BIOMARKER-BASED COVID SEVERITY PREDICTION AND DATA QUALITY EXPLORATION

Mohamed Waleed El Habashy (s2525550)

mohamedwaleedmohamedelhabashy@student.utwente.nl

Supervised By: Faizan Ahmed

University of Twente. PO Box 217, 7500 AE Enschede, Netherlands

***Abstract* — This report focuses on investigating the significance of Data Quality (DQ) on COVID severity prediction models and how these models can affect resource allocation management. By understanding the impact of DQ on COVID datasets, valuable insights can be gained to enhance the allocation of resources. The main research question of this project is: "What is the importance of Data Quality in predicting the severity and progression of COVID?". The research methodology employed for this study is a Design Science Research Methodology (DSRM). The findings reveal the prevalence of DQ issues in COVID data, with Missing Data and Imbalanced Data being the most common issues. To evaluate the effects of data quality we developed a COVID severity prediction model using a Support Vector Machine (SVM), and a feature importance analysis using permutation importance to demonstrate the correlation between biomarkers and COVID severity. Among the biomarkers, Leuco (Leukocytes) exhibited the strongest correlation. The model achieved an accuracy of 76%, precision of 91%, recall of 69%, and an F1 score of 79%. The findings underscore the critical role of Data Quality in influencing model outcomes, highlighting the importance of proper preprocessing to ensure accurate and reliable results for the machine learning model. These results are crucial for the effective allocation of resources.**

Keywords—COVID-19; biomarkers; Data Quality; Resource Allocation; Machine Learning (ML); Severity

1. Introduction

Severe acute respiratory syndrome (SARS) was caused by a new virus called SARS-CoV-1, which first emerged in China in 2002. After spreading to other countries, it was eventually contained in 2003 after causing over 8,000 cases and 700 deaths [1]. In late 2019, a new coronavirus, SARS-CoV-2, emerged in Wuhan, China, causing a global outbreak of the coronavirus disease (COVID-19). The virus has since spread to infect millions of people worldwide, resulting in a pandemic that continues to impact the world today [2]. SARS-CoV-2 is highly contagious and can cause severe respiratory illness, particularly in older individuals and those with pre-existing health conditions. Although vaccines have been developed, the emergence of new variants and vaccine hesitancy pose ongoing challenges for controlling the pandemic. Scientists continue to research SARS-CoV-2 to better understand its biology, epidemiology, and clinical management to combat the ongoing COVID-19 pandemic. The pandemic has also had great impact on societies across the world, disrupting economies, impacting healthcare systems [3] and tragically claiming the lives of millions [4].

The surge in infections has posed immense challenges for healthcare providers, particularly in terms of resource allocation, with a critical emphasis on intensive care unit (ICU) beds and mechanical ventilators. The scarcity of these life-saving resources has necessitated the implementation of effective allocation strategies [5]. This research project aims to address this issue. Resource allocation can be effectively managed using Machine Learning methodologies. This study proposes the utilization of biological markers (biomarkers) present in the blood to determine the severity of COVID-19 in patients which can serve as an indicator to how the disease will progress.

Machine Learning can be utilized to create a severity prediction model based on the existing biomarker data of patients. Physicians can use these models to predict the severity of the disease in the patient that in turn enables them to decide whether a patient should be discharged or admitted to the hospital for monitoring. The severity of the disease is also a component of its progression, therefore physicians can make informed decisions on how the disease will progress based on its severity.

By using such a model, the severity of the disease can be detected immediately, and the necessary medical precautions will be executed in order to improve the patient's chances of survival. Efficient allocation of resources can be achieved based on such results.

However, the accuracy of this approach heavily depends on the quality of patient data. This research will outline the importance of data quality in predicting the severity of COVID and how that will ultimately affect the management of resources in hospitals. The COVID pandemic serves as a case study for this project, and determining how to properly allocate resources using Data Quality techniques and ML methodologies will help in effectively allocating resources in future pandemics. The main research question of this project is: "What is the importance of Data Quality in predicting the severity and progression of COVID?". This research question will be answered by answering these three research questions:

- RQ1: What are the data quality issues in COVID data?
- RQ2: How can the severity of COVID-19 be predicted using biomarker data?
- RQ3: How can COVID-19 disease progression improve resource allocation decisions for patient management?

The report will follow a specific structure. It will begin by outlining the research methodology used to address the three main questions. Each research question will have its own dedicated section, where the necessary methods employed to answer them will be explained in detail. Finally, the paper will conclude by summarizing the overall findings presented throughout the document.

2. Research Methodology

This section will provide an overview of the efforts that will be taken to answer the main research question. A Design Science Research Methodology (DSRM) [6] will be utilized. The DSRM consists of a 6-step process which will be discussed in section 2.2. Within the DSRM, a retrospective cohort study [7] will also be conducted.

2.1 The MST Dataset

The dataset of the case study for this project consists of anonymized user data from Medisch Spectrum Twente (MST). The dataset consists of over 3000 rows, but only a few hundred patients who have been tested multiple times throughout multiple dates. Each row consists of 25 biological markers that are based on real clinical blood tests. The data was collected at random with no pattern as per request of a physician working at MST. The severity of COVID for each patient is labeled as a number from 0 to 4. Where 0 represents patients who tested negative for the PCR test, and 1 to 4 represent patients who have Mild, Moderate, Severe and Critical cases of COVID severity (respectively).

2.2 DSRM Process

Problem Identification: The COVID pandemic has resulted in the collection and analysis of data from diverse origins. However, these various sources often have dissimilar structures and formats, which can introduce inherent data problems. Moreover, biological datasets are known to have extensive data quality issues [24] which require exploration and curation. It is essential to address and rectify these issues to ensure the reliability of the data. Moreover, accurately predicting the severity of COVID using machine learning techniques is crucial for efficient resource allocation and management. The quality of the data plays a significant role in the performance and outcomes of these prediction models.

Objectives and Requirements: Our primary objective is to address three research questions: RQ1 - identifying prevalent data quality issues in COVID data and the MST data, RQ2 - predicting the severity of COVID using ML techniques and RQ3 - Exploring the correlation between COVID severity and progression. To assess the quality of the data effectively, we will design a comprehensive data quality framework specifically tailored for assessing COVID data. This framework will be evaluated by applying it to different COVID datasets that were mentioned in [8] (our research serves as an addition to this paper), which were also used to develop a predictive model related to COVID data. Subsequently, the framework will be applied to the MST dataset, identifying any data quality issues present. After identifying the data quality issues, the dataset will be preprocessed to develop a predictive model for COVID severity using ML techniques. Additionally, a retrospective cohort study will be conducted to examine the correlation between COVID severity and its progression, this will be achieved using a meta-analysis.

Design and Development: The design process will consist of a Literature Review to gather relevant papers that discuss Data Quality Frameworks. A framework will be created based on the key qualities of other existing data quality frameworks that have been successful throughout the years. Moreover, the success of different types of ML models (that are mentioned in [8]) that were used to create COVID severity prediction models will be investigated. Literature will also be gathered regarding specific biomarkers and their correlation with COVID severity. Literature regarding the Outcome Evaluation of COVID patients with different severity levels will also be reviewed to determine how the disease progressed with different severity levels..

Demonstration: The created Data Quality framework will be applied to various COVID datasets including the MST dataset to assess the quality of the data and to identify the most prevalent data quality issues. The MST dataset will be preprocessed using techniques that were mentioned in the other datasets. A predictive model will be developed to predict COVID severity.

Evaluation: The performance of the predictive model will be evaluated using the following metrics: Accuracy, Precision, Recall and F1 score.

Communication: These findings will be presented in the Twente Graduate Conference.

3. What Are The Data Quality Issues in COVID Data?

In order to effectively evaluate the reliability of COVID data, we propose the development of a Data Quality Framework, this can be observed in Figure 1. This framework will incorporate key elements from two existing frameworks: the Data Quality Assessment (DQAF) [9] and the Total Data Quality Management (TDQM) [10] frameworks. By combining the strengths of these frameworks, we aim to create a more comprehensive framework. The DQAF offers a structured approach to assessing data quality by defining specific dimensions of data quality and associating them with relevant data quality issues. This allows for a systematic evaluation of the data's quality. On the other hand, the TDQM framework follows a cyclical process that involves iterative cycles to continuously monitor and enhance the data's quality. According to [11], the TDQM consists of four main cycles: Define, Measure, Analyze, and Improve.

The Define cycle involves identifying the relevant data quality dimensions that are applicable to the dataset being assessed. This step helps establish a clear understanding of the specific aspects of data quality that need to be considered.

The Measure cycle focuses on identifying the potential data quality issues that may arise within the dataset, based on the selected data quality dimensions. This step allows for a comprehensive examination of the dataset's quality, considering various aspects that could affect its reliability.

The Analyze cycle aims to uncover the root causes behind the identified data quality issues. By understanding the underlying factors contributing to these issues, it becomes possible to develop targeted strategies for improvement.

Finally, the Improve cycle provides techniques and approaches for enhancing the overall data quality. This step involves implementing corrective measures, refining data collection processes, and adopting best practices to ensure the data's accuracy and reliability.

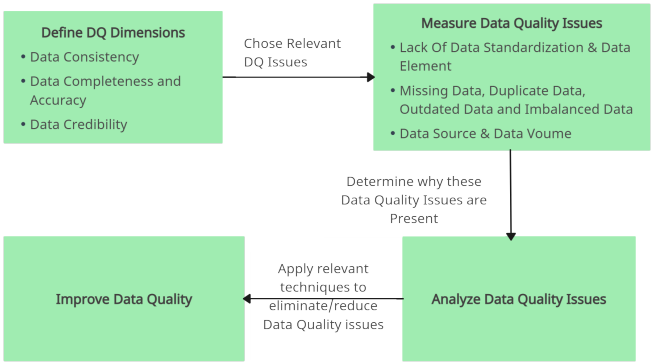

Figure 1: The Data Quality Framework

3.1: Motivation

The selection of data quality dimensions and issues in our framework is motivated by the need to address specific challenges encountered in biological datasets. Our framework consists of three main data quality dimensions: Consistency, Credibility, and Completeness and Accuracy. Each dimension includes potential data quality issues that were identified through research.

Data Consistency is the first dimension we will be exploring. Datasets can be categorized as Single-sourced or Multi-sourced. Single-sourced datasets are collected by a single organization, while Multi-sourced datasets are a compilation of various datasets. However, due to differing standards and rules across these datasets, data standardization is often lacking during collection. This results in diverse data types and complex structures during the data integration process, as highlighted by Cai and Zhu, 2015 [12]. Establishing a standard becomes crucial to handle variations in rules and regulations from different data sources. The combination of datasets with diverse rules introduces complexity and challenges in effectively managing the data. The importance of data standardization is emphasized in [13] and [14]. Data collection can also lead to issues in certain values, causing inconsistencies in the data. These issues, known as Data Element problems, require curation. Both the lack of data standardization and Data Element problems fall within the Data Consistency dimension.

The second dimension, Data Completeness and Accuracy, encompasses crucial data quality issues applicable to any dataset. Missing data is a common issue that complicates analysis and introduces bias [15]. Duplicate data introduces bias in ML models and can lead to data imbalance [16]. The impact of outdated data on ML models is not extensively studied, but it can significantly affect the accuracy of predictions as the model will learn from data that does not reflect the current environment. Additionally, imbalanced data, where target classes have uneven distribution [17], can introduce bias and negatively impact model performance.
The Data Credibility dimension governs two main data quality issues: Data Source and Data Volume. The Data Source refers to the source(s) that the data was collected from, the source(s) must be checked for reliability to ensure that the data is correct and trustworthy. Data Source also refers to the problems associated with the data that is from these sources. Data Volume refers to the scalability and size of the dataset, the size of the dataset can either be a limitation of a certain study, or/and can cause the results of a ML model to be inaccurate. If the dataset size is large, this can also be an issue since the cleaning process of the dataset can be very difficult [12]. The Data Volume is checked after all preprocessing methods are applied to the dataset.

3.2: Application of Framework on COVID Datasets

In this section, the framework from Figure 1 will be used to assess the data quality of COVID datasets that were used to create a predictive model related to COVID datasets as mentioned in section 2.2.

All of the datasets discussed the issue of missing data and how to solve it [18]-[26], the most common way to handle missing data was to remove the rows where the missing data existed or impute it with another value.

One of the other major data quality issues was the imbalance of data as mentioned in [19],[24], the most common way to rebalance the dataset was to use a SMOTE algorithm during the training process. SMOTE is a statistical technique for increasing the number of the cases in a balanced way. The Data Volume was way lower after preprocessing the datasets, this was mentioned in [18],[22] and [26]. The Data Volume was a limitation in these studies.

The only paper that explored Data Element issues was [23], the datasets used showed that there was a consistency and standardization issue in some fields, there were also incorrect fields which were corrected when the data was preprocessed.

Although most of the datasets are multi-sourced datasets, only one paper discusses the standardization issues related to collecting each of the datasets and combining them into one [20]. After combining the datasets, there were a lot of formatting issues that needed to be corrected. Additionally, only one paper discusses the data quality issue related to the source of the Data (Data Source), since this data was collected from a single source, the data was not generalizable and hence the model would produce biased results if other datasets were applied to it, this was discussed in [21].

3.3: Application of Framework on MST Dataset

In order to evaluate the quality of the MST dataset, we will be using the Data Quality framework (Figure 1) to assess the data quality issues that exist in this dataset. We will follow the four-step process that was mentioned in section 3 to analyze data dimensions, identify data quality issues, and propose solutions for improvement.

Define: Since the dataset is single sourced, there is a clear data standard set by the MST hospital itself, therefore there is no lack of data standardization. Similarly, there are no Data

Element issues. The identified dimensions for this dataset are Data Completeness and Accuracy, and Data Credibility.

Measure: From the selected dimensions, we will then figure out which data quality issues exist in our own dataset. We will start with the Data Completeness and Accuracy dimension. After screening the data, it is obvious that portions of the data are missing, there are duplicates and that the data is imbalanced. The Data Credibility dimension only has one data quality issue related to this dataset, and that is the Data Volume. The Measure phase will explain how to possibly identify such data quality problems.

1. **Missing Data**: To determine whether there are missing values, a bar chart can simply be plotted to show the percentage of missingness per Biomarker, the results can be visualized in Figure 2.

2. Duplicate Data: A scatter plot was used in order to visualize the number of duplicate values per patient, the results can be seen in Figure 3.

3. Imbalanced Data: To determine whether or the dataset is balanced, we can look at the distribution of the COVID severity classes and determine whether they are balanced. To determine this, we find the number of patients belonging to each COVID severity class (from 0 to 4) and plot it on a bar graph to visualize the distribution. The results are in Figure 4.

4. Data Volume: Before processing the data, there are 3609 rows and 701 patients. The results after preprocessing are discussed further in section 4.1.

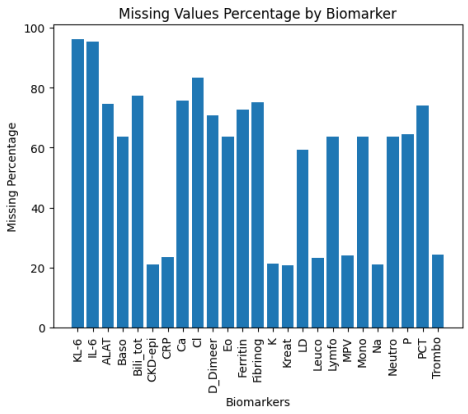

Figure 2: Bar Chart of Missing Values % per Biomarker

Analyze: The root cause of each of the mentioned data quality issues can be explained. Firstly, missing data values are represented by zeroes in the dataset. This occurs because different patients are tested for different biomarkers based on factors like severity, symptoms, and other variables. The decision to conduct specific tests depends on the patient's condition and the medical opinion of the doctor.
Secondly, duplicates exist in the dataset due to daily testing of patients throughout their stay. This repetitive testing leads to multiple entries for the same patient.
The imbalance in data is common in biological datasets, it is typical for the number of patients who test negative for the COVID virus to be higher than those to test positive.

The Data Volume after preprocessing is low, because the curation of the mentioned data quality issues eliminated a large number of rows.

Improve: This will be discussed in section 4.1 which discusses how the data quality issues will be handled.

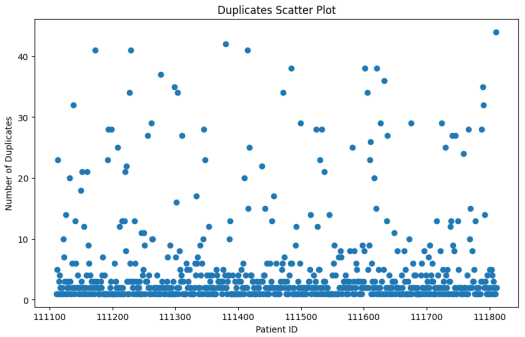

Figure 3: Scatter plot of duplicate rows per patient

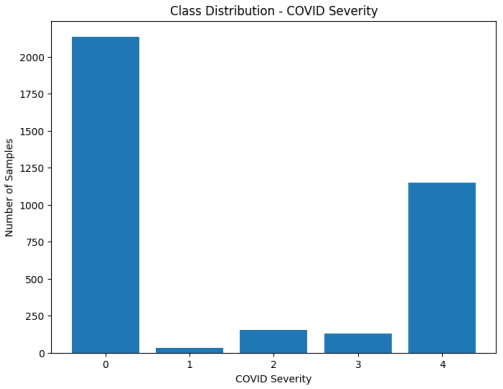

Figure 4: COVID Severity Class Distribution

3.4: Conclusion

We have developed a comprehensive framework by analyzing research findings from literature regarding Data Quality and Data Quality Frameworks. The effectiveness of this framework was evaluated by applying it to research conducted on COVID data to create ML models, as discussed in section 3.2. The research and the application of the framework on COVID data collectively highlight the alarming presence of data quality issues within COVID data, with the most cited dimension being Data Completeness and Accuracy. Specifically, an issue that consistently emerged across the discussed papers was the presence of missing data, which was observed in all the datasets mentioned in section 3.2. Imbalanced data was also a very common data quality issue in some of the datasets [19],[24]. The same issues were also present in the MST dataset as discussed in section 3.3.

In conclusion, it is evident that data quality issues exist within COVID data. Unfortunately, these issues are not adequately addressed and their impact on ML models is often overlooked and underestimated. Extensive research papers have discussed the effects of data quality issues on COVID data and various datasets, highlighting their significance. However, in research papers where data quality is not the primary focus, such as those aiming to build predictive models based on COVID

data, the mention of data quality is typically minimal despite its crucial importance. While most papers include a section on Data Preprocessing, the approaches employed are often overly simplistic, lacking in-depth exploration of potential data quality issues. Additionally, there is a noticeable absence of thorough dataset profiling in the reviewed papers. The proposed framework (Figure 1) can help in simplifying the task of dealing with Data Quality issues.

4. How Can The Severity of COVID-19 be Predicted using Biomarker Data?

Based on the findings from various studies [18]-[26], it is clear that biomarker data can be utilized to develop a predictive model for COVID-related datasets. To support this claim, we will employ our own machine learning model on the MST dataset and analyze the outcomes. These results will be thoroughly explained. However, before creating the model, it is essential to address the necessary data preprocessing steps.

4.1 Data Preprocessing

In order to predict COVID severity, patients who tested negative were excluded, indicated by a severity score of zero in the dataset. By removing these cases, a total of 2136 out of 3608 rows were eliminated from the dataset. As a result, the number of patients decreased from 701 to 133. Therefore, the only severity scores available are 1 to 4 which represent Mild, Moderate, Severe and Critical cases, respectively.

Due to a significant percentage of missing values in certain biomarkers, it is necessary to establish a cut-off point to ensure the inclusion of biomarkers with sufficient data values. Therefore, a threshold of 50% missingness will be set, and only the biomarkers below this threshold will be retained for further analysis. As a result, the following biomarkers will be retained for analysis: CKD-epi, CRP, K, Kreat, Leuco, MPV, Na, and Trombo. Moreover, certain rows in the dataset where all biomarker values were equal to zero were removed.

Duplicates exist for some patients due to daily testing that was made throughout their stay. This repetitive testing leads to multiple entries for the same patient. To tackle this issue, the duplicate values per patient were merged into a single value by calculating the average of the biomarker values. During this calculation, the value of zero was excluded to ensure accuracy. As a result of merging the duplicates, the percentage of missing values decreased significantly. The average percentage of missing values for CKD-epi, CRP, K, Kreat and Na is 4.1%. The average percentage of missing values for Leuco, MPV and Trombo is 18.4%. The number of rows after handling missing values and duplicates is now 121.

Finding the average values per biomarker may reduce accuracy due to the presence of outliers. Boxplots were used to visualize the data distribution for different biomarkers and identify outliers, revealing a considerable number of values falling outside the normal range. Defining outliers in biological datasets is challenging as they may indeed represent valid values, emphasizing the importance of developing enhanced methods for outlier evaluation in future studies.

One more important data quality issue that exists in this dataset is Data imbalance as mentioned in section 3.3. To resolve this, we will be balancing the data by applying the SMOTE algorithm. The success of SMOTE was seen in [19],[24].

Before the dataset imbalance issue is addressed, an extra preprocessing step was implemented to improve resource allocation decisions. Our goal was to determine if a patient's blood test results could effectively classify them as either severe or non-severe. Since there were only two primary classes, a decision was made to merge COVID severity classes 1, 2, and 3, 4 into two distinct categories: non-severe and severe. This allowed us to transform the problem into a binary classification task, simplifying the analysis and decision-making process.

4.2 Machine Learning Model

To predict COVID severity, we employed a machine learning model known as a Support Vector Machine (SVM). SVM models showed great results when predicting COVID severity in [18] and [25] and had one of the highest result metrics out of all the datasets [18]-[26]. We employed a Support Vector Classifier (SVC) model with a linear kernel, this is suitable because we transformed the problem into a binary classification task with two classes (non-severe and severe). The dataset was split into a training set and a testing set using a 70/30 ratio to assess how well the model performs to unseen data.

Along with the SVM model, we also calculated the feature importance for each biomarker. Feature importance refers to techniques that assign a score to input features based on how useful they are at predicting a target variable. Permutation importance was used to find out which features affected the severity the most. The feature importance of each biomarker can be seen in Figure 5.

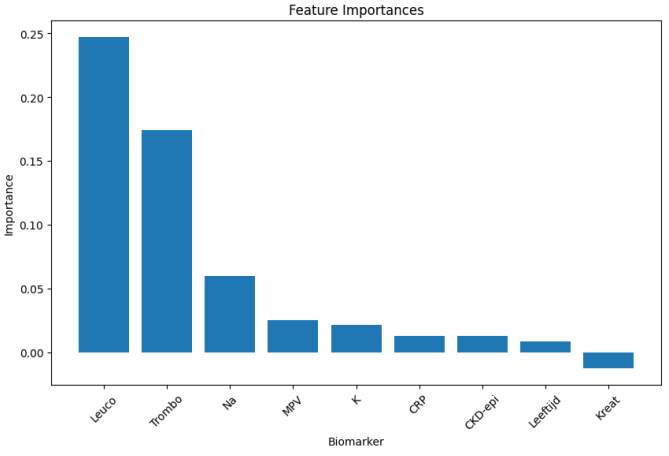

Figure 5: Permutation Importance based on Biomarker

The performance of the SVM model in predicting COVID severity can be evaluated by analyzing the Confusion Matrix shown in Figure 6. The Confusion Matrix provides a comprehensive summary of the model's predictions and actual outcomes. From the matrix, we observe an overall Accuracy

of 76.19%, indicating the proportion of correct predictions made by the model. A Precision score of 90.91% which indicates the model's ability to accurately identify the non-severe and severe classes, out of all the positive predictions made. The Recall score is 68.97%, representing the proportion of actual positive instances correctly classified by the model. Finally, the F1 Score, which combines Precision and Recall into a single metric, is calculated to be 78.57%.

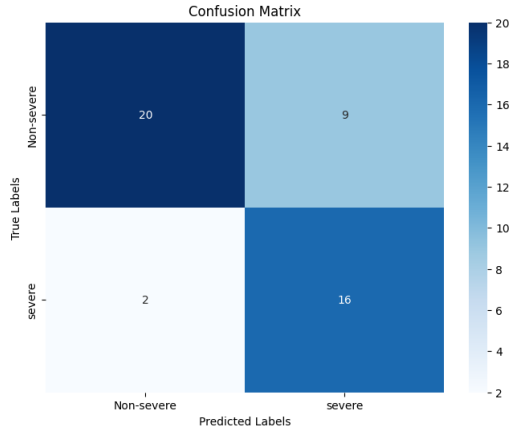

Figure 6: Confusion Matrix of SVM Model

4.3 Conclusion

Our model yielded promising results in predicting the severity of COVID, highlighting the potential correlation between two biomarkers: Leukocytes (Leuco) and Thrombocytes (Trombo). These biomarkers exhibited the highest feature importance, as depicted in Figure 5. Notably, Leukocytes have been identified as indicators of severity in other research papers [28], [29]. They are widely used, inexpensive biomarkers of inflammation [28], further supporting our model's findings. Furthermore, our model's accuracy of 76% on a larger number of test cases (121 vs 53) performed similarly to another SVM model with 80% accuracy [18]. Our model demonstrated good performance on a larger number of unseen data. In another study [25], an SVM model achieved an accuracy of 81.48%, comparable to our results. It's worth noting that their dataset had more features (26 vs 8) for training, indicating a considerable amount of additional data available for comparison which was lacking in the MST dataset.

It is important to note that drawing such a definitive conclusion is difficult to prove without considering the other 17 biomarkers that have been removed due to their missingness in the dataset. Some research suggests that biomarkers such as KL-6 [30],[31] and PCT [32]-[34] can be used as severity indicators. The absence of these biomarkers in our dataset hinders the ability to determine what key biomarkers correlate most with COVID severity. The absence of the relevant biomarkers for different patients as well as the Data Volume of this dataset after preprocessing is the limitation of this study, and more useful information could have been derived if the data was more complete. The success of this model in predicting COVID severity is due to the improvements made in the quality of the COVID data, demonstrating the crucial role of data quality in influencing machine learning model results. Moreover, it is also important to mention that the COVID severity labels assigned to the patients do not reflect the initial severity of the disease but rather the severity after the patient's outcome was already determined. This distinction is important because it means that the severity labeling does not accurately represent the true state of the disease at the onset. As a result, it is highly likely that these misleading severity labels had a substantial impact on the model's results, potentially skewing them significantly. Such a problem would also affect the ability to allocate resources effectively.

5. How can COVID-19 disease progression improve resource allocation decisions for patient management?

To understand the progression of COVID, it is crucial to define "disease progression" in the context of COVID. Disease progression refers to the transition from a mild or moderate stage of the disease to a more severe stage [35].

It is evident that there exists a correlation between COVID progression and the severity of the disease. In order to investigate whether COVID-19 disease progression can be predicted, it is necessary to address a specific sub-research question:

- RQ3.1: Can the severity of COVID help forecast its progression?

Having defined COVID disease progression, it is also important to define COVID severity. COVID severity refers to the intensity and seriousness of the disease's impact on an individual's health, categorized as mild, moderate, severe, or critical. COVID severity is a component of disease progression.

Consequently, if the severity of the disease in a patient is known, it becomes possible to predict the trajectory and progression of the disease's outcome. A meta-analysis will be conducted to collect and analyze data from multiple studies, aiming to determine the outcomes of patients at different levels of COVID severity. Another method that could be used to predict COVID progression is to use Machine Learning. If the outcome of the disease in the patient is known and is part of a dataset, it could be possible to determine if the disease will progress to a more severe case based on biomarker data, such a model was created in [27]. However, since the MST dataset does not contain the outcomes of patients, such a model would be impossible to create because the data needed to create this model is unavailable, which is why a meta-analysis will be used along with the statistical data it will provide to see how COVID severity can affect its progression. The results of the meta-analysis can be seen in Table 1.

The meta-analysis included 7169 patients which have been positively tested for COVID-19 with varying severity levels from mild-moderate to critical. 1172 patients had mild-moderate severity, 8% of these patients died and 77% recovered fully, while the rest had unchanged conditions. 583 patients had Severe COVID-19, out of which 21% passed away and 75% recovered fully. Critical COVID cases were the most reported in this meta-analysis study, with 5414 patients. Around 39% of these patients died and 52% recovered, with some papers reporting complications even after recovery. It is evident from this study that as the severity of the disease increases, the percentage of deaths increase, and the

percentage of recoveries decrease. Hospitals can use such information to improve the allocation of resources, higher priority can be given to those that show more severe reactions to the virus. Mild-moderate patients do not reach the stage where they need to be hospitalized, so they can be dismissed.

Paper	Severity	No. of Patients	1*	2*	3*
[36]	Mild-Moderate	869	616	158	95
[37]	Mild-Moderate	303	289	10	2
[38]	Severe	275	205	0	70
[39]	Severe	260	183	52	52
[40]	Critical	1616	842	391	383
[41]	Critical	164	79	0	85
[42]	Critical	468	275	38	155
[43]	Severe	48	47	0	1
	Critical	26	13		13
[44]	Critical	3140	1594	0	1483

1: Represents the number of patients who have been discharged or are in the process of recovering.*
2: Represents the number of patients who have not experienced a change in their condition.*
3: Represents the number of patients who have had a poor disease progression/death.*

Table 1: Outcome Evaluation of Mild-Moderate, Severe and Critical Patient

In conclusion, we have discovered a link between the severity of COVID and how the disease progresses and affects patients. By examining data from numerous studies, we can predict the outcome of a patient if we know the severity of their condition. Patients with more severe cases of COVID are more likely to have a poorer prognosis compared to those with milder cases.

However, the method used in this section to determine how COVID severity impacts disease progression may not be ideal. It is important to note that even patients with mild symptoms can experience a significant deterioration in their condition. As previously mentioned, the study found that 8% of patients initially classified as having mild to moderate symptoms unfortunately passed away. This finding highlights the risk of relying solely on statistics. If hospitals were to solely rely on statistics, they might discharge patients who actually have a high risk of poor disease prognosis but initially exhibited mild to moderate symptoms. This approach could be extremely risky and could potentially result in the loss of many lives.

A more effective approach to measure disease progression in patients would involve the use of an ML model. By utilizing such a model, it becomes possible to predict how the disease will progress and what the outcome is likely to be, based on factors like severity and biomarker data. However, the success of this method depends on the availability of relevant data. This is how Data Quality plays a significant role.

For instance, if we consider the MST dataset, it would not be suitable for training an ML model because it lacks a specific column indicating the outcome of the disease in patients. Without this crucial information, the model would not be able to learn and make accurate predictions. Therefore, it is essential to have a comprehensive dataset that encompasses the necessary variables in order to successfully develop and train an ML model for this purpose.

6. Conclusion

This research project's main focus was to determine the impact that Data Quality has on severity prediction models, and how this can ultimately affect resource allocation for COVID-19 patient management.

The first research question identified the most prevalent data quality issues in COVID datasets as well as the MST dataset; the quality of these datasets were assessed using a Data Quality Framework (Figure 1) that we created. The framework provided a systematic and structural way to assess the data.

For the second research question, we identified that Machine Learning methods can be used in order to predict the severity of COVID in patients. After researching different machine learning models, we decided that a Support Vector Machine (SVM) was the most appropriate to create a COVID severity prediction model. The model was applied to the MST dataset and the results of the model were promising. Such models can be used by physicians to predict the severity of COVID in patients that in turn enables them to decide whether a patient should be admitted for further monitoring or discharged since there is no severe risk, which ultimately improves the utilization of resources in hospitals. However, the results of the model could not have been achieved if the necessary preprocessing techniques had not been applied. Preprocessing was necessary due to the abundant data quality issues that existed in the dataset, this helps show the importance of data quality in achieving accurate and reliable results for a Machine Learning model, this also answers the main research question of this study.

The final research question highlighted the correlation between COVID severity and COVID progression, we also determined how resources can be allocated based on how the disease will progress.

To conclude, we identified common data quality issues in COVID data by implementing a Data Quality Framework. We also established that machine learning models can be effective in predicting COVID severity using patient biomarker data, with the model's accuracy relying on the quality of the data.

Furthermore, we discovered a correlation between COVID severity and disease progression, where a higher severity level corresponds to a higher likelihood of poor disease progression.

7. Future Work

There are still many areas of research that have not been explored in this project. It would be interesting to focus on identifying which biomarkers are most closely related to COVID severity using a more complete dataset. Unfortunately, 17 biomarkers had to be removed from the MST dataset because there was not enough information about them. Another idea that could improve resource allocation efficiency is to create a model that predicts how the disease progresses based on biomarker data and severity labels. However, this would require a detailed dataset that shows what happened to different patients. Exploring these research areas further could provide valuable insights into the COVID-19 virus and how Machine Learning methodologies could potentially help reduce its spreading by improving how resources can be allocated.

8. Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Faizan Ahmed, for his invaluable guidance and support throughout the course of this project. Dr. Ahmed's expertise and insightful advice played a pivotal role in shaping the direction of my research. I would also like to extend my appreciation to Jeewanie Jayasinghe Arachchige, whose support greatly contributed to the successful completion of this project. Her expertise in the subject and her willingness to share their knowledge have been instrumental in enhancing my understanding and skills in different research areas.

References

[1] LeDuc JW, Barry MA. SARS, the First Pandemic of the 21st Century. Emerg Infect Dis. 2004 Nov;10(11):e26. doi: 10.3201/eid1011.040797_02. PMCID: PMC3329048.

[2] Muralidar S, Ambi SV, Sekaran S, Krishnan UM. The emergence of COVID-19 as a global pandemic: Understanding the epidemiology, immune response and potential therapeutic targets of SARS-CoV-2. Biochimie. 2020 Dec;179:85-100. doi: 10.1016/j.biochi.2020.09.018. Epub 2020 Sep 22. PMID: 32971147; PMCID: PMC7505773.

[3] Nistha Shrestha, Muhammad Yousaf Shad, Osman Ulvi, Modasser Hossain Khan, Ajlina Karamehic-Muratovic, Uyen-Sa D.T. Nguyen, Mahdi Baghbanzadeh, Robert Wardrup, Nasrin Aghamohammadi, Diana Cervantes, Kh. Md Nahiduzzaman, Rafdzah Ahmad Zaki, Ubydul Haque, The impact of COVID-19 on globalization, One Health, Volume 11, 2020, 100180, ISSN 2352-7714, https://doi.org/10.1016/j.onehlt.2020.100180.

[4] WHO coronavirus (covid-19) dashboard.: https://covid19.who.int/table.

[5] Yuk-Chiu Yip J. Healthcare resource allocation in the COVID-19 pandemic: Ethical considerations from the perspective of distributive justice within public health. Public Health Pract (Oxf). 2021 Nov;2:100111. doi: 10.1016/j.puhip.2021.100111. Epub 2021 Mar 28. PMID: 33817679; PMCID: PMC8005252.

[6] Peffers, Ken & Tuunanen, Tuure & Rothenberger, Marcus & Chatterjee, S.. (2007). A design science research methodology for information systems research. Journal of Management Information Systems. 24. 45-77.

[7] Setia MS. Methodology Series Module 1: Cohort Studies. Indian J Dermatol. 2016 Jan-Feb;61(1):21-5. doi: 10.4103/0019-5154.174011. PMID: 26955090; PMCID: PMC4763690.

[8] T. Nae, J. Krabbe, F. A. Bukhsh, J. Jayasinghe Arachchige and F. Ahmed, "Covid severity prediction: Who cares about the data quality?," 2022 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 2022, pp. 225-230, doi: 10.1109/FIT57066.2022.00049.

[9] Fifth Review of the fund's Data Standards Initiatives -- Data Quality Assessment Framework and Data Quality Program. International Monetary Fund. (n.d.). https://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm

[10] Francisco, Maritza & Souza, Solange N A & Campos, Edit & Souza, Luiz. (2017). Total Data Quality Management and Total Information Quality Management Applied to Customer Relationship Management. 40-45. 10.1145/3149572.3149575.

[11] Baskarada, Sasa & Koronios, Andy & Gao, Jing. (2006). Towards a Capability Maturity Model for Information Quality Management: A TDQM Approach. 499-510.

[12] Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(0), p.2.DOI: https://doi.org/10.5334/dsj-2015-002

[13] Rachel L. Richesson, PhD, MPH , Jeffrey Krischer, PhD, Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions, Journal of the American Medical Informatics Association, Volume 14, Issue 6, November 2007, Pages 687–696, https://doi.org/10.1197/jamia.M2470

[14] Rachel L Richesson , Prakash Nadkarni, Data standards for clinical research data collection forms: current status and challenges, Journal of the American Medical Informatics Association, Volume 18, Issue 3, May 2011, Pages 341–346, https://doi.org/10.1136/amiajnl-2011-000107

[15] Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402-6. doi: 10.4097/kjae.2013.64.5.402. Epub 2013 May 24. PMID: 23741561; PMCID: PMC3668100.

[16] Chowdhury, A., & Alspector, J. (2003). Data duplication: an imbalance problem?. In ICML 2003 workshop on learning from imbalanced data sets (II), Washington, DC.

[17] Mazumder, S. (2022, December 1). 5 techniques to handle imbalanced data for a classification problem. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/

[18] X. Jiang, M. Coffee, A. Bari, et al., "Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity," Computers, Materials and Continua, vol. 62, no. 3, pp. 537–551, 2020. DOI: https://doi.org/10.32604/cmc.2020.010691

[19] M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, "Ensemble learning model for diagnosing covid-19 from routine blood tests," Informatics in Medicine Unlocked, vol. 21, p. 100 449, 2020. DOI: 10.1016/j.imu.2020.100449

[20] M. Nemati, J. Ansary, and N. Nemati, "Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data," Patterns, vol. 1, no. 5, p. 100 074, 2020. DOI: 10.1016/j.patter.2020.100074.

[21] T. J. Levy, S. Richardson, K. Coppa, et al., "A predictive model to estimate survival of hospitalized covid-19 patients from admission data," 2020. DOI: 10.1101/2020.04. 22.20075416.

[22] Y. Li, M. A. Horowitz, J. Liu, et al., "Individual-level fatality prediction of covid-19 patients using ai methods," Frontiers in Public Health, vol. 8, 2020. DOI: 10.3389/ fpubh.2020.58793

[23] C. Costa-Santos, A. Luísa Neves, R. Correia, et al., "Covid-19 surveillance - a descriptive study on data quality issues," 2020. DOI: 10.1101/2020.11.03.20225565.

[24] M. Naseem, H. Arshad, S. A. Hashmi, F. Irfan, and F. S. Ahmed, "Predicting mortality in sars-cov-2 (covid-19) positive patients in the inpatient setting using a novel deep neural network," International Journal of Medical Informatics, vol. 154, p. 104 556, 2021. DOI: 10.1016/j.ijmedinf.2021.104556.

[25] H. Yao, N. Zhang, R. Zhang, et al., "Severity detection for the coronavirus disease 2019 (covid-19) patients using a machine learning model based on the blood and urine tests," Frontiers in Cell and Developmental Biology, vol. 8, 2020. DOI: 10.3389/fcell. 2020.00683.

[26] M. Laatifi, S. Douzi, A. Bouklouz, et al., "Machine learning approaches in covid-19 severity risk prediction in morocco," Journal of Big Data, vol. 9, no. 1, 2022. DOI: 10.1186/s40537-021-00557-0.

[27] Wang, M., Wu, D., Liu, CH. *et al.* Predicting progression to severe COVID-19 using the PAINT score. *BMC Infect Dis* 22, 498 (2022). https://doi.org/10.1186/s12879-022-07466-4

[28] Ramos-Hernández, W.M., Soto, L.F., Del Rosario-Trinidad, M. et al. Leukocyte glucose index as a novel biomarker for COVID-19 severity. Sci Rep 12, 14956 (2022). https://doi.org/10.1038/s41598-022-18786-5

[29] Hottz, ED, Bozza, PT. Platelet-leukocyte interactions in COVID-19: Contributions to hypercoagulability, inflammation, and disease severity. Res Pract Thromb Haemost. 2022; 6:e12709. doi:10.1002/rth2.12709

[30] d'Alessandro M, Cameli P, Refini RM, Bergantini L, Alonzi V, Lanzarone N, Bennett D, Rana GD, Montagnani F, Scolletta S, Franchi F, Frediani B, Valente S, Mazzei MA, Bonella F, Bargagli E. Serum KL-6 concentrations as a novel biomarker of severe COVID-19. J Med Virol. 2020 Oct;92(10):2216-2220. doi: 10.1002/jmv.26087. Epub 2020 Jun 9. PMID: 32470148; PMCID: PMC7283867. Serum KL-6 concentrations as a novel biomarker of severe COVID-19 - PMC

[31] Maruyama, S., Nakamori, Y., Nakano, H. *et al.* Peak value of serum KL-6 may be useful for predicting poor prognosis of severe COVID-19 patients. *Eur J Med Res* 27, 69 (2022). https://doi.org/10.1186/s40001-022-00690-3

[32] Hodges G, Pallisgaard J, Schjerning Olsen AM, McGettigan P, Andersen M, Krogager M, Kragholm K, Køber L, Gislason GH, Torp-Pedersen C, Bang CN. Association between biomarkers and COVID-19 severity and mortality: a nationwide Danish cohort study. BMJ Open. 2020 Dec 2;10(12):e041295. doi:

10.1136/bmjopen-2020-041295. PMID: 33268425; PMCID: PMC7712929. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7712929/
[33] Bivona G, Agnello L, Ciaccio M. Biomarkers for Prognosis and Treatment Response in COVID-19 Patients. Ann Lab Med. 2021 Nov 1;41(6):540-548. doi: 10.3343/alm.2021.41.6.540. PMID: 34108281; PMCID: PMC8203437.
[34] Danwang, C., Endomba, F.T., Nkeck, J.R. *et al.* A meta-analysis of potential biomarkers associated with severity of coronavirus disease 2019 (COVID-19). *Biomark Res* 8, 37 (2020). https://doi.org/10.1186/s40364-020-00217-0
[35] Cen Y, Chen X, Shen Y, Zhang XH, Lei Y, Xu C, Jiang WR, Xu HT, Chen Y, Zhu J, Zhang LL, Liu YH. Risk factors for disease progression in patients with mild to moderate coronavirus disease 2019-a multi-centre observational study. Clin Microbiol Infect. 2020 Sep;26(9):1242-1247. doi: 10.1016/j.cmi.2020.05.041. Epub 2020 Jun 9. PMID: 32526275; PMCID: PMC7280135.
[36] Zhang J, Wang M, Zhao M, Guo S, Xu Y, Ye J, Ding W, Wang Z, Ye D, Pan W, Liu M, Li D, Luo Z, Liu J, Wan J. The Clinical Characteristics and Prognosis Factors of Mild-Moderate Patients With COVID-19 in a Mobile Cabin Hospital: A Retrospective, Single-Center Study. Front Public Health. 2020 Jun 5;8:264. doi: 10.3389/fpubh.2020.00264. PMID: 32582615; PMCID: PMC7291856.
[37] James, E., Wanume, B., Musaba, M.W. et al. Characteristics, treatment outcomes and experiences of COVID-19 patients under home-based care in Kapelebyong district in Uganda: a mixed-methods study. Trop Med Health 50, 93 (2022). https://doi.org/10.1186/s41182-022-00486-5
[38] Neville TH, Hays RD, Tseng CH, Gonzalez CA, Chen L, Hong A, Yamamoto M, Santoso L, Kung A, Schwab K, Chang SY, Qadir N, Wang T, Wenger NS. Survival After Severe COVID-19: Long-Term Outcomes of Patients Admitted to an Intensive Care Unit. J Intensive Care Med. 2022 Aug;37(8):1019-1028. doi: 10.1177/08850666221092687. Epub 2022 Apr 5. PMID: 35382627; PMCID: PMC8990100.
[39] Wang, J., Zheng, X. & Chen, J. Clinical progression and outcomes of 260 patients with severe COVID-19: an observational study. Sci Rep 11, 3166 (2021). https://doi.org/10.1038/s41598-021-82943-5
[40] Martin-Villares, C., Perez Molina-Ramirez, C., Bartolome-Benito, M. et al. Outcome of 1890 tracheostomies for critical COVID-19 patients: a national cohort study in Spain. Eur Arch Otorhinolaryngol 278, 1605–1612 (2021). https://doi.org/10.1007/s00405-020-06220-3
[41] Silvio A. Ñamendys-Silva, Pedro E. Alvarado-Ávila, Guillermo Domínguez-Cherit, Eduardo Rivero-Sigarroa, Luis A. Sánchez-Hurtado, Alan Gutiérrez-Villaseñor, Juan P. Romero-González, Heber Rodríguez-Bautista, Alondra García-Briones, César E. Garnica-Camacho, Néstor G. Cruz-Ruiz, María O. González-Herrera, Francisco J. García-Guillén, Manuel A. Guerrero-Gutiérrez, José D. Salmerón-González, Laura Romero-Gutiérrez, José L. Canto-Castro, Victor H. Cervantes, Outcomes of patients with COVID-19 in the intensive care unit in Mexico: A multicenter observational study, Heart & Lung, Volume 50, Issue 1, 2021, Pages 28-32, ISSN 0147-9563, https://doi.org/10.1016/j.hrtlng.2020.10.013.
[42] Anesi GL et al. Characteristics, Outcomes, and Trends of Patients With COVID-19-Related Critical Illness at a Learning Health System in the United States. Ann Intern Med. 2021 May;174(5):613-621. doi: 10.7326/M20-5327. Epub 2021 Jan 19. PMID: 33460330; PMCID: PMC7901669.
[43] Junli Li, Ge Xu, Heping Yu, Xiang Peng, Yongwen Luo, Cheng'an Cao, Clinical Characteristics and Outcomes of 74 Patients With Severe or Critical COVID-19, The American Journal of the Medical Sciences, Volume 360, Issue 3, 2020, Pages 229-235, ISSN 0002-9629, https://doi.org/10.1016/j.amjms.2020.05.040.
[44] Chisala et al, Patient care and clinical outcomes for patients with COVID-19 infection admitted to African high-care or intensive care units (ACCCOS): a multicentre, prospective, observational cohort study, The Lancet, Volume 397, Issue 10288, 2021, Pages 1885-1894, ISSN 0140-6736, https://doi.org/10.1016/S0140-6736(21)00441-4.