of the logistic regression model discussed in the article is compared to other classifiers to make sure the best approach is used for patient diagnosed with COVID-19.

II. Logistic Regression: Why to use it?

1. Successful and simple to use: It is applicable for fitting binary classification models and frankly easy to use.

2. Probability Output: Supplies a probability value that can be employed so as to gauge the likelihood of the forecast.

3. Interpretable: Relative values of several aspects within the models can be determined with the help of model scales.

LITERATURE REVIEW

Shreshth et.al (2020) developed an AI based framework called Weibull Robust Model using iterative weighting to foresees covid-19 using cloud computing platform which performed better than baseline Gaussian model. Five distributions on confirmed cases of COVID-19 each day has been performed where GIW Distribution (Weibull) performs the best.

E. Gothai et.al (2021) formulated a model using 172,479 documents where they extracted only confirmed COVID-19 cases using dataset from John Hopkin's university database and after refining given as training data to the model. They recommended Time series Holt's model as it provided better results than Support vector regression and Linear regression. They have also mentioned the techniques through which Covid-19 patients forecasting can be increased.

Deepak et.al (2021) Deployed two models for two different problems, one is to forecast COVID-19 trend and other is to predict COVID-19 patients. They have used classifiers where Extra tree classifier outperform Logistic regression by giving 93% accuracy rate to diagnose covid infected patients. They have also used ARIMA (autoregressive integrated moving average) by using time series for forecasting the growing trend of COVID-19 in coming future.

Rifat et.al (2020) conducted survey among various machine learning models using PR (Polynomial regression), MLP (Multilayer perception), LSTM (Long Short-Term memory) and SIR (Susceptible, Infected and Removed) to calculate the prediction of COVID 19 crisis in Bangladesh to help the government by implementing new instructions and prescriptions to control the spread and vaccination distribution strategies.

Narinder Singh Punn et.al (2020) analyzed the combinational method by mixing machine learning and deep learning models to analysis the confirmed cases of COVID 19 by comparing various algorithms such as SVR, DNN, LSTM, PR under two parameters Root mean square error and forecast. Polynomial regression offered minimum value for Root Mean Square Error (RMSE) confirmed exponential growth of the pandemic globally.

Atta-ur-Rahman (2020) proposed a mathematical model name CSDC-SVM (Cloud-based smart detection algorithm using Support vector machine) using patient data gathered from various parameters such as headache, fever, shortness of breath, chest paint etc. Details related to patients are added into cloud and pre-processed. To predict covid 19 author used many cross-fold validation to diagnose and forecast the patient's illness provided accuracy rate of 98.4%.

Sina F. Ardabili (2020) performed a comprehensive survey of soft computing models and analytical models to prognose the covid-19 patients spread and impact.They proved that MLP (Multi-layer perceptron) and ANFI system performed well and replaced the traditional SEIR and SIR model.

Bodul et.al (2021) Wrote book analysing each and every perspective about study, evaluation, security threats, government policies regarding the COVID-19 pandemic. In chapter 15 author analysed an open-source ML software WEKA and the algorithms such as LR, SMOreg, MP etc. MAPE and RMSE are two error criteria to calculate the prediction,forecasting and spread of covid

around 12 selected countries under which gaussian and multilayer perceptron gave the most precise results.

R. Sujath et.al (2020) introduced a new method to predict the expand of COVID 19 effected patients extracting confirmed cases, deaths and recovered cases dataset from Kaggle and verify result by comparing it with john Hopkins university repository. They proved that MLP performs better than LR and VAR as it can handle non linear complex data patterns after data analysis using Orange and WEKA tools.

Zohair Malki et.al (2021) proposed a prototype based on decision tree and compared with various models (deep learning and ARIMA) where root mean square error and mean absolute error is determined on training dataset of 15 countries with most precise parameter R2 value 0.99.

III. ALGORITHMS

Data preparation:

• **Collect data:** Collect information from various sources (eg clinical symptoms of patients).

• **Clean Data:** Correct missing or incorrect data to ensure model performance

• **Split the data:** Split the data into two parts: one for training the model and one for testing it. Here we split the dataset 7:3. 70% of the data will be used for training and the rest will be used for testing purposes.

Forest construction:

• **Create Multiple Trees:** For each small part of the training data (so-called "bootstrap sample"), create a decision tree. Each tree looks at a random subset of features (features) and splits the data to make a prediction.

• **Tree growth:** Each tree makes decisions based on the data it sees. It continues to split the data until it can no longer be split or until it reaches a certain limit.

Model Validation:

• Check Accuracy: Check how accurate the forest's results are when tested on the test data set to the real outcomes.

• Measure Performance: They include; Accuracy, Precision, Recall and other techniques needed to check how well or bad the model is.

Optimization:

• **Improve the model:** Change something within the structure (for instance if the number of trees must be adapted or the depth each tree must develop) to improve the model.

• **Feature tuning:** Sometimes one can try to introduce or modify parameters in order to make predictions better.

Final Rating:

• **Compare with other models:** In other words, see how the performance of your new model looks in comparison with other models such as logistic regression model or support vector machines.

• **Real-world test:** It is vital to confirm that the model performs well on unrealistic new data and plays a role in practice.

RESULTS

Confusion matrix

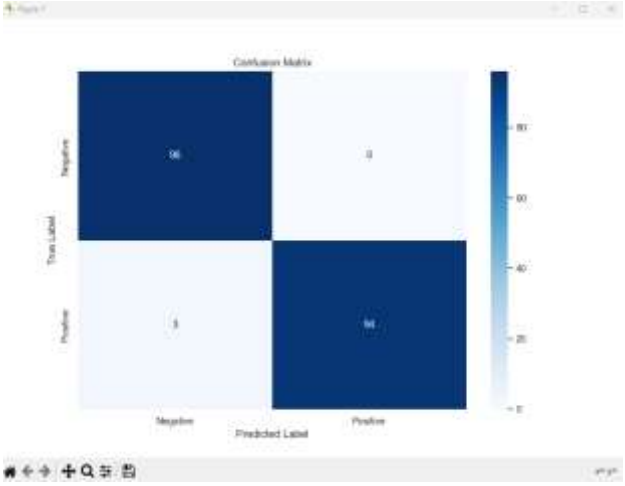

Fig. 1 Confusion matrix

True Positives (TP): There were 94 right classifications of positive instances out of which it shows that the model has accurately classified ninety-four positive inputs.

True Negatives (TN): Pre-identified negative cases = 96. The model state that it correctly predicted negative 96 times hence demonstrating that it identified true negative 96 times.

False Positives (FP): The number of false positives is 0 which means that there is no situation that the model classified any negative instances as positive.

False Negatives (FN): It classified three positive cases as negative thus it is off by three. This means that three positive cases were upwardly classified as negative cases now that the actual image is clear.

Performance Metrics:

Using the confusion matrix, you can calculate several performance metrics:

1. **Accuracy:**

Accuracy= (94+96)/(94 + 96 + 0 + 3)= 190/193= 0.985

Precision= 94/(94+0)= 1.0

Recall=94/(94+3)=0.969

F1 Score=2*(1.0*0.969)/(1.0+0.969)=0.984

ROC curve

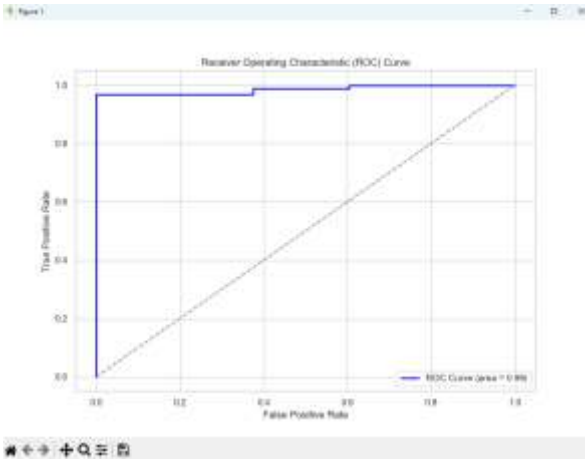

Fig. 2 ROC curve

The receiver operating characteristic (ROC) curve provides the most complete overview of the solution's effectiveness in recognizing positive and negative samples. From the confusion matrix the true positives are 94, while the true negatives are 96, false positives have been noted at 0 while the false negatives came out to be 3. These results are depicted from the ROC curve since it presents a high True Positive Rate (TPR) and low False Positive Rate (FPR). As a result of no false positive predictions, the curve should still lie close to the upper left quadrant of graph proving just how accurate this model is. AUC will also be high almost equal to one. 0 which indicates that the model based on MASES has high ability of the classification of positive and negative classes. This high AUC
value indicates that the model effectively
identifies positive cases with minimal misclassification, thus confirming its strong predictive capabilities.

Precision and recall curve

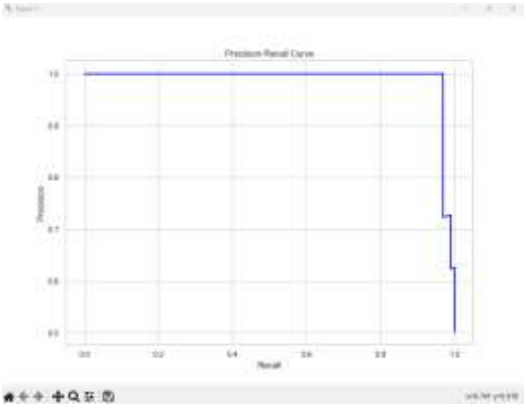

Fig. 3 precision and recall curve

Accuracy				0.98	193
Macro	Avg	0.98	0.98	0.98	193
Weighted	Avg	0.98	0.98	0.98	193

This 100% accuracy reflects that every event that occurred, predicted by the model to be positive, is actually positive, indicating perfect accuracy of positive predictions. This recall of approximately 96.9% shows that the model effectively identifies the majority of true positive cases, although there are a few positive cases that are missed. When plotted, the Precision-Recall curve will typically start at a high precision value (because there are no false positives) and gradually decrease as recall increases. In this case, the curve is likely to be near the upper right corner, where high precision and high recall intersect, reflecting the effectiveness of the model in both correctly predicting positive cases and capturing a high proportion of true positive cases. The high precision and relatively high recall indicate that the model is well calibrated for positive predictions, making it reliable in scenarios where accurate identification of the positive class is critical. **Tabular representation**

Table:1 accuracy

	Precision	recall	f1-score	support
0	0.97	1	0.98	96
1	1	0.97	0.98	97

Table:2 precision and recall values

IV. CONCLUSION

Implementation of a logistic regression model for rapid diagnosis of COVID-19 from clinical symptoms has shown promising results. Model performance metrics show high efficiency in accurately classifying cases. Specifically, the confusion matrix shows that the model achieved 94 true positives, 96 true negatives, 0 false positives, and 3 false negatives. This translates to an impressive accuracy of 98.5%, with a perfect precision (100%) and a recall of around 96.9%. The high F1 score of 98.4% further highlights the balanced and accurate performance of the model in predicting positive cases of COVID-19. ROC curve analysis supports these findings and shows a high area under the curve (AUC), indicating a strong ability of the model to discriminate between positive and negative cases. Moreover, the Precision-Recall curve highlights the effectiveness of the model, starting with high precision and maintaining robust recall values, demonstrating its reliability in identifying positive cases.

V. FUTURE WORK

Future Work:

To further enhance the model and ensure its robustness, several avenues for future work can be explored:

1. **Expanded Dataset:**
    - **Diversity:** Incorporate a more diverse dataset with varied demographic and clinical profiles to improve the model's generalizability across different populations.
    - **Data Augmentation:** Utilize techniques to augment the dataset, such as synthetic data generation, to address any imbalances or data limitations.
2. **Model Comparison and Ensemble Approaches:**
    - **Benchmarking:** Compare the Logistic regression model with other classifiers, such as Support Vector Machines (SVM) and Gradient Boosting Machines (GBM), to evaluate relative performance and identify potential improvements.
    - **Ensemble Methods:** Explore the use of ensemble methods that combine multiple models to leverage their strengths and enhance overall accuracy.
3. **Hyperparameter Tuning and Feature Engineering:**
    - **Advanced Tuning:** Implement advanced hyperparameter tuning techniques, such as Bayesian optimization, to further refine model parameters and improve performance.
    - **Feature Selection:** Investigate additional feature engineering techniques to enhance the relevance of input features and improve the model’s predictive power.
4. **Real-World Testing and Validation:**
    - **Clinical Trials:** Conduct real-world testing in clinical settings to validate the model's performance in practical scenarios and assess its efficacy in live diagnostics.
    - **User Feedback:** Gather feedback from healthcare professionals to ensure the model's usability and integration into diagnostic workflows.
5. **Handling Evolving Data:**
    - **Adaptive Learning:** Develop mechanisms for the model to adapt to evolving data and emerging variants of COVID-19, ensuring ongoing accuracy and relevance.

VI. REFERENCES

Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*, *11*, 100222. https://doi.org/10.1016/j.iot.2020.100222

Gothai, E., Thamilselvan, R., Rajalaxmi, R., Sadana, R., Ragavi, A., & Sakthivel, R. (2023). Prediction of COVID-19 growth and trend using machine learning approach. *Materials Today: Proceedings*, *81*, 597–601. https://doi.org/10.1016/j.matpr.2021.04.051

Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. In *Elsevier eBooks* (pp. 381–397). https://doi.org/10.1016/b978-0-12-824536-1.00027-7

Sadik, R., Reza, M. L., Noman, A. A., Mamun, S. A., Kaiser, M. S., & Rahman, M. A. (2020). COVID-19 Pandemic: A Comparative Prediction using Machine Learning. *International Journal of Automation, Artificial Intelligence and Machine Learning*, 01–19. https://doi.org/10.61797/ijaaiml.v1i1.44

Arora, N., Banerjee, A. K., & Narasu, M. L. (2020). The role of artificial intelligence in tackling COVID-19. *Future Virology*, *15*(11), 717–724. https://doi.org/10.2217/fvl-2020-0130

Atta-Ur-Rahman, N., Sultan, K., Naseer, I., Majeed, R., Musleh, D., Gollapalli, M. a. S., Chabani, S., Ibrahim, N., Siddiqui, S. Y., & Khan, M. A. (2021). Supervised Machine Learning-Based Prediction of COVID-19. *Computers, Materials & Continua/Computers, Materials & Continua (Print)*, *69*(1), 21–34. https://doi.org/10.32604/cmc.2021.013453

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T., & Atkinson, P. M. (2020). COVID-19 Outbreak Prediction with Machine Learning. *Algorithms*, *13*(10), 249. https://doi.org/10.3390/a13100249

Bodul, D., Jakovac, P., Gaspari, F., Poščić, A., Martinović, A., Zafiroski, J., Stanković, J., Marjanović, I., Drezgić, S., Bogović, N. D., Čegar, S., Stojanović, B., Kostić, Z., Vučić, V., Rendulić, D., Mihanović, D., Troković, R., Sameti, M., Redzepagić, S., . . . Čolig, M. (2021). *Contemporary economic and business issues*. https://dabar.srce.hr/en/islandora/object/efri%3A3002

Sujath, R., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, *34*(7), 959–972. https://doi.org/10.1007/s00477-020-01827-8

Malki, Z., Atlam, E. S., Ewis, A., Dagnew, G., Ghoneim, O. A., Mohamed, A. A., Abdel-Daim, M. M., & Gad, I. (2021). The COVID-19 pandemic: prediction study based on machine learning models. *Environmental Science and Pollution Research International*, *28*(30), 40496–40506. https://doi.org/10.1007/s11356-021-13824-7