covMAG: Patient Allocation Management Tool for Infectious Disease Based on Data Mining with Machine Learning Algorithms for COVID-19

Kazi Md Shahiduzzaman[1], Sumaiya Islam[2], Samia Sultana Mohua[2]

Abstract: The Global Pandemic Preparedness ranking indicates that our country was unprepared to manage a pandemic as the number of infectious cases increased exponentially during the Pandemic, namely SARS-CoV-2. Separating the infectious from the level of infectiousness becomes so difficult in a short period. Since treating everyone needs a vast amount of clinical assessment, many patients will generate considerable data. That results in much more difficulty in manually analyzing datasets by compensating for treatment time. Through this study, we will provide a management tool by analyzing datasets with machine learning algorithms using MATLAB. This management tool will employ a highly efficient and effective hospital management system for the COVID-19 situation to tackle severe waves and regular COVID-19. The average evaluation of performance parameters gives the best result for both SVM and KNN algorithms which is 92% and 93% respectively.

Keywords: Pandemic, COVID-19, Severity, Patient Allocation tool, Machine Learning, SVM, KNN, RF Classifier

1. Introduction

Modern medicine's challenge is compiling, interpreting, and using the considerable knowledge needed to address challenging clinical issues. AI applications that support physicians in diagnosing patients, choosing treatments, and forecasting their prognoses have been closely linked to the advancement of medical AI. AI programs have been developed to assist clinicians in making diagnoses, treatment decisions, and predicting patient outcomes. These programs, including machine learning and deep learning in

[1] Associate Professor, Department of Electrical and Electronic Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh. Email: shahiduzzaman@jkkniu.edu.nd

[2] Department of Electrical and Electronic Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh.

data science, aim to simplify data and knowledge management and aid healthcare professionals in their daily work (Ramesh et al., 2004). A significant challenge is the need for high-quality data, as medical data is often complex, diverse, and unstructured.

Additionally, privacy concerns related to patient data can limit access to high-quality datasets for machine learning algorithms. The need to develop machine learning models that are explainable and transparent is another obstacle in medical applications; it is critical to understand how the algorithm arrived at a particular decision, as this information can have significant implications for patient care. There is also a need for standardization in machine learning practices, such as data preprocessing, feature selection, and model evaluation. Lack of standardization can lead to inconsistencies in results and hinder the deployment of machine-learning models in clinical settings.

Finally, machine learning models must be flexible enough to adjust to evolving data. Machine learning algorithms in medical applications must be able to update and adapt to the potential for changes in patient populations and treatment regimens.

Because AI technology has implications in more psychological domains, such as experience, intelligence, and expert judgment, than prior technologies, it is revolutionary. Artificial intelligence (AI) technology can now analyze data patterns comparable to the typical human worker for any given task, thanks to the significant improvement in pattern recognition accuracy brought about by machine learning technology. [Erickson and others, 2017]. Because machine learning algorithms are based on data mining, which reconstructs the network of neurons in the human brain and can learn highly complex non-linear correlations, they are actively used in roles involving medical data (Hu et al., 2019). Artificial intelligence has many applications in the healthcare industry, ranging from medication creation and diagnostics to workflow management in clinics and hospitals. Artificial intelligence is employed explicitly in the following areas: disease diagnosis and detection, cancer detection, chronic condition management, drug development, robot-assisted surgery, health care delivery, and scientific research and experiments [Aggarwal et al., 2022].

COVID-19 is among the best illustrations of how artificial intelligence may help fight disease. The COVID-19 epidemic has been plaguing the planet since March 2020, resulting in notable rates of illness and mortality (Yassin et al., 2021). This disease has forced countries to enact stringent response

measures to reduce the spread of the illness. These measures include mandatory quarantines for asymptomatic contacts, curfews, border closures, social distancing, travel restrictions, school closures, and hospital patient isolation (Violato et al., 2021). All people have had various health, psychological, financial, social, and cultural repercussions due to this pandemic and the related health procedures, but COVID-19 patients, students, and healthcare staff have been most affected (Ramadan et al., 2021). As a result, specialists are working feverishly to employ AI to lessen the pandemic's impacts. Various applications, deep learning methods, and artificial intelligence have contributed to the fight against the COVID-19 pandemic (Jamshidi et al., 2020). These applications involve the following: contact tracing of clusters, identification of high-risk patients, automatic monitoring and prediction of the virus's spread, rapid and accurate diagnosis, and detection of the infection through the use of practical algorithms and medical imaging technologies, and the prediction of COVID-19 case count and mortality rates in any given region through thorough patient data analysis (Rivas et al., 2020). Moreover, artificial intelligence has been applied to social control by enforcing social separation and lockdown procedures and employing thermal imaging to search public areas for possibly infected individuals.

As of March 29, 2023, at 9:19 a.m. CEST, 761,402,282 COVID-19 confirmed cases—including 6,887,000 deaths—had been reported to WHO worldwide. Bangladesh ranked 95th in the 2021 Global Security Health Index, averaging 35.5 across all six categories, up +0.01 from the previous year. According to the Global Pandemic Preparedness Ranking, our nation was ill-prepared to contain a pandemic (Bell et al., 2021). Bangladesh, a country with a land area of 147,570 square kilometers, is among the most densely populated places on Earth (56,977 square miles). Death rates could be lowered with early detection and appropriate care. Enhancing survivability through early and sufficient identification using machine learning techniques is essential. Whether or not we can apply this to improve pathologists' diagnostic process will be a critical first step toward increasing early and appropriate detection.

This study aims to sort patients based on infection severity. This study employs a Machine Learning Algorithm to analyze data sets using MATLAB. Machine learning algorithms can help analyze large data sets and identify patterns and trends. Using this feature, patients can be sorted according to their severity. By doing so, limited healthcare resources can be

allocated more efficiently, and patients can receive the appropriate care based on their needs. Developing countries such as Bangladesh face significant challenges regarding pandemic preparedness, and it is crucial to sort patients based on infection severity. This management tool aims to provide an efficient and effective management system to this challenge to ensure proper utilization of limited healthcare resources.

2. Literature Review

We examined a few studies that were similar to ours in this area. We have examined the literature based on the samples used, the quantity of samples collected for analysis, and the study period's constraints.

The study of Wu et al. highlights the importance of early diagnosis for prompt treatment of COVID-19. The researchers used a random forest algorithm to analyze 11 important blood indexes and derived 49 clinical blood test data using commercially available blood test equipment to develop a tool for assisting in discrimination. However, the study faced a significant challenge as it took time to obtain COVID-19 cases with standard symptoms due to the current situation (Wu et al., 2021).

Yue et al. created a machine learning-based model to forecast how long COVID-19 pneumonia patients would stay in the hospital. The model included random forest and logistic regression methods. 52 COVID-19-positive patients and their CT scans from five different Chinese hospitals were the data source used in the model between January 23, 2020, and February 8, 2020. The suggested models accurately predicted the length of hospital stays for COVID-19 pneumonia patients, which could be less than ten days or more (Yue et al., 2020).

Rahul Kumar et al. created a technique that uses asymptomatic patients' chest X-rays to forecast the spread of COVID-19. This strategy could facilitate more efficient monitoring and aid in the early detection of outbreaks. Utilizing machine learning techniques, the researchers constructed and classified the model on a dataset of 5,840 X-ray scans from Italy. The study divided participants into three groups: regular, COVID-19-positive, and pneumonia-positive (Kumar et al., 2020).

The research uses machine learning classification algorithms titled "Severity Prediction of COVID-19 Patients: A Case Study of Small City in Pakistan with Minimal Health Facility." The authors utilized a dataset from Kaggle for their study, and the paper was published in IEEE. To estimate the severity

of COVID-19 patients, they employed the SVM algorithm and Python for data analysis. This methodology categorized patients into three severity levels—mild, moderate, and severe—with a 60% accuracy rate. They exclusively considered COVID-19 symptoms, which include fever, exhaustion, dry cough, aches, dyspnea, and other associated symptoms (Gull et al., 2020).

Based on the findings of the literature review we have selected the following objectives for this research work:

i. Designing a machine learning-based tool (covMAG) that can determine Covid-19 as well as classify its severity.
ii. Development of a data set based on a few clinical data.
iii. Performance analysis of the proposed covMAG tool.

3. Proposed covMAG Tool

Machine learning, a subset of artificial intelligence, has led to significant advances in various fields, including medicine. To train computer models properly, machine learning methods require big datasets. Although medical datasets are accessible, more datasets from multiple medical entities are needed, particularly in COVID-19.

Our proposed model aims to classify the severity of a medical condition as mild, moderate, or severe after confirming a positive diagnosis. If the state is classified as soft, the patient will not be recommended for hospitalization, as shown in the following block diagram (Figure 1). However, if the condition is severe, hospitalization will be recommended. In the case of moderate severity, the model will further analyze the patient's condition to determine whether hospitalization is necessary. This architecture proposes an adaptive algorithm (See) that can select the best fit to make COVID-19 detection and severity prediction highly considerable.

4. Dataset Development and Data Processing

The datasets used in this experiment are demi-datasets from secondary sources for primary detection of COVID-19 (0/1) (Wang et al., 2020). The blood and clinical sampled datasets from patients confirmed to be infected with COVID-19 and taking references from different secondary sources were generated into two sets of demi-data to detect and predict severity cases (Rampal & Seng, 2020). Two hundred records were taken for binary

classification and 150 for the multiclass classifier to indicate the severity level into the Mild, Moderate, and Severe categories. The features comprise various clinical parameters broadly categorized under CRP, platelets, lymphocytes, monocytes, neutrophils, basophils, eosinophils, leukocytes, and hemoglobin (Poggiali et al., 2020).

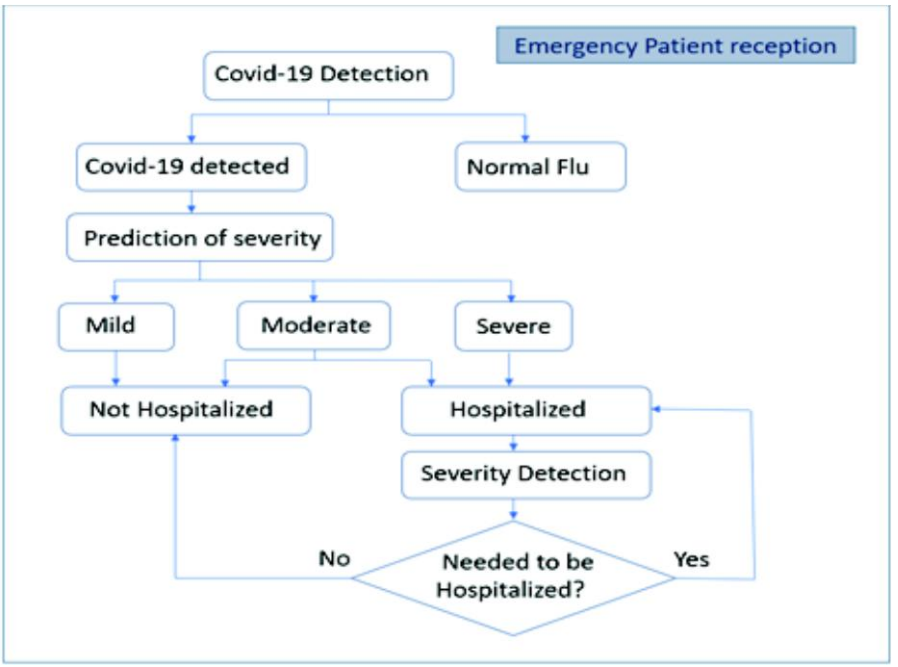

Figure 1: Flow diagram of covMAG tool for patient management plan during COVID-19

We have collected primary data from various sources such as medical prescriptions, medical reports, journal papers, news articles, medical journals, the Directorate General of Health Services (DGHS) Dashboard, etc. By sampling the primary Data, extracting the feature of the data, attributing, and labeling were done to generate the secondary data (Blomme et al., 2022), as shown in Figure 2

Our demi-data set contains 200 records for binary classification to detect COVID-19 positive and negative cases. Among these, 150 were seen as positive cases, and 50 were negative cases. Nine attributes were taken for symptoms like fever, cough, headache, breathing problems, muscle pain, Fatigue, smell and test disturbance, sore throat, and abdominal pain (Darapaneni et al., 2021).

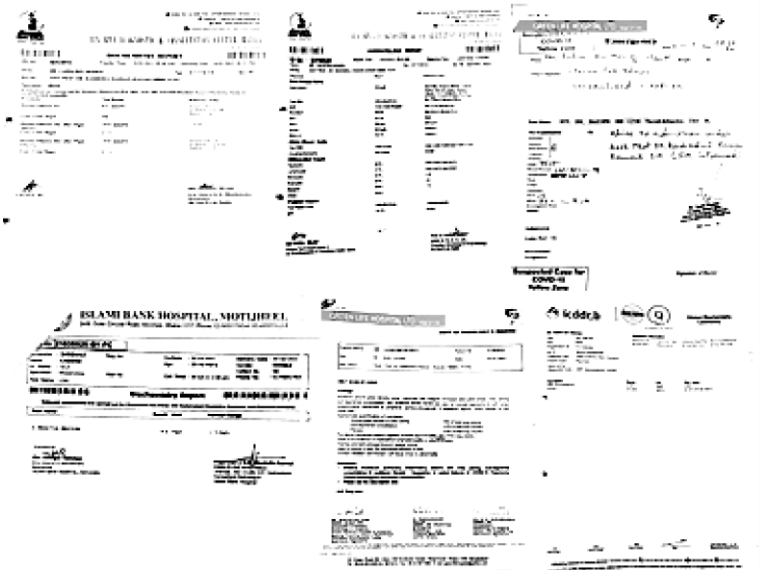

Figure 2: Data Generation Sources

The following shows the schematic of the data generation process. In severity cases, 150 positive points were divided between mild and moderate, and nine clinical parameters were taken. The features comprise various clinical parameters broadly categorized under CRP, platelets, lymphocytes, monocytes, neutrophils, basophils, eosinophils, leukocytes, hemoglobin, and severe cases.

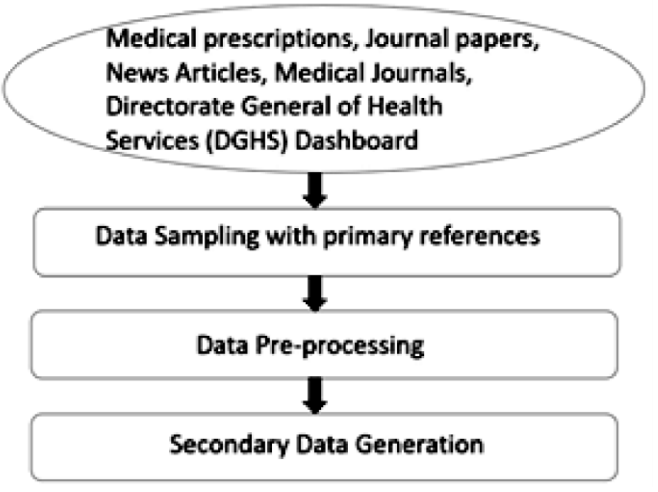

Figure 3: Schematic of Data Generation

Figure shows the overall distribution of the dataset. Our dataset must be divided into training and testing sets. To improve the accuracy of our model, we have employed two sets of train-test ratios in this paper: training data, 80 percent; testing data, 20 percent; training data, 70 percent; and testing data, 30 percent. We have also included a 5-fold cross-validation and Holdout method.

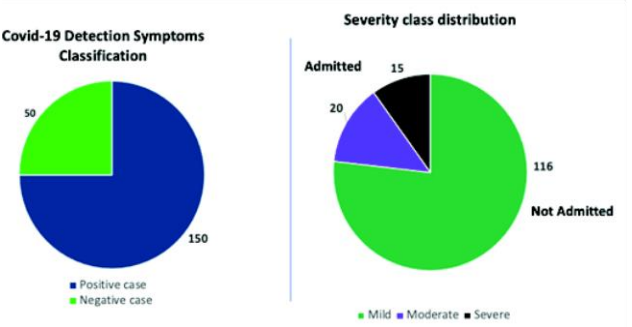

Figure 4: Distribution of Developed Dataset

5. Methodology of covMAG Tool

Before starting this section, we must admit that our goal for this research is to develop a patient management tool to manage not only the COVID-19 situation but also other contiguous diseases. Therefore, this tool utilized three already set and well-established supervised machine learning algorithms, named Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF). The tool selected the best fit trained model based on the validation accuracy. The covMAG tool's general process is depicted in Figure 5.

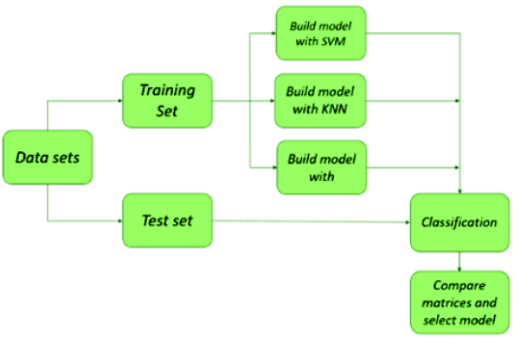

Figure 5: Methodology of covMAG tool

6. Result and Discussion

This section will discuss the experimental setup, performance evaluating matrices, and obtained results.

6.1 ***Experimental setup*** Figure 5 summarizes the experimental setup used in this paper. The raw demi-datasets were divided into two sets for binary and multiclass classification. Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF) classifier techniques were implemented on the Binary and Multiclass datasets. Performance evaluation will be based on accuracy, precision, recall, and f1-score.

We have used a desktop PC with an Intel 5 8th Gen processor with 24 GB RAM, 25g GB M.2 SSD, and NVIDIA GeForce GT 1030 GPU for running simulation. MATLAB-2022a was used on this desktop to determine how SVM, KNN, and RF algorithms can detect COVID-19 and predict severity.

6.2 *Performance Evaluating Matrices*

In our research, we have used the following metrics to evaluate the tool's effectiveness.

- Accuracy: The accuracy is determined by taking the total number of examples that have been labeled and dividing that by the number of models that have been correctly labeled. In terms of the confusion matrix, it is given by:

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

- Precision: The percentage of correct positive predictions to all optimistic forecasts is known as "Precision" and is expressed as follows:

$$\text{Precision}=\frac{TP}{TP+FP}$$

- Recall: In testing data, recall is the ratio of accurate optimistic predictions to the total number of positive examples.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: The F1-score is a statistic that is produced by summing the precision and recall scores of a classifier using their harmonic mean. Gaining a high F1 score necessitates having a solid recall and precision base. The precision and recall numbers are averaged to determine the F1 score. The choice to use the harmonic mean is clear-cut and logical because they are both rates. This is the formula for the F1 score:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

From the confusion matrix, we obtain the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). By comparing the model's predictions with real data, the confusion matrix is a table that shows how well the classification model predicts cases that fall into different classes. In the confusion matrix, the real title appears along one axis, while the label predicted by the model appears along the other.

6.3 *Result and Evaluation*

This section will discuss how our proposed covMAG tool detects COVID-19 and classify it into three severity stages. Figure shows that our proposed tool selects the classifier based on the performance-evaluating matrices. For COVID-19 detection, covMAG chooses SVM.

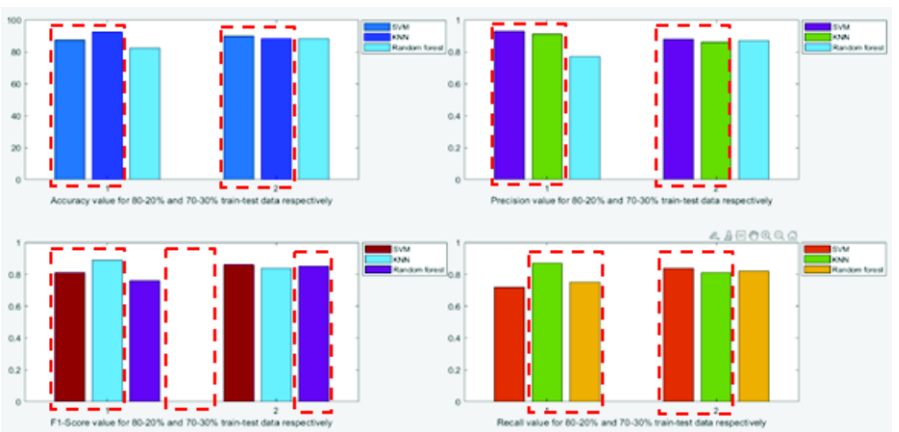

Figure 6: Performance Evaluation of COVID-19 Detection

The confusion matrix in Figure shows the efficiency of detecting COVID-19. We can easily observe that covMAG can differentiate Normal flu from COVID-19 in both cases of data split with an accuracy of more than 92%.

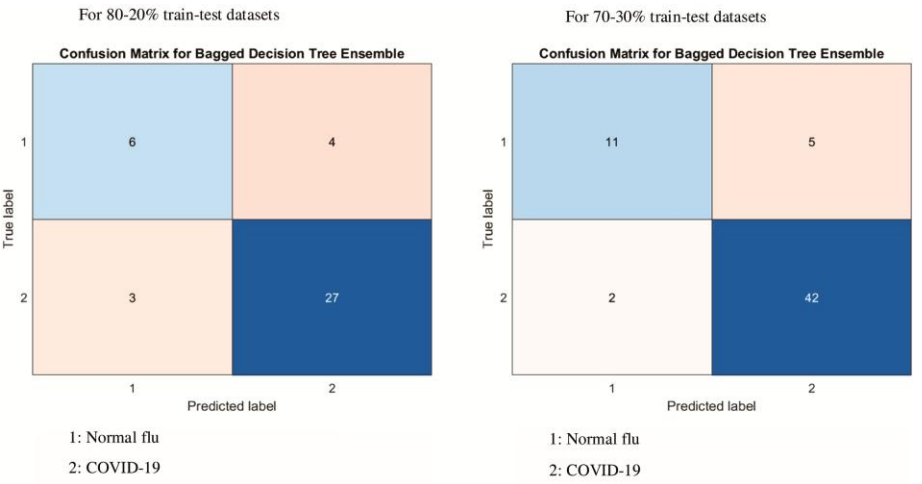

Figure 7: Confusion Matrices of COVID-19 Detection

After the COVID-19 detection, covMAG checks the severity to suggest to a patient what to do next in the treatment. The severity is categorized into three categories: 1) Mild, 2) Moderate, and 3) Severe. Like detection, covMAG again uses an adaptive mechanism to select the best-fit algorithm to recognize the severity of COVID-19. The medical personnel can decide on the treatment plan based on the severity classification. The following bar chart (Figure) shows how the RF classifier outperforms SVM and k-NN.

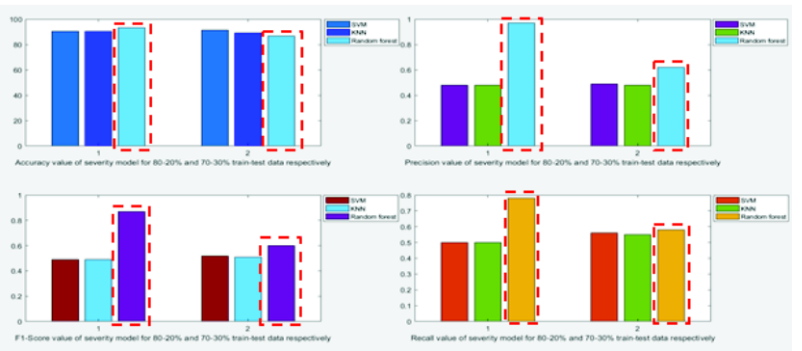

Figure 8: Performance Evaluation of COVID-19 Severity Recognition

After selecting the RF classifier, the tool started recognizing the severity. We have considered more than 30 cases. The confusion matrix shows that the proposed covMAG tool can accurately classify more than 93% severity.

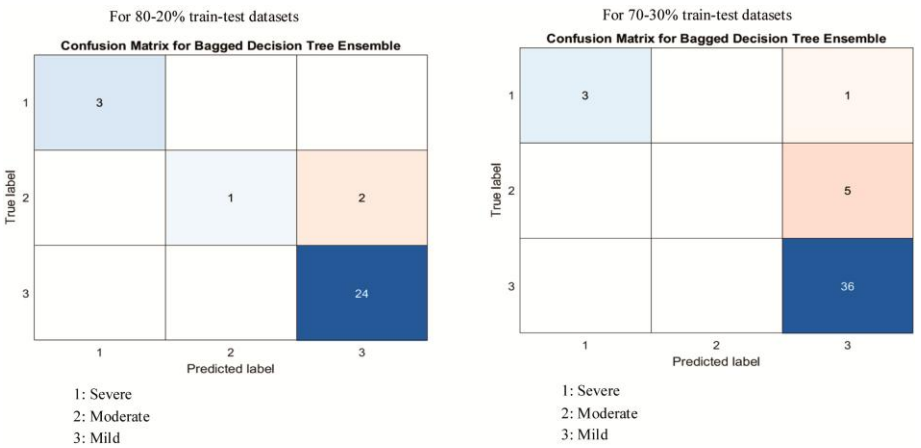

Figure 9: Confusion Matrix of Severity Recognition

For severity detection cases, the RF classifier shows the best fit. Specifically, for the RF classifier, the confusion matrix shows that the actual label matched with a predicted tag for both 80-20% and 70-30% train-test ratios. However, in the case of severity detection, severe and mild classes are in two distinct levels, which is why they match the predicted label. For the moderate class, the actual label is mismatched with the predicted label for a 70-30% ratio, and 80-20%, one case is matched.

6.4 Discussion

covMAG shows promising results in detecting COVID-19 and correctly identifying its severity. From the obtained result, we can come up with some enlisted discussions.

1. Adaptive machine learning algorithm selection is beneficial for handling highly contiguous diseases like COVID-19.
2. We have determined COVID-19 with a percentage of 92.45 accurately while accuracy reaches 93.33 percent in severity identification.
3. 80-20 data split shows better performance in detection and classification.

7. Limitation of our Study

One of the biggest challenges we face is the unavailability of real datasets. For an accurate data set to train and further test our model, we applied for permission to provide accurate data online and in the Directorate General of Health Services and private and public Medical Hospitals. After continuously visiting DGHS for six months to collect real datasets, authorities could not

provide us with the data required to test our model. Real datasets are complex to come by as they need a lot of effort and resources to collect, clean, and label. Hence, We have to settle for demi datasets with smaller datasets to train in the machine learning model.

7.1 Bias in Data

Moreover, even the actual data may not be representative of the population or may have biases that can affect the performance of the machine learning models trained on it. This problem is known as bias in data, and it can lead to unfair decisions and predictions.

7.2 Limited Resources

Another challenge that we face is limited resources. Resources refer to the Medical Equipment, Bed allocation, ICUs, and infrastructure required to process and analyze our data.

7.3 Small Data Sample

Finally, another challenge that we face is the small data sample. A small data sample refers to a dataset that has a limited number of observations. Small datasets lead our result in overfitting, where the machine learning model fits the data too closely and needs to generalize to new data. Small datasets can also limit the diversity of the data and make it harder to detect patterns and correlations. This can affect the accuracy of the machine learning model and lead to unreliable predictions.

8. Conclusion

Throughout this study, we demonstrated a machine learning approach toward detecting COVID-19 and predicting its severance using clinical spectrum datasets. As we add more attributes to our training data, the accuracy, precision, f1-score, and recall improve significantly as the countries struggle with managing more patients than their capacity in any hospital. Proper utilization of patient allocation tools can enable minimizing risk for a patient. This can be further implemented to reduce the contact for the healthcare workers as the number of patients visiting the hospital to check can increase significantly as the wave surges. The model output is a binary indicator for SARS-CoV-2 infection and indicates severance measurement for patient allocation. An AI-powered COVID-19 classification website should be built.

In conclusion, we are optimistic that by using this process flow and training it on raw datasets resulting in rich data, this study's prediction can have a broad impact and achieve greater accuracy at less risk to human lives. Proper

implementation of real datasets to test our model by adding more attributes to our training data.

References:

A. Ramesh, C. Kambhampati, J. R. Monson, and P. Drew, "Artificial Intelligence in Medicine," Annals of the Royal College of Surgeons of England, vol. 86, no. 5, p. 334, 2004.

B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine Learning for Medical Imaging," Radiographics, vol. 37, no. 2, pp. 505–515, 2017.

W. Hu, B. Cai, A. Zhang, V. D. Calhoun, and Y.-P. Wang, "Deep Collaborative Learning with Application to the Study of Multimodal Brain Development," IEEE Transactions on Biomedical Engineering, vol. 66, no. 12, pp. 3346–3359, 2019.

K. Aggarwal, M. M. Mijwil, A.-H. Al-Mistarehi, et al., "Has the Future Started? The Current Growth of Artificial Intelligence, Machine Learning, and Deep Learning," Iraqi Journal for Computer Science and Mathematics, vol. 3, no. 1, pp. 115–123, 2022.

A. Yassin, A. Ghzawi, A.-H. Al-Mistarehi, et al., "Mortality Rate and Biomarker Expression within COVID-19 Patients Who Develop Acute Ischemic Stroke: A Systematic Review and Meta-Analysis," Future Science OA, vol. 7, no. 07, FSO713, 2021.

C. Violato, E. M. Violato, and E. M. Violato, "Impact of The Stringency of Lockdown Measures on COVID-19: A Theoretical Model of A Pandemic," PloS one, vol. 16, no. 10, e0258205, 2021.

M. Ramadan, Z. Hasan, T. Saleh, et al., "Beyond Knowledge: Evaluating the Practices and Precautionary Measures Towards COVID-19 amongst Medical Doctors in Jordan," International Journal of Clinical Practice, vol. 75, no. 6, e14122, 2021.

M. Jamshidi, A. Lalbakhsh, J. Talla, et al., "Artificial Intelligence and COVID-19: Deep learning Approaches for Diagnosis and Treatment," IEEE Access, vol. 8, pp. 109 581–109 595, 2020.

A. Rivas, "Drones and Artificial Intelligence to Enforce Social Isolation During COVID-19 Outbreak," Medium Towards Data Sci, vol. 26, 2020.

J. A. Bell and J. B. Nuzzo, "Global Health Security Index: Advancing Collective Action and Accountability Amid Global Crisis," NTI and Johns Hopkins Center for Health Security, 2021.

J. Wu, P. Zhang, L. Zhang, et al., "Rapid and Accurate Identification of COVID-19 Infection Through Machine Learning Based on Clinically Available Blood Test Results," MedRxiv, pp. 2020–04, 2020.

Yue H, Yu Q, Liu C, Huang Y, Jiang Z, Shao C, Zhang H, Ma B, Wang Y, Xie G, Zhang H, Li X, Kang N, Meng X, Huang S, Xu D, Lei J, Huang H, Yang J, Ji J, Pan H, Zou S, Ju S, Qi X. Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. Ann Transl Med 2020;8(14):859. doi: 10.21037/atm-20-3026

R. Kumar, R. Arora, V. Bansal, et al., "Accurate Prediction of COVID-19 Using Chest X-Ray Images Through Deep Feature Learning Model with Smote And Machine Learning Classifiers," MedRxiv, pp. 2020–04, 2020.

H. Gull, G. Krishna, M. I. Aldossary, and S. Z. Iqbal, "Severity Prediction of Covid-19 Patients Using Machine Learning Classification Algorithms: A Case Study of A Small City In Pakistan With Minimal Health Facility," in 2020 IEEE 6th International Conference on computer and communications (ICCC), IEEE, 2020, pp. 1537–1541.

G. Wang, C. Wu, Q. Zhang, et al., "C-reactive Protein Level May Predict the Risk of COVID-19 Aggravation," in Open forum infectious diseases, Oxford University Press US, vol. 7, 2020, ofaa153.

M. Lekhraj Rampal and L. B. Seng, "Coronavirus disease (covid-19) pan- demic," Med. J. Malays, vol. 75, no. 2, p. 95, 2020.

E. Poggiali, D. Zaino, P. Immovilli, et al., "Lactate Dehydrogenase and C-Reactive Protein as Predictors of Respiratory Failure In COVID-19 Patients," Clinicachimica Acta, vol. 509, pp. 135–138, 2020.

S. Blomme, L. Smets, M. Van Ranst, N. Boeckx, and C. Van Laer, "The Influence of COVID-19 on Routine Hematological Parameters of Hospitalized Patients," Acta Clinica Belgica, vol. 77, no. 2, pp. 241–246, 2022.

N. Darapaneni, A. Singh, A. Paduri, et al., "A Machine Learning Approach to Predicting COVID-19 Cases amongst Suspected Cases and Their Category of Admission," in 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), IEEE, 2020, pp. 375–380.