Bio Marker Based COVID Severity Prediction and Data Quality Analysis

Veselin Daskalov[Author], Dr. F.A. Bukhsh[Supervisor], and Dr. F. Ahmed[Critical Observer]

University of Twente
EEMCS
Enschede, Netherlands

July 22, 2022

Contents

Abstract

Introduction

This paper explores the data quality issues and State of the Art of the COVID-19 diagnostics, mortality prediction and severity prediction to create a Machine Learning model that can be used by hospitals in order to properly manage resources due to the strain the pandemic forced on the healthcare system of the Netherlands.

Method

To facilitate our goals, we use the Kitchenham method to perform and Systemic Literature Review used to identify the common data quality problems in COVID-19 data sets, along with the current State of the Art. After which a Machine Learning model is formed using Logistic + LASSO Regression for Feature Selection, SMOTE for data rebalancing and Binary Decision Trees for predicting the severity of a patient.

Results

Our paper shows that the positive predictors for high severity are; increased age, MPV, and high CRP, elevated Trombo. While negatively linked are decreased age (data set cannot account for below 18 years old), Na, K, Kreat and Leuco. With the most common data quality issues found are missing data, data format incongruency and data imbalance. Furthermore, our paper also shows a negative result on our predictive model, with it performing worse than most models found in the State of the Art. Our model scores an Accuracy of 79.9%, Precision at 69.0%, Recall at 72.2% and an F1 score of 70.6%.

Discussion and Conclusion

Our paper shows that Machine Learning models can be effective in predicting mortality and severity in COVID-19 patients. We believe that given a high-quality data set along with implementation of a more nuanced pre-processing step our Machine Learning model can achieve the metrics required to be an effective tool in alleviating the pressures on the healthcare system. Nevertheless, we believe that our research shows the potential of Machine Learning based tools and encourage further study into the topic.

Keywords Machine Learning, COVID-19, Severity Prediction, Data Quality, Kitchenham, Literature Review.

Chapter 1

Introduction

From its start, the SARS-CoV-2 (COVID-19) pandemic has generated more than 400 million cases and has resulted in a mortality level in excess of 5.5 million since the start [1]. Studies have shown that both age, gender and pre-existing health issues can become contributing factors to an individual's survival rate when infected with the virus [2]. Thus, it is imperative to accurately predict severe cases of the infection so that so that the proper medical attention can be given, and the resources used more efficiently.

One of the fields that has shown to produce prediction results consistent with medical studies is Machine Learning (ML) [3]. However, it is heavily dependent on the quality of data used to train it. Having biased or limited data would result in poor predictive power [3].

This paper aims to determine what are the data quality issues associated with COVID-19 data sets along with which pre-processing techniques are used to improve it. This is to be done by exploring the different ways pre-existing research, focused on predicting severity of COVID-19 cases, and how data and data quality were handled and pre-processed. Since we are concerned with both the data and techniques used on it, this paper will focus only on material that provides the data set used in training and pre-processing.

The paper will then go on to create its own implementation of an ML based COVID-19 severity prediction model, that can be used by hospitals in order to streamline resource allocation. It is hoped that the implementation will be able to ease the strain on the healthcare system during the pandemic.

Chapter 2

Research Method

This chapter will focus on the Literature Review along with the State of the Art within the sphere of COVID-19 data sets and ML models. The aim of this chapter is to establish the most effective means and methods for dealing with COVID-19 and building the new models, with the goal of setting up the implementation stage of the research.

2.1 Research Questions

The Research Questions (RQs) in this stage are aimed at to further the goal of understanding the pre-existing academic literature within the sphere of COVID-19. As such the following research questions were constructed:

1. What is the data parameter quality reporting in COVID-19 papers?
2. What is the impact of data quality on ML and it reported enough?
3. Are the results of COVID-19 research papers reproducible?

2.2 Approach

The approach that we will use in order to satisfy the goals set out in the start of the chapter is by using the Systemic Literature Review (SLR) technique "Kitchenham" [4]. The Kitchenham method identifies 3 3 phases of the literature review, them being "Planning the Review", "Conducting the Review", and "Documenting the Review". The Kitchenham method calls for the employment of multiple academic article selection based on a constant key word search strategy, over multiple databases.

For purpose of our review is to both acquaint ourselves with the data quality issues of COVID-19 data sets along with their implications and to ascertain the current "State of the Art". Thus, the keywords chosen for our search will be:

1. Machine Learning
2. COVID-19
3. Bio-Marker

The databases chosen are; ScienceDirect and Scopus accessed via the University of Twente databases portal. The process of the academic paper selection can be found below in figure 2.1.

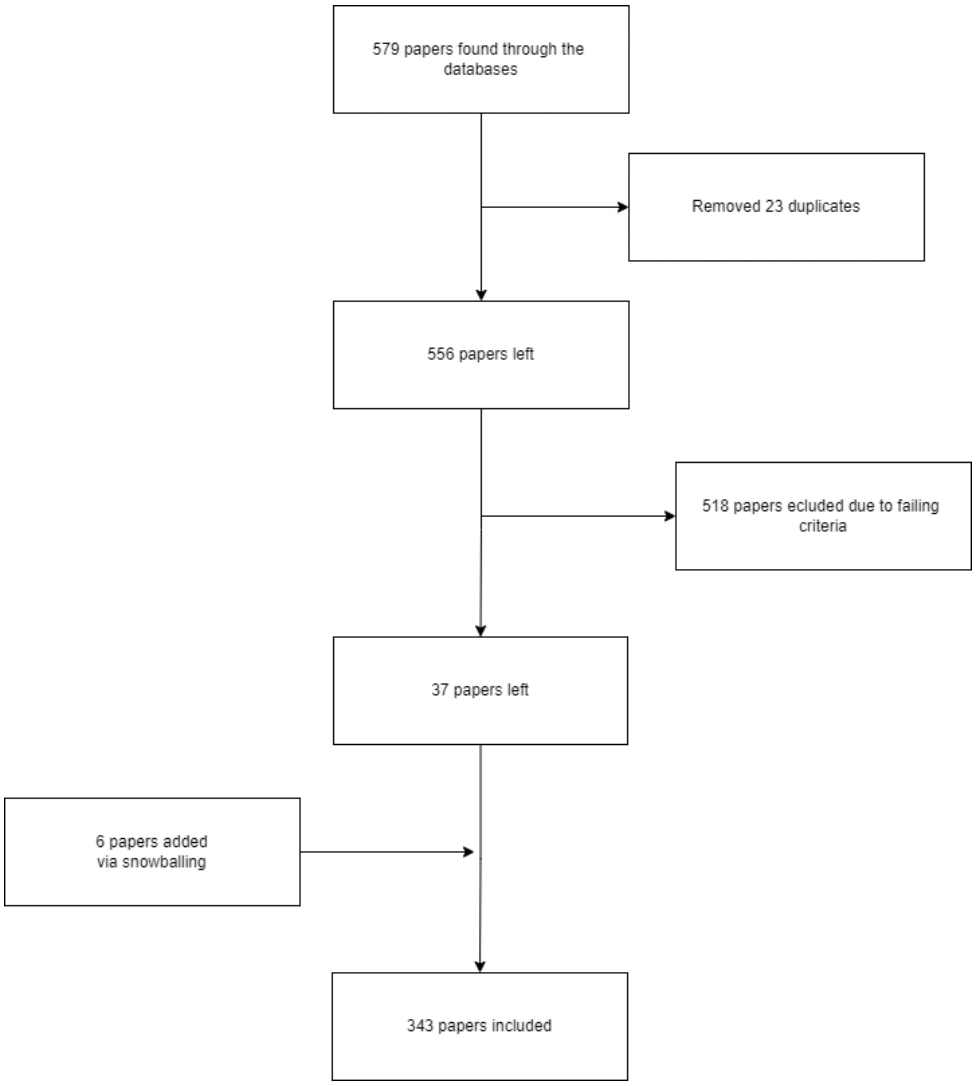

Figure 2.1: Flow Chart Depicting Academic Paper Selection for Literature Review

2.3 Literature Review

2.3.1 Introduction

As mentioned in the introduction the COVID-19 pandemic had both economic [5] and sociological [6] consequences. The Netherlands is no exception with lockdowns causing closing down the catering and leisure industries along with schools, universities and even gyms. The all had a measurable negative effect on the health [6]. Given that the one of the main reasons stated by the governmental bodies, for the lockdowns, is the potential lack resource such as ventilators and ICU beds, thus the reason for a system that can aid in the efficient distribution of these limited resources.

However, for an ML system to have high enough accuracy factors need to be taken into account. Data quality, quantity, and balance along with the model and training set can all

influence the accuracy of an ML implementation. Thus, it is vital for these variables to be taken into consideration. Therefore, a Systemic Literature Review will be performed in order to ascertain the common challenges COVID-19 data presents along with any remedies applied within the current sphere of research

2.3.2 Results

What is the data parameter quality reporting in COVID-19 papers?

The data quality parameter reporting refers to the data quality problems and their reporting, specifically how much they are reported. This is important since this will form the basis for both the subsequent RQs along with the expected steps that are needed to be taken for correcting any data quality issues. From the research done, we have found that that COVID-19 data sets suffer from three main quality problems: Intra-Set, Inter-Set and Balance problems.

Intra-Set quality problems are those that only effect a single data set. These problems typically manifest themselves as missing or non-consistently formatted data points as well duplicate or inconsistent data. These types errors can and do cause problems for ML solutions since they can reduce training and validations data, class diversity and can bias the final model. [7] [8]. Furthermore, format variation or inconsistent can also skew results since some classes might gain an artificially higher magnitude due to this and must be corrected. Direct examples of this can be seen in Levy et. al. [9, 10]

Inter-set quality problems include all problems with Inter-Set quality problems; however, these problems are distributed between multiple data steps that are linked with each other. Most of the time when talking about multiple data sets, we either refer to cumulative data sets that are released monthly and or independent data sets that are looked at with the aim for research. The standardization of data becomes very important, Li et. al. [11] shows this well with the usage of two independent data sets. The smaller, Wolfram data set, has more bio-marker data points that the far larger, GitHub, data set. Furthermore, inconsistency between cumulative data sets with data suddenly changing between monthly data sets as described by Costa-Santos et. al. [12].

The final data quality issue of note is data balance. Data balance refers to the distribution of data points per category or class. In COVID-19, balance problems can be seen mainly in mortality [13] along with severity numbers with many data sets having a disparity between severities [14, 15]. As Naseem et. al. [13] points out data balance problems are quite common in biological data sets and regularly need to be corrected for via techniques such as undersupplying or oversampling, with techniques such as SMOTE being the most common within the set of papers retrieved though some papers elected to create a cut-off date in order to balance the patient numbers like Cabitza et. al. [16].

Given the data quality issues mentioned above the, from the research done during the literature review the most prevalent data quality issues seems to be missing data along with balance issues. References to missing data and subsequent corrections for it can be found in [7], [9]–[11], [14]–[42] and the same goes for data balance. Furthermore, there are additional data quality issues that do not fall within the three categories specified. These data quality issues mostly involve the small sample sizes of the data sets along with a lack of validation data, a problem stemming from the small sets of data [16], [22], [23], [31], [33], [37], [41], [43].

In summary, COVID-19 are prone to a numerous data quality issues at current time. Of the major problems missing data, due to non-standardization of test, and data imbalance are the most prevalent. These two issues cover the vast majority of the data quality problems found within COVID-19 data sets. There are additions issues like the small group of patients within the data set leading to a lack of validation data or limited diversity of patients due to the origin of the data set.

What is the impact of data quality on ML and is it reported enough?

After we have identified the most common problems with the data quality of COVID-19 data sets, it is to ascertain how these problems effect ML models. This RQ will be vital in the implementation stage of this paper later, as it will show which Feature Selection and Preprocessing and Processing steps works best for the given data quality concerns.

The first data problems that needs to be addressed is that of the missing data. Missing data refers to the data points within the class that are not populated by any data, most of the time this will take the form of a blank spot in the data set, though for numeric data a 0 or -1 can also be seen. Missing data can cause loss of predictive power during the analysis part of the ML solution. As Laatifi et. al. [17] states "Data analysis outputs may be wrong and erroneous if missing values are not handled, resulting in bias in later phases and inadequate models used in decision-making processes.", this assertion is also mentioned as well in Li et. al. [11]. In Li et. al. problems with missing data arose due to using two data sets, with the later and larger data set not containing certain bio-markers that the more detailed one they used.

It can thus be concluded that missing data has a large impact on ML algorithms, thus papers try to minimize missing data. To achieve this two main approached are normally taken, first is the exclusion of the data based on a certain threshold, for example 70% missing data, during pre-processing. This tactic is used within the following papers [9], [11], [13], [17], [25], [35], [37], [38] with the rest of the papers choosing to use a data imputation approach with two notable examples being data imputation based in means with Soltan et. al. [23] and imputation using nearest neighbor and Iterative Imputation in Goodman-Meza et. al. [24]. There is also a third though less common technique which involve value subsection with the with the substitute value being either a 0 or -1 such as with Yao et. al. [15] and Yan et. al. [44].

An additional issue that is linked, though not the same as, Missing Data issue is the variance of format, or data format inconsistency. This refers to variables not having the same range as other classes, which can result in overrepresentation of the higher range classes. would have the same impact as missing data. Three papers specifically, note and deal with this issue, Laatifi et. al [17], Nemati et. al. [9] and Levy et. al. [10], with Levy et. al. specifically using Z-Score normalization.

The second large data quality issue is that of data balance. Though as stated by Naseem et al. [13] that biological data sets were highly imbalanced there were more than a few occasions where the paper reported balance issues or did not specify any. The papers in question are [9], [10], [14], [20], [24], [26], [27], [31], [34], [35], [37]–[40], [42]–[47], Cabitza et. al. [16] also does not report balance problems within the data though this is due them creating a cutoff point in order to preserve data balance. It is in fact only Cabitza et. al. that reports to use this type balancing the rest either use a form of oversampling or undersupplying with Nasseem

et. al. [13] and Soares [25] using SMOTE for data balancing. Missing data can lead to heavy biasing during both Feature Selection and Analysis stages, this is due to a certain class being overrepresented within the training data set along with the data set. Such biasing can lead to large deviations in the predictive power of ML solutions [13].

The rest of the quality issues within the data sets being mentioned within the "Limitations" sections of the papers, with it being stated that the generalization of the ML model would suffer due to this.

To conclude, of the two main data quality issue concerning data COVID-19 data sets, are missing data and data imbalance, with the missing data being reported in almost every paper. Both issues can cause bias and erroneous prediction during the Analysis and Feature Selection step of the ML model which would lead to a loss in predictive power. Though as mentioned earlier missing data has been reported, or stated as not being a factor expressly, in most papers this is not the case for data imbalance. Almost 40% of the papers do not report any data balance issues not concerning which is a typical of biological datasets. Data balance is not the only issues that have a lacking in reporting, all but 3 papers mention any problems with data format consistency and most issues with selection bias or model generalization are put in the "Limitations" section of the paper. Thus, it can be concluded that for some issues that affect the predictive power of the ML model, the report is lacking.

Are the results of COVID-19 research papers reproducible?

The final part of the Literature Review concerns itself with the replicability of the COVID-19 research papers. This is being asked in order to ascertain the general quality of the data surrounding COVID-19 and aid in answering the second part of this paper's thesis. That of the data quality analysis.

For a paper to be considered reproducible it should clearly state the pre-processing of its data, any normalization or additional data quality corrections balance correction and such. Along with that their Feature Selection and ML algorithm should also be known. Finally, the data set used should be available publicly. In order to encompass as many papers as possible, both the raw dataset and the processed version can be accepted as part of the criteria. Further, easing of the criteria comes from the availability as this paper will consider situations where there must be a specific request sent to an email as valid as well.

As stated in the "Approach" section, a total of 43 papers were selected from the databases, of those the number which are considered reproducible by this paper is 16, or 37.2%. The reproducible papers are the following, [8]–[10], [13], [15]–[19], [21], [24], [25], [29], [31], [47], [48]. With this number it can safely be said that the general reproducibility of COVID-19 research paper results, is low. This can be due to many reasons, one of the most common is the withholding of data due to patient anonymity, though understandable it still must be noted, as part of this paper.

2.4 State of the Art

The State of the Art will attempt to establishes the current work being done COVID-19 severity and infection prediction. There are two main reasons for making this section. The first, it so establishes the effectiveness of different supervised ML algorithms, Feature Selection techniques and pre-processing steps that works best for the COVID-19. The second is to establish a base line against which the custom implementation of this paper will be judged. This will give an objective measure of the performance of the ML solution with regards to the existing body of work

2.4.1 Performance Metrics

The metrics used for the evaluation of the ML algorithms are going to be the following:

- Accuracy - Number of Correct Predictions VS Total Number of Predictions
- Precision - Number of True Positives VS Total Number of Predictions
- Recall - Number of True Positives VS Total Number of Positives
- F1 Score - Harmonic Mean of Precision and Recall (Best indication of performance due to the high imbalance nature of COVID-19 datasets)

However, if the paper itself does not specify any of the aforementioned metrics, then the metric used within the paper will be used in substitution.

In Laatifi et. al. [17], a ML was developed in order to predict case severity. Data is 337 COVID-19 patients with the data set combining both biological and non-biological data, they purport to be the first study to do this. The Feature Selection technique used is Uniform Manifold Approximation and Projection (UMAP), this is reported to be the most optimal of the ones tested. The ML model included a Feature Selection step that goes into a Feature Engineering model after which a derision tree algorithm XGBoost is used. The reported Accuracy, Recall, Precision and F1 Score are all 100%. This is the highest reported metric of all papers looked at.

In Nasseem et. al. [13], a "Novel 3-phase deep learning ML framework is developed", with the aim to predict COVID-19 severity. The three phases are, in order; Variable Selection, Involved New Variable Creation and the application the novel ML framework along with other supervised ML algorithms. The paper had a dataset of 1214 patients, selected from 1228 admitted patients. The Neo-V framework resulted in an accuracy of 87.67% with a precision of 61%, a recall of 33.33% and an F1 score of 43.11%.

In Jaing et al. [7], the paper aims to create a ML framework which can be used to identify the characteristics during the illness that predict a move from mild to severe symptoms. The identification of the characteristics will allow for the creation of an AI powered tool which would be able to predict which patients are at risk of developing severe symptoms. The paper uses a Feature Selection technique that combines Information Gain which is used to ascribe a rank to a feature, Gini index based on the impurity of the dataset, Chi-Squared used to establish dependence between the given feature and the class of severity. Based on the Feature Selection above the characteristics identified are as follows; ALT, Myalgias, Hemoglobin, Gender/Sex, Fever, Na+, K+, Lymphocyte Count, Creatinine, Age, White Blood Count. Additionally, several

ML algorithm were applied upon these characteristics with the top performing being Support Vector Machine (SVM), with accuracy of 80%, and Decision Tree (DT), with an accuracy of 70%. No recall, precision or F1 score were calculated.

In Batista et. al. [45], the paper aims to create a ML based COVID-19 infection predictor based on Emergency Care (ER) admission exams. The paper applied multiple algorithms to the problem those including Neural Networks (NN), Random Forest (RF), Gradient Boosting, Logistic Regression (LR) and SVM, with the most performant algorithm being SVM at and accuracy of 84.7%, precision at 77.8%, recall at 67.7% and an F1 score of 72.4%.

In Schwab et. al. [18], the paper uses a systematic approach to evaluating ML based clinical predictive models that predict the outcome of a COVID-19 infection test, chances of hospitalization and or expected severity. All models were applied to a data set of 5644 patients with the models chosen being XGBoost (Extreme Gradient Boosting Tree), RF, NN, LR and SVM. From their results the most performative model for COVID-19 test was XGBoost at acc. of 66%, precision of 21%, recall at 75% and F1 score of 32.8%. For hospital admissions RF was the best with accuracy at 92%, precision 43%, recall at 55% and F1 score of 48.3%. Finally, for case severity SVM was the most performant with an accuracy of 98%, precision of 53%, recall at 90% and F1 score of 66.7%.

In Alakus et. al. [21], a ML based predictive algorithm is made in order to identify positive COVID-19 cases. For that purpose, a combination of different deep learning models were tested in order to identify the most effective. The deep learning models used are Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), Convolutional Neural Networks with Long-Term Short-Term Memory (CNNLSTM), Convolutional Neural Network with a Recurrent Neural Network (CNNRNN), Long-Term Short-Term Memory and a Recurrent Neural Network (RNN). The most performant of which is the CNNLSTM at an accuracy of 92.30%, precision of 92.35$, recall at 93.68 and F1 score of 93%.

In Brinati et. al. [19], a ML based system was created in order to detect COVID-19 infections based on routine blood exams. Two models are developed based on data from 279 patient. An additional ML was also created that as a decision aid. The two main ML models were created using RF and LR with the most performant of the two being RF with an accuracy of 82%, precision of 83%, recall at 92% and an F1 score of 87.3%. The DT algorithm has an accuracy of 76% with no further information on precision and recall.

In Shoer at. al. [20], a model is developed for estimating the probability of a PCR test returning a positive result based on a 9 feature survey with features from the Nation Symptom Survey. The survey generated has been filled out by over 43 752 people in Israel with 498 of them having actually tested positive. The model uses a LR model with only an AUG score given at 73.7%.

In Yang et. al. [22], creates a ML based model that is able to predict COVID-19 infections based on laboratory blood test, along with demographic data such as age, sex and race. The model was trained on a data set consisting of 3356 COVID-19 test of which 1402 positive and 1954 negative. The paper tested multiple algorithms such as RF, LR and Gradient Boost

(GBDT) of which the GBDT model was the most performant at an AUC of 87.9% recall of 80% and a specificity of 82.5%, no data on accuracy and or precision.

In Sotlan et. al. [23], the development of two early-detection models is facilitated powered by ML models and with data provided by hospitals. The data a is gathered during patient visitation to the ER of which of which the people admitted to the hospital would have their laboratory data, blood gas measurements and vitals collected. Three models were used for training, the linear LR and the non-linear "ensemble" RF and XGBoost. The XGBoost model ultimately proved to be the most accurate and thus was used on the Emergency Department model and the Admissions model. On the ED model XGboost achieved a AUC of 94% with a recall of 77.4% and specificity of 95.7%. The Admissions Model proved to be less performant.

In Cabitza et. al. [16], the aim is to create a alternative to the PCR test using the hemato-chemical data withing standard blood tests. The aim of this is to create and alternative method of identifying COVID-19 infections that do not have the pitfalls of PCR test. Five different algorithms were used on a data set of 1624 patients of which 786 had a positive PCR test. The five algorithms chosen were RF, Naive Bayes (NB), LR, SVM and K-Nearest Neighbor (KNN). The most performant of the in terms of accuracy and recall is SVM with 88% and 92% respectively, though RF managed to equal SVM in accuracy whilst having a work sensitivity at 86% and a better specificity at 91%, better AUC.

In Goodman-Meza et. al. [24], a diagnostic tool was developed for the sake of diagnosing COVID-19 patients, in the hopes of removing the bottleneck that is COVID-19 testing capacity. This is done using a data set containing laboratory data from ER room admissions and hospital admissions. the size of the data set is 1455 features, with 182 positive and 1273 negative. The ML algorithms tested on the model are RF, LR, SVM. NN, Stochastic Gradient Descent (SGD), XGBoost and ADABoost. The most performant of which was the SVM model with it achieving an AUC of 85%, recall at 68% and specificity at 85

In Aljame et. al. [8], a diagnostic is also developed for COVID-19 in order to make diagnosing COVID-19 patient become much quicker. For this they used a data set of 5644 data points of which 559 were confirmed COVID-19 cases, to train three different classifiers; Extra Trees, RF and LR, the three classifiers are used together in what the paper calls the ERLX model which uses KNN for imputing missing data and SMOTE for re-balancing. The classifier outputs are piped into an XGBoost model in the "second level" of the model with the final performance being, 99.73% AUC, 99/47% sensitivity and 99.99% specificity.

In Feng et. al. [48], the development of a COVID-19 diagnostic tool is being facilitated that does not use Compute Tomography (CT) images in the detection of early onset COVID-19 pneumonia. The data set used has a size of 164 patients of which 32 are set for the validation group thus only 132 are used for developing the model. The paper uses the LASSO model achieved the best performance against the validation set with an AUC of 93.8%, recall of 100% and a specificity of 77.8%. It should also be noted that the paper identified the following biomarkers as highly predictive for the diagnostics; age, IL-6, blood pressure, monocyte ration and fever classification.

In Soares [25], developed a diagnostic model with the hopes of circumventing the problems with testing capacity, especially in less fortunate countries, whilst also helping ease the burden on the healthcare sector. The data set used to do this consists of 599 subjects taken from the data set of Albert Einstein Hospital's data set of 5644 subjects originally, the reason for the culling is to minimize missing data in patient records. The model created is referred to as ER-Cov with three classifiers being used in order to train the model, them being SMOTEBoost (Gradient Boosting with the addition of SMOTE), SVM and ensembling. The framework achieves a specificity of 92.16%, Negative Predictive Value (NPV) of 95.29% and recall at 63.98%.

In Gao et. al. [26], the creation of a model that predicts COVID-19 mortality risk, the data used come from clinical data on admission in the hospital with 2160 patient being included in the data set. The model was trained using three ML classifiers; LR, SVM, GBDT and NN with the model AUC of 97.60%. The model further used L1 LASSO for feature selection via which the following bio-markers were found to be negatively correlated to mortality; lymphocyte, CKD, fever, platelet count and albumin. With the high risk bio-markers being; sex (specifically being male), sputum, BUN, respiratory rate, D-dimer, co-morbidities, and age.

In Vaid et. al. [29], creates a severity prediction framework based on the Electronic Health Records (EHR) of patients who have tested positively for COVID-19. All patient was collected from the Mount Sinai Health System in New York, with a 1514 patient training model being validated against 2202 patients from other hospitals. The ML algorithms used in paper are XGBoost and baseline comparator models with XGBoost showing the best predictive abilities with an 89% AUC for mortality at 3 days along with an AUC of 80% for critical even prediction again at 3 days.

In Yan et. al. [44], a severity prediction model was created in order to alleviate pressure for healthcare services. The paper uses a data set of 404 infected patients with ML models being used to identify three bio-markers that had the highest predictive power for case severity, them being; Lactic Dehydrogenase (LDH), Lymphocyte and C-Reactive Protein (CRP). A single-tree XGBoost algorithm was applied to the data set which had a precision of 89%, a recall at 100% and an F1 score of 94% for predicting "Death", with the survival prediction being at precision 100%, recall 83% and F1 score of 91%.

In Wang et. al. [43], the development of mortality prediction models is carried out using data from the People's Hospital of Jiangxia District in Wuhan, China. The data set used has a total of 296 COVID-19 positive patients, with 277 discharged from the hospital. The paper used Stepwise Akaike Information Criterion to select "base line data" and used XGBoost for mortality prediction. The results for XGBoost came in at 88% AUC, recall at 92.31% and a specificity of 77.44.

In Rechtman et. al. [47], developed a model to predict mortality rates with data from the hospital system of New York. The data set used had 8770 confirmed cases of COVID-19, coming from 53 hospitals in New York City, with 1114 of those cases being fatalities. The paper identified that the following bio-markers were the most predictive for high mortality; age (old age), sex (male), elevated Body Mass Index (BMI), elevated respiratory rate and Chronic Kidney Disease (CKD). The paper used XGBoost for mortality prediction with a resulting AUC of 86%.

In Bertsimas et. al. [31], the paper aims to create a mortality prediction tool for the improvement of patient management in hospitals. The ML development data used within the paper comes from a total of 33 hospitals and medical facilities within Europe and U.S. with the total amount of patients in the data set being 3927 all positive for COVID-19. XGBoost was the ML algorithm of choice with the resulting AUC of anywhere between 81% and 92%. With the bio-markers identified as predictive for higher mortality being; older age, higher CRP levels, blood urea nitrogen, creatinine and decreased oxygen saturation.

In Guan et. al. [32], a model was developed for the prediction of COVID-19 mortality. A data set of 1270 patients was used with the patient being dispersed between the Sino French New City Branch and the Optical Valley Branch of the Wuhan hospital. Feature Selection was done using LASSO with XGBoost being employed to rank the features output from whilst also being used for risk prediction. The most predictive features turned were found to be; age, levels of high-sensitivity C-reactive protein (hs-CRP), lactate dehydrogenase (LDH), ferritin, and interleukin-10 (IL-10), with the subsequent risk prediction achieving a score of 90% accuracy, 85% precision, recall at 96% and an F1 score at 90%.

In Booth et. al. [33], the usage LR and SVM is facilitated in the prediction COVID-19 mortality in COVID-19 positive patients. The data set consists of 398 patients, of which 43 died and 355 lived, gathered from the University of Texas Medical Branch. During the study both LR and SVM were used to predict mortality with the SVM model achieving the better metrics with 91% recall, 91% specificity and an AUC of 93%.

In Sun et. al. [34], created a severity prediction tool based on the data received from Shanghai medical centers. The data set used, has 336 COVID-19 positive patients with an additional 220 clinical and laboratory observations added. SVM was used for the prediction for severity with the algorithm achieving 97.57% AUC and a recall are of 100%.

In Yao et. al. [15], develops a severity prediction model using data gathered from Tongji Hospital Affiliated to Huazhong University of Science and Technology. The data set used consists of 137 confirmed cases of COVID-19 with 17 of them are mild. 45 moderate and 75 sever, with 21 of the severe cases resulting in a fatality. The ML algorithm used SVM for severity prediction with being an overall accuracy metric of 81.48%.

In Hu et. al. [35], developed a ML model that predicts mortality risk of severe COVID-19 patients, with the data set of 183 patients with 115 survivors and 68 fatalities with an additional 64 cases of severe infection were gathered form the Optical Valley Branch of Tongji Hospital, Wuhan. Two models were used for severity prediction, LR and RF, with both performing similarly with respect to their AUC scores thus the LR model was chosen for the final due to its simplicity and interoperability. The LR model had an AUC pf 88.1% with a specificity of 83.9%.

In Levy et. al. [10], the paper developed a calculator for the survival rate of a patient in the subsequent 7 days. The prediction model was created using LASSO regression with a data set of 11 095 patients, of which there were 2596 fatalities. The prediction was achieved an AUC of 86% for internal and 82% for validation data sets.

2.5 Future Implementation

Based on the State of the Art research it is clear that for severity prediction; SVM and XGBoost a the proffered algorithms to use. With most of the missing data problems either being removed or corrected with imputation, though as Laatifi et al. [17], the imputation of data might not be appropriate due to the potential introduction of false measurements. Data balance on the other hand is either corrected via biasing selection criteria such, such as time period of in which data selection was done, random oversampling or via more complex techniques like SMOTE.

For the subsequent implementation phase the aim is to first identify the most predictive biomarkers after which training, and severity prediction would be done by a Decision Tree algorithm. The DT was used here due to several reasons. The first, was the generally high metrics of tree-based algorithms with reference to XGBoost, however DT is both simpler and interoperable.

Chapter 3

Methodology

3.1 Introduction

The severity prediction model will leverage a data set given by Medisch Spectrum Twente (MST) anonymised before being used for training or processing. The data set given consists of 3081 data points consisting of 627 patients. The data set consists of 26 bio-markers collected from clinical blood tests, along with gender and age this makes 28 bio-markers. Additionally, the severity of the sickness within the data set, something this paper will follow, is divided in 5 distinct numeric variables. A severity of 0 denoting a negative COVID-19 disposition, severity of 1 is very mild, severity of 2 moderate, 3 moderate to serious, 4 sever/life threatening.

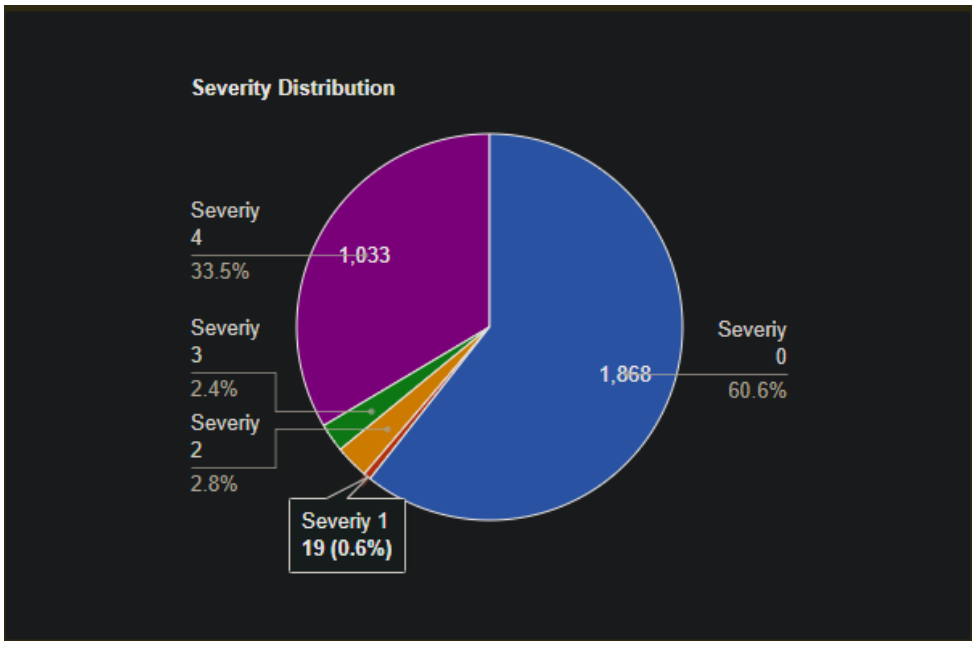

Figure 3.1: Distribution of Severity Variables Within the COVID-19 Data Set

From figure above it can be seen that the data set is highly imbalanced, with the severity 0 and 4 categories being far too overrepresented. This will have to be remedied within during the data processing and pre-processing stages.

3.2 Evaluation and Metrics

As stated on the section in the "Approach" section in the Literature Review, the main evaluation criteria will be Accuracy, Precision, Recall and F1 Score with F1 score being taken as the cumulative metric for predictive power of the model, Area Under the Receiver Operating Curve or AUC, though used by many of the papers in the "State of the Art", will not be used here. This was decided due to the high level of imbalance within the data set making the AUC far less indicative of the performance as opposed to the F1 score, this was further discussed with the Critical Observer of this research.

3.3 Data Pre-Processing

The main aim of this section is to try and correct any grave data quality issues that need to be fixed before re-balancing and Feature Selection can take place. The pre-processing step is meant to correct missing values and normalize the data formats.

The first steps taken was the removal of duplicate data for each patient. The data set given consisted of blood tests taken for a patient at a given time, thus some patients have more than data point associated with them. To remove the duplicates the most recent data blood test was taken of the patient and the rest was discarded. The new value distribution can be seen below.

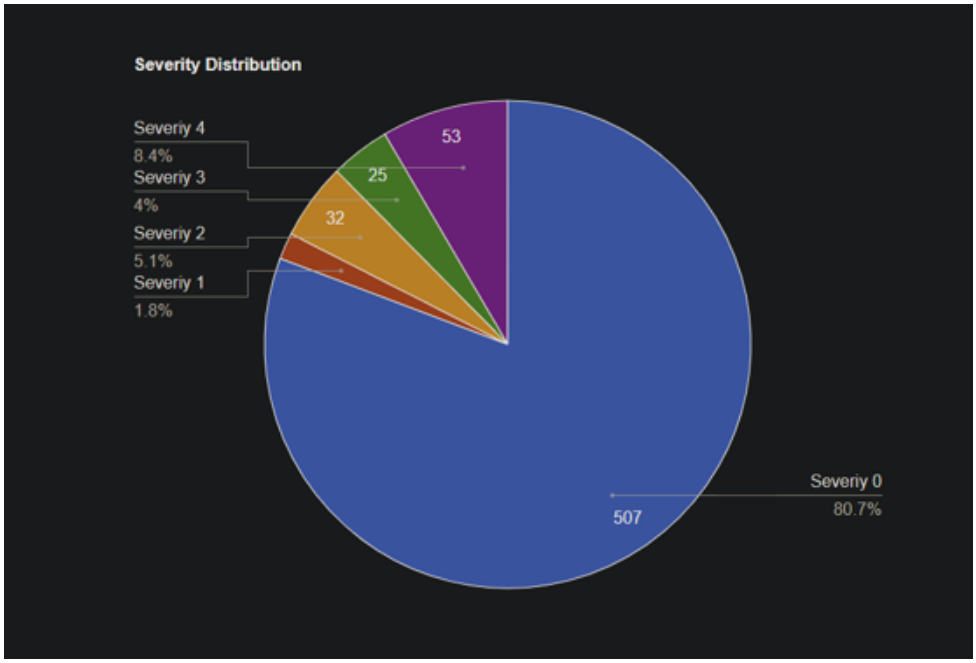

Figure 3.2: Distribution of Severity Variables After Duplicate Removal

From 3.2 we can still see that our data is still heavily imbalanced, thus the decision was taken to remove the severity 0 data points. The decision was taken based on two factors, the first is that correction of data imbalance within the data set. Furthermore, as can be seen from the "State of the Art" the state of research concerning itself with diagnostic prediction is already well formed, thus this paper concerns itself with severity prediction rather than diagnostics.

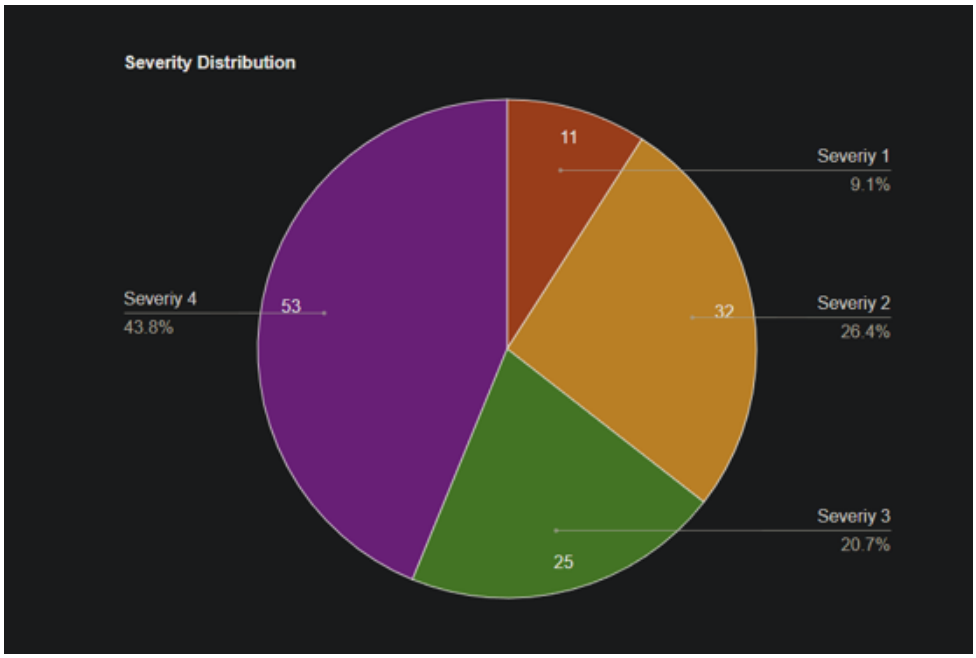

Figure 3.3: Distribution of Severity Variables After the Removal of Negative COVID-19 Cases

As can be seen in 3.3 the removal of severity 0 has allowed the rest of the categories to be better represented, though there is still a balance problem with overrepresented severity 4. The data set can now be put rebalanced in order to correct that, this could not be done before due to the overwhelming amount of data for severity 0.

An additional step that must be taken is to normalize the data due to format variation, The way this was done was via z-score normalization.

Bio-Marker	% of Missing Data
KL-6	96.69%
IL-6	97.52%
ALAT	90.91%
Baso	85.12%
Bili_tot	90.91%
CKD-epi	19.91%
CRP	28.10%
Ca	92.56%
Cl	95.04%
D_Dimeer	86.78%
Eo	85.12%
Ferretin	86.78%
Fibrinog	87.60%
K	17.36%
Kreat	19.01%
LD	79.34%
Leuco	38.02%
Lymfo	85.12%
MPV	38.02%
Mono	85.12%
Na	19.83%
Neutro	85.12%
P	80.17%
PCT	80.43%
Trombo	38.02%

Table 3.1: Table Showing Each Bio-Marker's Proportion of Missing Values

The final, step that must be taken in the pre-processing steps is to remove the missing values that can be seen in 3.1. Based on the data from the table it can be seen that only 8 of the 26 original blood test-based bio-makers have a proportion of missing data below 50%. Thus, the cutoff point of for trimming the classes that have too much missing data was set at 50%.

Though many of the paper choose to use data imputation for filling in the missing data, this becomes impractical for this situation due to the large amount of missing data, this would result in the imputed data being biased. A further consideration must also be given to the fact that SMOTE will be used later to re-balance the data set, this would have further exasperated any biases introduced during data imputation.

After pre-processing that data set left was now having the following bio-makers left:

- Age
- Sex
- K - Potassium
- CKD-epi - Chronic Kidney Disease

- Kreat
- Na - Sodium
- CRP - C-Reactive Protein
- Leuco - Leucocyte
- MPV - Mean platelet volume
- Trombo - Thrombocytes

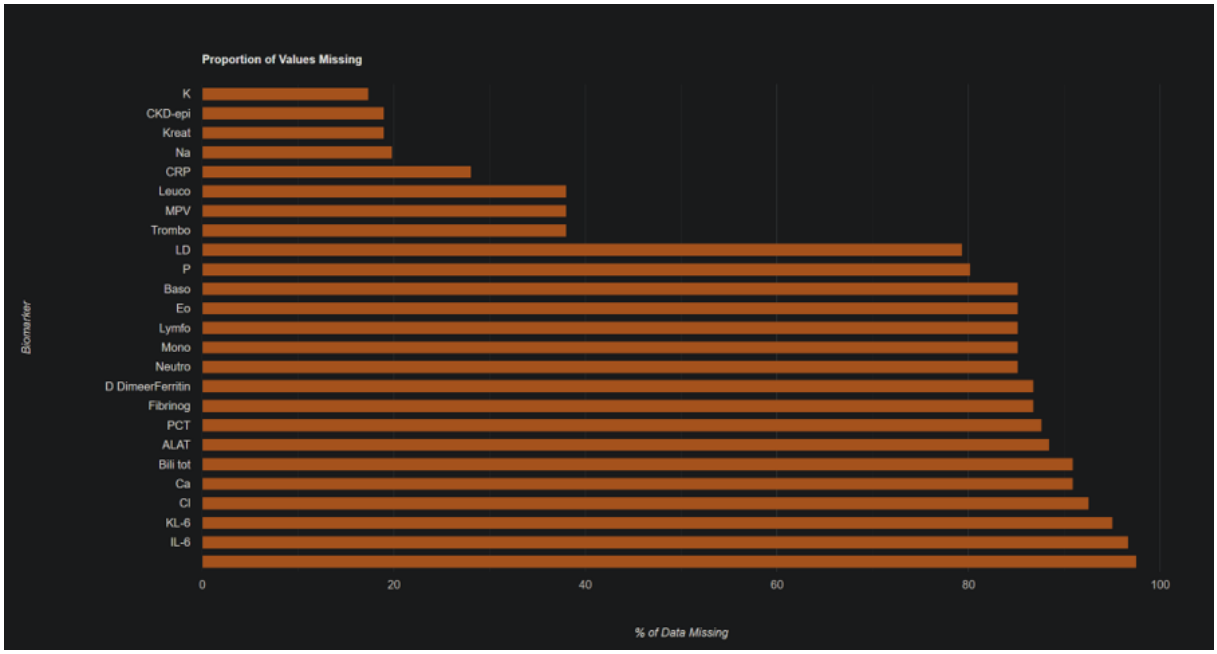

Figure 3.4: Proportion of Each Bio-Marker's Missing Data Graphed

3.4 Data Balancing and Feature Selection

As mentioned in the data pre-processing in order to rebalance the data SMOTE will be used. SMOTE was chosen over other techniques like Random Undersupplying due to the good results it showed within the papers that used it [13], [25]. With the introduction of SMOTE the data generated by it, is used only within training data, with the training/validation data being a 30%/70% split.

Feature Selection will be done via Logistic + LASSO Regression, this is done for two reasons. First is due to the linear nature of the technique meaning that interoperability is preserved, the second is the proven effectiveness in the use of LASSO in Feature Selection contexts with regards to COVID-19, as see in Feng et al., Gao et al., Guan et al. and Levy et al. [10], [26], [32], [48]. Furthermore, Logistic + LASSO has also been shown to works well in genetic selection circumstances as well as seen in Ma and Huan [49].

3.5 Severity Prediction Model

Finally, for the severity prediction the Binary Decision Trees (BCT) will be used. BCTs were chosen due to their simplicity in along with interoperability, Furthermore Decision Tree based algorithms have shown to be quite effective in COVID-19 severity and mortality prediction as seen from the wide usage of XGBoost and GBDT being widely use in the "State of the Art". The BCTs are setup where one branch denotes severity 4 while the other groups severity 1, 2 and 3, this is done since we are looking for severe case for which resources must be prioritized for. Thus, eliminating the factor requiring granular severity classification.

Chapter 4

Results

4.1 Results of Model vs State of the Art

From the Feature Selection the positively predictive bio-markers for a severity of 4 are; increased age, MPV, and elevated CRP, elevated Trombo. While negatively linked are decreased age (data set cannot account for below 18 years old), Na, K, Kreat and Leuco. This goes in line with what has been observed in the "State of the Art".

The results of the prediction when against the validation data for the Binary Decision Tree model comes with an Accuracy of 79.9%, Precision, Recall of 72.2% and an F1 score of 70.6%. As compared to the "State of the Art" these results are within the lower half if the predictive models, indicating that this is a negative result.

4.2 Discussion

As mentioned in the "Results" chapter, the accuracy and F1 score received by the ML model are on the lower end of the spectrum seen in the "State of the Art", which would indicate a negative result. Though the steps outlined in the methodology section there may be room for improvement of the model, before any limiting factors are considered. One of the first things that comes to mind when discussing what can done differently is the duplicate data removal. The problem that comes with this is that the data set is considered, per data point rather than per patient. This means that changes in severity of a patient over time, along with the addition training and validation, data is not considered. This can have an effect on the predictive power of the model which would cause a negative result.

Another consideration is the lack of variables after the trimming of the missing data. It is posited that, just as SMOTE was used to correct the data set balance by generating additional data, SMOTE could also be used to correct the proportion of missing data, by adding additional data points. Indeed, this was considered during the research as well though do the high level of missing data, it was deeded that something like this would introduced too much bias into the model.

Finally, though the model is fairly removed from any major biases there is indeed one bias within the model that is introduced due to the age bio-marker. The Literature Review shows that age is a powerful predictor of COVID-19 severity, this can lead to a higher accuracy however,

it also introduces bias. The bias comes in the form of data points win an increased age marker would get influence the over Feature Selection and predictive model due to the age. Thus, this can be another point where predictive power is lost.

4.3 Limitations

One limitation that is forthcoming is missing data within the data set. As it currently stands the majority of bio-markers have had to be removed due to not having enough data in the class. This is a problem because many of the highly predictive bio-makers that were identified in the "State of the Art" section were within the range of trimmed data. Thus, it is expected that with an expanded data set the predictive power of the model would be improved.

Furthermore, the amount of data in the data set was also rathe limited at 627 patients, this resulted in both less training and validation data. An expanded data set is expected to further improve the predictive of the model.

Finally, the data has major balance issues which forced us to delete material along with using SMOTE to generate new data points. A more balanced data set, just like with a larger one, is expected to improve the model. In this case because during feature selection the model will be able to be trained on negative COVID-19 patients.

Another limitation that is present is the lack of diversity in patients, with most patients coming from the Eastern regions of the Netherlands, this would have a negative impact on the generalizability of the model. Though this is somewhat compensated for during Feature Selection it the problem is only mitigated rather than fully corrected.

4.4 Future Work

From the research done it is clear that the usage of multiple ML algorithms during training, in order to isolate the most performant, is a popular approach. Thus, one aspect of this work that can be expanded upon in the addition of more ML algorithms one of note would be SVM, given it is the consistent choice amongst most papers. In addition to SVM, XGBoost also provide to be both performant and widely used, this would also give us a good comparison point between XGBoost and the Binary Decision Tree models. Lastly, though limited in the "State of the Art" Deep Learning algorithms have also shown to have great promise in predicting COVID-19 severity and or mortality, thus a good addition as well.

Chapter 5

Conclusion

To conclude the research is safe to say that the ML solutions can be applied to machine learning models and could produce very promising results. However, COVID-19 data sets are very prone to data quality issues, thus effective steps must be taken in order to transform the data into a state where ML classifiers can be applied.

The research has shown that the most common data quality concerns in COVID-19 data sets are missing data, data format incongruence and imbalance of classes. These issues can be fixed via data imputation, normalization, and various types of under/over-sampling. However, care must be given to their usage since some techniques can introduce biases and others could lead to erroneous outputs.

However, both the 'State of the Art" our custom implementation show that even with the various data quality pitfalls, ML models must be effective tools in diagnosis, mortality prediction and severity predictions. Keeping in mind that negative result achieved in this paper their certain pitfalls that exists where predictive power can be easily lost.

With that in mind, there is evidence aplenty showing that ML models are effective tools mortality and severity prediction. Based on this study's resulting certain correction must be done in order to achieve a positive result. Overall, further work within this field is suggested as it promises to effectively alleviate stresses of the healthcare industry during the pandemic and beyond.

Bibliography

[1] *Who coronaviurs (covid-19) dashboard*. [Online]. Available: https://covid19.who.int/table.

[2] *Report of the who-china joint mission on coronavirus disease 2019 (covid-19)*. [Online]. Available: https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19).

[3] N. Alballa and I. Al-Turaiki, "Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: A review," *Informatics in Medicine Unlocked*, vol. 24, p. 100 564, 2021. DOI: 10.1016/j.imu.2021.100564.

[4] B. Kitchenham, "Procedures for performing systematic reviews," Department of Computer Science, Keele University, UK, Keele University. Technical Report TR/SE-0401, 2004.

[5] B. N. Ashraf, "Economic impact of government interventions during the covid-19 pandemic: International evidence from financial markets," *Journal of Behavioral and Experimental Finance*, vol. 27, p. 100 371, 2020. DOI: 10.1016/j.jbef.2020.100371.

[6] E. de Jonge, R. Kloppenburg, and P. Hendriks, "The impact of the covid-19 pandemic on social work education and practice in the netherlands," *Social Work Education*, vol. 39, no. 8, pp. 1027–1036, 2020. DOI: 10.1080/02615479.2020.1823363.

[7] 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: 10.32604/cmc.2020.010691.

[8] 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.

[9] 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.

[10] 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.

[11] 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.2020.587937.

[12] 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.

[13] 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.

[14] Z. Zhu, T. Cai, L. Fan, *et al.*, "Clinical value of immune-inflammatory parameters to assess the severity of coronavirus disease 2019," *International Journal of Infectious Diseases*, vol. 95, pp. 332–339, 2020. DOI: 10.1016/j.ijid.2020.04.041.

[15] 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.

[16] F. Cabitza, A. Campagner, D. Ferrari, *et al.*, "Development, evaluation, and validation of machine learning models for covid-19 detection based on routine blood tests," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 59, no. 2, pp. 421–431, 2020. DOI: 10.1515/cclm-2020-1294.

[17] 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.

[18] P. Schwab, A. DuMont Schütte, B. Dietz, and S. Bauer, "Clinical predictive models for covid-19: Systematic study," *Journal of Medical Internet Research*, vol. 22, no. 10, 2020. DOI: 10.2196/21439.

[19] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of covid-19 infection from routine blood exams with machine learning: A feasibility study," *Journal of Medical Systems*, vol. 44, no. 8, 2020. DOI: 10.1007/s10916-020-01597-4.

[20] S. Shoer, T. Karady, A. Keshet, *et al.*, "A prediction model to prioritize individuals for a sars-cov-2 test built from national symptom surveys," *Med*, vol. 2, no. 2, 2021. DOI: 10.1016/j.medj.2020.10.002.

[21] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict covid-19 infection," *Chaos, Solitons and Fractals*, vol. 140, p. 110 120, 2020. DOI: 10.1016/j.chaos.2020.110120.

[22] H. S. Yang, Y. Hou, L. V. Vasovic, *et al.*, "Routine laboratory blood tests predict sars-cov-2 infection using machine learning," *Clinical Chemistry*, vol. 66, no. 11, pp. 1396–1404, 2020. DOI: 10.1093/clinchem/hvaa200.

[23] A. A. Soltan, S. Kouchaki, T. Zhu, *et al.*, "Rapid triage for covid-19 using routine clinical data for patients attending hospital: Development and prospective validation of an artificial intelligence screening test," *The Lancet Digital Health*, vol. 3, no. 2, 2021. DOI: 10.1016/s2589-7500(20)30274-0.

[24] D. Goodman-Meza, A. Rudas, J. N. Chiang, *et al.*, "A machine learning algorithm to increase covid-19 inpatient diagnostic capacity," *PLOS ONE*, vol. 15, no. 9, 2020. DOI: 10.1371/journal.pone.0239474.

[25] F. Soares, "A novel specific artificial intelligence-based method to identify covid-19 cases using simple blood exams," 2020. DOI: 10.1101/2020.04.10.20061036.

[26] Y. Gao, G.-Y. Cai, W. Fang, *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for covid-19," *Nature Communications*, vol. 11, no. 1, 2020. DOI: 10.1038/s41467-020-18684-2.

[27] C. Satici, M. A. Demirkol, E. Sargin Altunok, *et al.*, "Performance of pneumonia severity index and curb-65 in predicting 30-day mortality in patients with covid-19," *International Journal of Infectious Diseases*, vol. 98, pp. 84–89, 2020. DOI: 10.1016/j.ijid.2020.06.038.

[28] P. Bradley, F. Frost, K. Tharmaratnam, and D. G. Wootton, "Utility of established prognostic scores in covid-19 hospital admissions: Multicentre prospective evaluation of curb-65, news2 and qsofa," *BMJ Open Respiratory Research*, vol. 7, no. 1, 2020. DOI: 10.1136/bmjresp-2020-000729.

[29] A. Vaid, S. Somani, A. J. Russak, *et al.*, "Machine learning to predict mortality and critical events in a cohort of patients with covid-19 in new york city: Model development and validation," *Journal of Medical Internet Research*, vol. 22, no. 11, 2020. DOI: 10.2196/24018.

[30] L. Yan, H.-T. Zhang, J. Goncalves, *et al.*, "A machine learning-based model for survival prediction in patients with severe covid-19 infection," 2020. DOI: 10.1101/2020.02.27.20028027.

[31] D. Bertsimas, G. Lukin, L. Mingardi, *et al.*, "Covid-19 mortality risk assessment: An international multi-center study," *PLOS ONE*, vol. 15, no. 12, 2020. DOI: 10.1371/journal.pone.0243262.

[32] X. Guan, B. Zhang, M. Fu, *et al.*, "Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized covid-19 patients: Results from a retrospective cohort study," *Annals of Medicine*, vol. 53, no. 1, pp. 257–266, 2021. DOI: 10.1080/07853890.2020.1868564.

[33] A. L. Booth, E. Abels, and P. McCaffrey, "Development of a prognostic model for mortality in covid-19 infection using machine learning," *Modern Pathology*, vol. 34, no. 3, pp. 522–531, 2020. DOI: 10.1038/s41379-020-00700-x.

[34] L. Sun, F. Song, N. Shi, *et al.*, "Combination of four clinical indicators predicts the severe/critical symptom of patients infected covid-19," *Journal of Clinical Virology*, vol. 128, p. 104 431, 2020. DOI: 10.1016/j.jcv.2020.104431.

[35] C. Hu, Z. Liu, Y. Jiang, *et al.*, "Early prediction of mortality risk among patients with severe covid-19, using machine learning," *International Journal of Epidemiology*, vol. 49, no. 6, pp. 1918–1929, 2020. DOI: 10.1093/ije/dyaa171.

[36] Z. Zhao, A. Chen, W. Hou, *et al.*, "Prediction model and risk scores of icu admission and mortality in covid-19," *PLOS ONE*, vol. 15, no. 7, 2020. DOI: 10.1371/journal.pone.0236618.

[37] J. Xie, D. Hungerford, H. Chen, *et al.*, "Development and external validation of a prognostic multivariable model on admission for hospitalized patients with covid-19," 2020. DOI: 10.1101/2020.03.28.20045997.

[38] Y. Zhou, Z. Yang, Y. Guo, *et al.*, "A new predictor of disease severity in patients with covid-19 in wuhan, china," 2020. DOI: 10.1101/2020.03.24.20042119.

[39] J. Gong, J. Ou, X. Qiu, *et al.*, "A tool for early prediction of severe coronavirus disease 2019 (covid-19): A multicenter study using the risk nomogram in wuhan and guangdong, china," *Clinical Infectious Diseases*, vol. 71, no. 15, pp. 833–840, 2020. DOI: 10.1093/cid/ciaa443.

[40] M. Luo, J. Liu, W. Jiang, S. Yue, H. Liu, and S. Wei, "Il-6 and cd8+ t cell counts combined are an early predictor of in-hospital mortality of patients with covid-19," *JCI Insight*, vol. 5, no. 13, 2020. DOI: 10.1172/jci.insight.139024.

[41] C. de Terwangne, J. Laouni, L. Jouffe, *et al.*, "Predictive accuracy of covid-19 world health organization (who) severity classification and comparison with a bayesian-method-based severity score (epi-score)," *Pathogens*, vol. 9, no. 11, p. 880, 2020. DOI: 10.3390/pathogens9110880.

[42] W. Liang, H. Liang, L. Ou, *et al.*, "Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with covid-19," *JAMA Internal Medicine*, vol. 180, no. 8, p. 1081, 2020. DOI: 10.1001/jamainternmed.2020.2033.

[43] K. Wang, P. Zuo, Y. Liu, *et al.*, "Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus disease-2019: A cohort study in wuhan, china," *Clinical Infectious Diseases*, vol. 71, no. 16, pp. 2079–2088, 2020. DOI: 10.1093/cid/ciaa538.

[44] L. Yan, H.-T. Zhang, J. Goncalves, *et al.*, "A machine learning-based model for survival prediction in patients with severe covid-19 infection," 2020. DOI: 10.1101/2020.02.27.20028027.

[45] A. F. de Moraes Batista, J. L. Miraglia, T. H. Rizzi Donato, and A. D. Porto Chiavegatto Filho, "Covid-19 diagnosis prediction in emergency care patients: A machine learning approach," 2020. DOI: 10.1101/2020.04.04.20052092.

[46] R. P. Joshi, V. Pejaver, N. E. Hammarlund, *et al.*, "A predictive tool for identification of sars-cov-2 pcr-negative emergency department patients using routine test results," *Journal of Clinical Virology*, vol. 129, p. 104 502, 2020. DOI: 10.1016/j.jcv.2020.104502.

[47] E. Rechtman, P. Curtin, E. Navarro, S. Nirenberg, and M. K. Horton, "Vital signs assessed in initial clinical encounters predict covid-19 mortality in an nyc hospital system," *Scientific Reports*, vol. 10, no. 1, 2020. DOI: 10.1038/s41598-020-78392-1.

[48] C. Feng, L. Wang, X. Chen, *et al.*, "A novel artificial intelligence-assisted triage tool to aid in the diagnosis of suspected covid-19 pneumonia cases in fever clinics," *Annals of Translational Medicine*, vol. 9, no. 3, pp. 201–201, 2021. DOI: 10.21037/atm-20-3073.

[49] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Briefings in Bioinformatics*, vol. 9, no. 5, pp. 392–403, 2008. DOI: 10.1093/bib/bbn027.